DEVELOPMENT OF AN EXPLAINABLE HYBRID MACHINE LEARNING FRAMEWORK FOR PROSTATE CANCER PREDICTION

BY

AMECHI ELOCHUKWU WESLEY
GOU/U22/CSC/874

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: PROFESSOR I. A AJAH

MAY, 2026.

APPROVAL PAGE

This research, titled "**DEVELOPMENT OF AN EXPLAINABLE HYBRID MACHINE LEARNING FRAMEWORK FOR PROSTATE CANCER PREDICTION**," 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.

	.....	.....
Supervisor	Signature	Date
External Examiner	.....	.....
External Examiner	Signature	Date
Dr. Frank Okebanama	.....	.....
HOD, Department of Computer Science and Mathematics	Signature	Date
Prof. I. A. Ajah	.....	.....
Dean, Faculty of Computing And Information Technology.	Signature	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.

_____

STUDENT'S NAME

REG. NUMBER

DEDICATION

This project is dedicated to God almighty and my amazing mother Mrs Mercy E. Amechi only through her constant moral, financial and spiritual support was I able to complete this project and I will forever be grateful to her.

(one paragraph sentence)

ACKNOWLEDGMENTS

My profound gratitude to God, who has guided and given me the knowledge and inspiration to embark on this project.

My sincere appreciation goes to my project supervisor Prof. Ifeyinwa A Ajah who provided me with valuable advices, ideas and guidance throughout the course of this project. Her unwavering support and expertise were instrumental in ensuring this project was a success.

To my parents, guardians, family and friends for their support, prayers and financial provision, to my friends for their care, love and support and to my colleagues who contributed one way or the other to the success of this project and making the journey a less-stressful and blissful one, I am sincerely grateful.

ABSTRACT

Prostate cancer accounts for the most male deaths due to cancer in Nigeria; however, the country suffers from lack of availability of accurate diagnosis infrastructure, especially in non-urban settings. Available AI-supported prostate cancer prediction systems rely heavily on imaging techniques such as Magnetic Resonance Imaging (MRI) and digital pathology slides, which make them hard to implement at locations lacking adequate infrastructure. The current research focuses on filling this void by designing a web-based machine learning system, ProstaPredict AI, for predicting the result of prostate cancer biopsies using clinical structured tabular data. This project used the Agile development approach and leveraged the Flask framework with an XGBoost classification model that was trained on 23 clinical and demographic variables. Data processing and scaling were done by the combination of StandardScaler and OneHotEncoder from scikit-learn package, while SHAP TreeExplainer was incorporated to provide visualization of important features along with each prediction. Data leakage test was performed to check the correctness of the feature selection process, and three models were tested using five performance criteria prior to deploying the XGBoost classifier. These three models all achieved random classification performance when evaluated with proper care in a leak-free environment, as indicated by their Reciever Operating Characteristics-Area Under Curve (ROC-AUC) values of around 0.50. This was believed to be because of the synthetic data used in training, as opposed to a failure on part of the system design. From the leakage test, it can clearly be seen that post-diagnosis attributes boosted the AUC of our models to above 0.97. ProstaPredict AI showcases how an explainable AI decision support system that is lightweight, role-based, and explainable can be developed and implemented without any specialized infrastructure setup. This system gives a proven platform that is instantly applicable on clinical data and develops an approach to identify data leakage problems in clinical datasets. It needs to be developed further for retraining on annotated biopsy data from Nigerian health facilities to improve prediction performance.

Table of Contents

List of Tables

List of Figures

CHAPTER ONE
INTRODUCION

1.1 Background of Study

Prostate cancer is a type of cancer that affects the prostate, which is reproductive gland in men that secretes alkaline fluid that constitutes semen. It affects men from the age of 60 and above, with 1.5 million new cases reported and 396,773 deaths worldwide in 2022 according to the World Health Organisation, prostate cancer is the second most diagnosed type of cancer and a major cause of cancer-related death among men.

Prostate cancer is the most common type of cancer among men in Nigeria, and like other black populations, Nigerian men present the disease at a more advanced stage, at an earlier age than in other ethnic groups (Akinremi et al, 2014). The World Health Organization estimates that the mortality rate for prostate cancer In Nigeria alone is 16.3 per ever 100,000 men, in 2022 an estimated 11,443 deaths was recorded. Prostate cancer is the major cause of cancer related death for men in Nigeria.

Over the years different techniques have been developed for the detection and diagnosis of prostate cancer, they include; Prostate Specific Antigen (PSA) screening, Digital Rectal Examination (DRE), Image testing (Ultrasound, MRI, Computer Tomography (CT) scans, Histopathology images etc), and Transrectal Ultrasound-guided (TRUS) needle biopsy, for a more definite diagnosis along with the Gleason grading system. Generaly these different techniques are used together to get accurate results.

There multiple issues that surround the use of these techniques, while PSA is highly sensitive it has low specificity and is imprecise, as both benign and malignant conditions elevate the serum marker meaning (David et al,2024). There are many different conditions like urinary tract infection, Benign Prostatic Hyperplasia (BPH) or Prostatitis can cause PSA levels to rise, all of which are prevalent conditions among the Nigerian population. So even though PSA screening is the most widely used tool for prostate cancer detection because of its high sensitivity it often produces false positives. DRE also often produces false negatives. A study carried out by Ozah et al, (2023) in 2023 shows that despite its high sensitivity, specificity and diagnostic accuracy, DRE still isn’t enough to diagnose prostate cancer diagnosis. In the study only 43 out of 94 patients

(45.7%) had abnormal DRE findings, despite all patients being under clinical suspicion of having prostate cancer.

Transrectal Ultrasound (TRUS) is the major technique used to guide biopsies in Nigeria, despite the fact that there are superior biopsy guidance techniques such as transperineal route which has low infectious complication risk (Tolani et al, 2025). TRUS has low sensitivity for cancer detection, and the diagnostic yield of additional biopsies performed on hypoechoic lesions is negligible (EAU Guidelines 2026). It is also generally unavailable in most Nigerian medical facilities. A biopsy has it's own set of issues such as the wide variations in prostate biopsy practice patterns among Nigerian urologists, these differences suggest that standardisation is necessary to increase consistency in patient outcomes (Biyouma et al, 2025). Carrying out biopsies also come with the risk of infections, bleeding and pain.

With the advancement of technologies such as; Computer-Aided Diagnosis (CAD) systems and Clinical Decision Support Systems (CDSS), artificial intelligence has become a very useful tool for the detection and diagnosis of prostate cancer (Olabanjo et al, 2023). Medical image datasets are now being used to train machine and deep learning models to help with detection of the disease. AI increases diagnostic accuracy by analyzing medical image datasets faster, more efficiently and more precisely than traditional methods (Arita et al, 2025).

The study conducted by Perera et al. in 2021 was focused on the development of a machine learning algorithm based on PSA, free-PSA, age, and free-to-total PSA ratio in the sample size of 4,548 individuals, with 3,638 cases used for model training, and 910 cases used for the test. This algorithm increased the accuracy of diagnosis of prostate cancer with the AUC score of 0.72 compared to the use of PSA (AUC 0.63).

Barlow et al showed that Accuracy of 91.5% was possible to reach in prediction of high-risk prostate cancer cases through the integration of standard scaling, SVMSMOTE sampling and AdaBoost algorithm in machine learning while taking into account changes of PSA and age, which improved AUC by 6.8%.

TRUS guided by AI is able to outperform human readers in detecting clinically significant prostate cancer. An intelligent and AI-assisted ultrasound can act as a good triage test and help detect cancer in community settings where the cost of MRI cannot be afforded. (Rajih et al, 2025).

Stanford University developed **ProCUSNet**, an AI tool that analyses standard ultrasound images collected during routine biopsy. It was able to detect 82% of clinically significant cancers and

identified 44% more lesions than human readers interpreting the same ultrasound images (Rusu et al, 2025).

Despite all these advancements, there has not been any application of Artificial Intelligence within the West African context, particularly Nigeria where just diagnosing prostate cancer is still a major challenge.

1.2 Problem Statement

Prostate cancer is the major cause of cancer related death among men in Nigeria, one major issue with the diagnosis and treatment of the disease is the lack of a centralized screening program in the country and the concentration of screening centers in major cities only. This has caused the testing for prostate cancer to be opportunistic, regardless of the increase in awareness of the disease and willingness to be screened. Despite the technological advancements with artificial intelligence which has led to the development of systems that can help with the diagnosis of prostate cancer, there has been no application of such system within the Nigerian context.

1.3 Aim and Objectives

The aim of this study is to develop an explainable hybrid machine learning framework that can accurately predict if a person has prostate cancer (benign or malignant).

The objectives of the study are to:

1. Collect and Preprocess prostate cancer patient data for machine model development.
2. Develop a model for Prostate cancer classification based on selected clinical features.
3. Design and implement a system that allows users to input patient data and obtain prostate cancer predictions.
4. Integrate explainability techniques for model interpretability.
5. Evaluate the performance of the proposed model using standard metrics.

1.4 Significance of the Study

This study is significant because Prostate cancer remains the number one cause of cancer related death in Males in the country. There is a huge gap between the number of pathology experts and diagnostic facilities compared to the number of reported cases, not to mention the unreported cases.

The study will also help to improve the accuracy and consistency of diagnosis even within areas where facilities and expertise is readily available.
Overall the study will contribute to AI research in the medical field in developing countries.

1.5 Scope of the Study

This study focuses on the development of an explainable hybrid machine learning framework
That is designed specifically to predict and classify prostate cancer using data from available datasets, with a primary emphasis on the Nigerian healthcare context.
The study focuses on solely on Prostate cancer and does not extend to other types of cancer or any others diseases.

1.6 Limitations of the Study

There are various limitations that may affect the scope, generalizability, and real-world application of the findings, which include:

1. The system is deployed as a prototype, a proof-of-concept demonstration and has not been tested with any formal clinical trials or validated with live patient data of any form.
2. This study focuses only on the detection and classification of prostate cancer, not full diagnosis. The framework is not intended to be an independent diagnostic system.
3. The lack of availability of Prostate cancer related datasets is a major hindrance to this study.

1.7 Operational Definition of Terms

For the purpose of this study, the following terms are defined as follows;

1 **Prostate Cancer** - A malignant tumor that forms in the prostate gland, which a small gland in the male reproductive system.

2 **Histopathology Images –** Images gotten from prostate tissue samples that have been removed through biopsy then processed, stained and scanned a at high resolution.

3 **Gleason Grading System -** A widely used pathological scoring system for assessing the aggressiveness of prostate cancer based on the patterns of cancer cells in histopathology images.

4 **Deep Learning -** A type of machine learning that uses multi-layered artificial neural networks to learn complex patterns and features from large datasets.

5 **Hybrid Deep Learning Framework -** An architecture model that combines two or more different deep learning approaches in order to combine their strengths.

6 **Computer-Aided Diagnosis (CAD) -** A system that helps clinicians make diagnostic decisions using computational methods.

7 **Clinical Decision Support System (CDSS) -** A health information technology tool designed to help healthcare professionals in make clinical decisions by providing relevant, organized, and analyzed information.

8 **Explainable Artificial Intelligence (XAI)** - A set of techniques and methods that allow humans understand the decisions and outputs of AI models.

9 **Convolutional Neural Network (CNN)** – This is a type of deep learning model specifically designed for processing grid-structured data such as images. CNNs use layers of filters to detect visual patterns such as edges, shapes, and textures.

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

The addition of artificial intelligence to medical diagnostics has created new ways to detect and manage life-threatening diseases, with cancer diagnosis being one of the most actively explored areas. Prostate cancer, in particular, has become a significant target for AI-based diagnostic solutions given the number of cases and deaths that occur globally and how complex it is to accurately classify it. Detection of prostate cancer through deep learning and conventional methods has attracted much attention in the recent past, indicating an increasing realization and acknowledgment that automated systems can play a significant role in the process of clinical decision-making in oncology (Iqbal et al., 2021). This literature review delves into the theoretical and empirical grounds on which the current study is built, particularly the conventional diagnosis methods, artificial intelligence applications in healthcare, and in particular, the application of deep learning in detecting prostate cancer.

2.2 Theoretical Background

The theoretical foundation of this study sits at the combination of digital pathology, computer vision, and explainable artificial intelligence. The growing availability of digitized histopathology slides, combined with advancements in deep neural network architectures, makes it ever more feasible to train models that can recognize and classify tissue-level patterns with clinical expert-level accuracy. AI and histopathological images play a growing role in detecting prostate cancer, with surveys highlighting the different ways machine learning approaches are applied to tissue-based diagnosis (Ayyad et al., 2021). An essential part of this study is the principle that AI models deployed in clinical settings must not only be accurate but also interpretable; this is a requirement that has driven the development of Explainable AI (XAI) as a distinct and necessary subfield within medical AI research.

2.3 Overview of Prostate Cancer

Prostate cancer is among the biggest medical challenges that men all over the world face, and it presents one of the most serious public health challenges in sub-Saharan Africa. As pointed out by GLOBOCAN 2022, prostate cancer is one of the most common cancers that are diagnosed all over the world, with millions of new cases registered each year, while the burden of mortality is high

in low and middle-income countries (IARC, 2022). In Nigeria, the situation is really serious. Prostate cancer is the most common cancer diagnosed in Nigerian men (IARC, 2022).

Beyond incidence figures, the factors that lead to late presentation and poor outcomes in Nigeria are well documented. Various factors influencing the uptake of prostate cancer treatment in Nigeria include limited awareness, cultural beliefs, financial constraints, and inadequate access to diagnostic and treatment facilities, all of which contribute to patients presenting at advanced and less treatable stages of the disease (Iheanacho & Odili, 2024). The situation is worsened by gaps in clinical infrastructure. A study examining the detection and management of localized prostate cancer in Nigeria found that patients, caregivers, and healthcare providers identified timely diagnosis, a shortage of specialist pathologists, weak referral systems, and limited availability of diagnostic equipment as major barriers across many Nigerian hospitals (Tolani et al., 2024).

From a clinical point of view, the disease is primarily characterized and staged using the Gleason grading system, which classifies cancer aggressiveness based on the appearance of prostate tissue under microscopic examination. Studies on Nigerian patients treated for prostate cancer have highlighted the importance of clinicopathological characterization, with histological grading playing a central role in guiding treatment decisions (Abdus-Salam Abbas et al., 2024).

2.4 Conventional Methods of Prostate Cancer Diagnosis

Traditionally, prostate cancer diagnosis has relied on a combination of clinical examination, biochemical testing, and tissue analysis. Prostate Specific Antigen (PSA) blood tests have served as a first-line screening tool, usually followed by a transrectal ultrasound-guided biopsy when PSA levels are elevated. The biopsy specimens are then examined by a pathologist using light microscopy in order to assign a Gleason grade. While this pathway remains the standard, it is associated with several well-known limitations including inter-observer grading variability, the subjective nature of visual slide interpretation, and long turnaround times; challenges that become significantly more problematic in resource-constrained settings like Nigeria.

A scoping review and meta-analysis conducted on prostate cancer data from Nigeria indicated that process indicators for both diagnosis and treatment were lacking in the whole nation. This indicated that there was no capability among the current diagnostic systems to be able to handle the extent of disease burden (Tolani et al., 2025). Imaging techniques like MRI, CT scans, and bone scans are used for the staging and tracking of disease progression. There have been attempts to employ deep learning models to diagnose and classify bone lesions from staging CT scans of

prostate cancer. This demonstrates how even conventional imaging modalities can be improved using artificial intelligence (Belue et al., 2024).

2.5 Artificial Intelligence and Machine Learning in Healthcare

The role of artificial intelligence has revolutionized the way medical diagnoses take place in the last decade. The learning capability of machines has shown to be capable of learning features from medical data that differentiate between them, at times even outperforming the human expert for the task. (Olabanjo et al., 2023).

Beyond raw performance, the integration of AI in the medical field has increasingly demanded transparency. Hassan et al. (2022) demonstrated that it is possible to build models for prostate cancer classification from ultrasound and MRI images that are not only accurate but also provide interpretable reasoning that medical experts can evaluate and trust. This shift toward explainability reflects a broader sentiment that clinicians do not entirely trust AI in real-world settings, especially in environments where medical experts need to understand and justify diagnostic decisions.

2.6 Review of Related Works

A number of studies have directly shaped the direction and design of this research, and reviewing them closely reveals both significant contributions and notable limitations that this study seeks to address.

The work carried out by Ayyad et al. (2021) involved conducting an extensive review of the use of AI and histopathological images in diagnosing prostate cancer, involving the examination of many deep learning models and their conclusion being that the analysis of tissue images through the automation of AI can help to boost the accuracy and reliability of the detection and grading of prostate cancer. This review is quite exhaustive in terms of exploring CNN approaches, but one of its major weaknesses is that it is basically descriptive since it reviews the available literature without presenting a new perspective on the subject. Furthermore, explainability was not considered in the article, nor was the applicability of the models to low-resource settings.

Hassan et al. (2022) proposed a deep learning explainable AI framework to classify prostate cancer based on images obtained through ultrasounds and MRIs. It should be mentioned that Hassan et al.'s research is one of the rare cases where an XAI framework was introduced, showing how incorporating it into the process can increase physicians' trust in AI-generated results while providing valuable insights to take action. However, there are some flaws in the design of the

paper. First of all, it is worth mentioning that Hassan et al. used images taken by radiologists as opposed to using the gold standard approach of histopathology slides. Secondly, the authors did not focus on using any hybrid architecture combining CNN and Transformers to improve feature extraction.

In their study, Abbasi et al. (2020) showed that with the help of transfer learning technique, a deep learning model based on CNN was able to perform effectively at detecting prostate cancer using pathology images despite training on relatively smaller datasets. The limitations of this particular study are the scarce availability of prostate histopathology data annotated from the clinical environment of Nigeria and in its use of the single architecture technique and lack of explainability techniques in the analysis.

Khosravi et al. (2021) demonstrated the effectiveness of using a deep learning-based method in classifying prostate cancer by combining pathology and radiology image data, proving that the combination of both images yields better results. The idea behind multimodal fusion is an important contributor to the design of the hybrid model in this work. One downside is the necessity to have both the pathology and radiology images available at the same time, which is challenging to achieve in regions with poor health care facilities and where radiology equipment is not always available. Another issue is the absence of interpretability in their model.

A systematic review on the application of machine learning and deep learning algorithms for diagnosing prostate cancer using medical imaging was conducted by Olabanjo et al. (2023). This systematic review offers a general overview of the state of affairs in the area, analyzing the various types of imaging modalities and algorithm structures. One drawback of this study is the fact that, similar to other systematic reviews, it is descriptive in nature; that is, it highlights the existing gaps without offering a practical solution.

With regards to the Nigerian clinical background, Tolani et al. (2024) conducted a study of barriers and facilitators for diagnosing and treating prostate cancer in Nigeria, involving feedback from patients, carers, and clinicians. This study unambiguously highlights the necessity of having a diagnostic test that is scalable and fits into the context of the healthcare system in Nigeria. This particular study does not constitute the deep learning aspect of the research. However, it serves as a basis for the clinical aspect and the relevance of this research study. The main drawback of the paper is its failure to provide a technological fix to the barriers raised.

Likewise, Tolani et al. (2025) conducted a more extensive scoping review and meta-analysis of prostate cancer process indicators in Nigeria. Once again, even though this study adds more weight to the reasons why there is a need for AI-based solutions, it fails to address the technical issues involved in building those solutions.

2.7 Research Gaps

The literature review identified the following gaps, presented in Table (2.1) below, alongside the studies that identified them and how this work addresses each gap.

Table 2.1: summary of literature review

Works	Techniques	Work done	Limitations (Gaps)
Ayyad et al (2021)	Deep learning from histopathological images	Shows that the use of computerized image analysis improves diagnosis and grading of prostate cancer.	The analysis remains mostly descriptive.
Hassan et al (2022)	Deep learning based Explainable AI using ultrasound and MRI images.	Proposed an XAI solution using deep learning that would classify prostate cancer using ultrasound and MRI scans.	The framework utilized radiological scans like ultrasound or MRI scan instead of histopathological scans or biopsy data.
Abbasi et al. (2020)	Deep learning (convolution neural network) with transfer learning approach.	Demonstrated that a CNN-based deep learning model using transfer learning could detect prostate cancer from pathology.	Focuses on The use of CNN-based deep learning only.
Khosravi et al. (2021)	Deep learning using pathology–radiology fusion.	Found that using different types of imaging techniques together provides better results than when used separately.	Both pathology and radiology information must be available simultaneously, and the proposed model does not provide any explainability.

Olabanjo et al. (2023)	Application of Machine Learning and Deep Learning Models in Prostate Cancer Diagnosis Using Medical Images	Performed a systematic review on machine learning and deep learning techniques used in prostate cancer diagnosis,	This is just a systematic review, which only highlights the gaps without providing a specific recommendation on how to overcome them.
Tolani et al. (2024)	Detection and management of localized prostate cancer in Nigeria	Identified barriers and enablers to the diagnosis and treatment of prostate cancer in Nigeria	This is neither an in-depth learning nor a technical examination that provides any technical solution.
Tolani et al. (2025)	A scoping review of Prostate cancer in Nigeria	Builds the case for the necessity of implementing AI-based diagnostic tools in Nigeria.	Does not deal with the technical issues involved in the development of AI diagnostic tools.

Overall, from the review of the literature, there emerges a significant gap, which is that although considerable progress has been made in relation to the development of AI systems for classifying prostate cancer, there have only been a handful of research works conducted specifically addressing the diagnostic challenges in Nigeria, including the unavailability of clinical data, as well as the combination of hybrid machine learning algorithms with methods of explainability. The vast majority of studies either do not provide any means of explainability, make use of imaging techniques that are uncommon in Nigeria, or generally neglect the constraints of applying their findings to a sub-Saharan setting.

CHAPTER THREE
METHODOLOGY AND SYSTEM ANALYSIS

3.1 Introduction

This chapter provides the explanation of the approach used in designing and developing an Explainable Machine Learning based clinical decision support system known as ProstaPredict AI for predicting biopsy outcomes in patients with prostate cancer. This chapter explores the overall system architecture, data gathering and pre-processing, machine learning model development, and user interface design in ProstaPredict AI. Each of these aspects is considered in relation to the specific clinical question they address and the resource-limited healthcare environment in Nigeria.

3.2 Ethical Statement

The study was carried out in strict accordance with the ethics standards that are applied to scientific investigations related to medicine and computing. Since the current proof-of-concept study entailed no involvement of patients through direct recruiting, no clinical trials, nor collecting any information related to patients' health status, it was exempt from institutional ethical approval. It should be noted that all the data sets utilized for the purposes of developing and validating the ProstaPredict AI model belong to the publicly available data sets released for academic purposes.

3.3 System Analysis

3.3.1 Analysis of Existing Systems

An analysis of existing clinical decision support systems for prostate cancer detection and AI-driven systems for prostate cancer detection was carried out before developing the proposed ProstaPredict AI system. This was done with the aim of understanding what features these existing systems possess, on what basis they operate, and what their limitations are.

System 1: ProCUSNet (Stanford University, 2025)

ProCUSNet is an artificial intelligence (AI) tool designed by researchers from Stanford University for evaluating conventional ultrasound images obtained during prostate biopsies. The software uses a deep convolutional neural network to recognize suspicious areas in transrectal ultrasound images and is able to detect 82% of all significant cancers as well as 44% more lesions compared to a human observer reading the same images.

ProCUSNet is a remarkable technological development in the field of biopsy guidance. Nevertheless, for ProCUSNet to be applicable in practice, a set-up of TRUS, which would be able

to generate high-resolution images needed for CNN, has to be standardised. This will be challenging to achieve in the context of Nigeria since TRUS devices are not readily available except in teaching hospitals in large cities, while the computational resources necessary for real-time CNN inference are out of reach of many hospitals in the country.

System 2: Hassan et al. - XAI Classification from Ultrasound and MRI (2022)

A deep learning explainable AI technique for prostate cancer detection with an emphasis on ultrasound and MRI imaging was proposed by Hassan et al. The algorithm includes Grad-CAM to produce a saliency map that indicates which areas of the image have the most impact on the classification process.

This is a significant addition to the system in terms of explicitly considering the aspect of explainability in addition to its accuracy in classification. However, one of the drawbacks of this system is that it requires the availability of both ultrasound images and MRI scans for each individual patient at the same time, which is almost impossible to do in most Nigerian hospitals due to the non-availability of MRI scanners except in some specialized hospitals.

System 3: Abbasi et al. - CNN with Transfer Learning (2020)

According to Abbasi et al., a classification system based on a deep CNN architecture that employed transfer learning with ImageNet-trained weights was capable of diagnosing prostate cancer from histopathology slides even with small training sets.

While the methodology of the transfer learning paradigm is valid, this particular system relies solely upon histopathology image analysis and thus depends upon a functional digital pathology pipeline, which entails high-resolution scanning of the biopsy slides. This infrastructure alone is demanding enough, and there is no mention of any web-based interface for the same.

System 4: Perera et al. - ML Risk Stratification from Biochemical Parameters (2021)

A machine learning approach was proposed by Perera et al., which utilized PSA, free PSA, age, and the ratio of free-to-total PSA based on a dataset comprising 4,548 patients to predict prostate cancer. This method yielded a better AUC of 0.72 than PSA alone (0.63).

It is the one that most closely aligns with the design concept of ProstaPredict AI, because it uses tabular clinical data, not image data, and therefore can be run without the need for special diagnostic devices. However, unlike ProstaPredict AI, it was developed and validated using patients that were not African, and it does not have a web interface.

Summary of Existing System Limitations

System	Imaging Dependency	XAI Integration	Deployment Layer
ProCUSNet (2025)	TRUS required	None	None described
Hassan et al. (2022)	Ultrasound + MRI both required	Grad-CAM (image saliency)	None described
Abbasi et al. (2020)	Digital pathology WSI required	None	None described
Perera et al. (2021)	None (tabular data)	None	None described
ProstaPredict AI (This Study)	None (tabular clinical data)	SHAP global feature importance	Full web app, role access, patient records

Table 3.1: Comparative Analysis of Existing Systems

3.3.2 Overview of the Proposed System

ProstaPredict AI is an internet-enabled, explainable machine learning tool intended to aid physicians in making decisions regarding the probability of a malignancy being present in the prostate based on the clinical risk profile of patients and their symptomatology. The program is tailored to operate in resource-scarce clinical settings, including primary care settings and community health centers within Nigeria where sophisticated medical imaging technology is lacking but clinical tests can be carried out.

There are 23 clinical input parameters considered by the algorithm, involving quantitative measures (e.g., PSA value, BMI, prostate size, and age) and qualitative measures (e.g., DRE test result, family medical history, presence of urinary symptoms, etc.). This data goes into a pretrained preprocessing pipeline, followed by an XGBoost gradient boosting classifier, resulting in a score of the malignancy probability. There is a preset decision threshold for the malignant/benign classification equal to 0.50. In addition to the classification and probability itself, the model provides SHAP visualization of feature importance to help the physician grasp the impact of clinical factors on prediction.

The system runs using a dual-role based access control architecture. The administrator controls the system and registers the doctors, while the doctor registers the patients, input patient data, performs the predictions and looks into the prediction history of their patients. Authentication is performed on the basis of sessions and hashed credentials are stored in the server. Data persistence is done locally using an SQLite database.

The designed model addresses the core deficiency noted in the review of current systems by making use of no specialized imaging machinery, being easily deployable as a portable web application in clinics, offering in-built explainability to ensure clinical clarity, and having an additional layer of management of patients' records and predictions made by the system.

3.4 User Requirements Analysis

User requirement specifications were collected from the existing literature regarding AI-based clinical decision support systems, analysis of the existing process of diagnosing prostate cancer patients in Nigeria mentioned in Chapter Two, and the user stories generated in Sprint 1 within the Agile development methodology. Requirements are classified into two main types: functional requirements, detailing what the system should perform; and non-functional requirements, detailing the qualities of the system.

Functional Requirements

ID	Role	Requirement	Priority
FR01	All Users	System shall provide a secure login interface with username and password authentication	High
FR02	All Users	System shall redirect users to a role-appropriate dashboard upon successful login	High
FR03	Admin	System shall allow an administrator to register new doctor accounts	High
FR04	Admin	System shall allow an administrator to delete a doctor account and all associated records	Medium
FR05	Admin	System shall display system-wide statistics including total doctors, patients, and predictions	Medium
FR06	Doctor	System shall allow a doctor to register a new patient with demographic information	High
FR07	Doctor	System shall allow a doctor to enter 23 clinical and risk parameters for a registered patient	High
FR08	Doctor	System shall automatically pre-fill the Age field from the patient record and set it to read-only	Medium
FR09	Doctor	System shall run a biopsy outcome prediction when the clinical form is submitted	High
FR10	Doctor	System shall display malignancy probability, classification outcome, and confidence score	High
FR11	Doctor	System shall display SHAP feature importance visualisations alongside every prediction result	High

FR12	Doctor	System shall save every prediction to the database with all input features and model outputs	High
FR13	Doctor	System shall allow a doctor to view the full prediction history for their own patients	Medium
FR14	Doctor	System shall allow a doctor to view a longitudinal prediction history for a specific patient	Medium
FR15	Doctor	System shall allow a doctor to delete patient records and associated predictions	Low

Table 3.2: Functional requirements.

Non-Functional Requirements

ID	Category	Requirement	Priority
NF01	Security	All passwords shall be stored as Werkzeug-generated salted hashes; no plaintext credentials stored	High
NF02	Security	All routes requiring authentication shall enforce login_required; role violations shall redirect with an access-denied message	High
NF03	Security	A doctor shall only be able to access records belonging to their own account	High
NF04	Performance	Model and preprocessor shall be loaded once at startup; inference shall complete within 2 seconds	High
NF05	Usability	Interface shall be responsive and operable on standard desktop browsers without client-side installation	High
NF06	Usability	Result screen shall use colour-coded indicators (red for Malignant, green for Benign)	Medium
NF07	Reliability	System shall handle invalid or missing form inputs gracefully without crashing	High
NF08	Deployability	System shall deploy on a standard PC without requiring a separate database server or internet connection	High
NF09	Maintainability	ML model and preprocessor shall be stored as serialised Joblib files replaceable without modifying source code	Medium
NF10	Transparency	System shall display SHAP explainability visualisations with every prediction result	High

Table 3.3: Non-Functional Requirements.

3.5 Development Methodology

ProstaPredict AI was developed using a hybrid methodology combined both the Agile and Object-Oriented Development (OOD) paradigms. The Agile approach guided the entire development a

single researcher. Object-Oriented Development (OOD) was used to handle the system analysis and design stage of software development. The Cross-industry Standard Process for Data Mining (CRISP-DM) was used for the machine learning model development.

3.5.1 Justification for the Hybrid Development Approach

Agile was chosen because the development of a machine learning model is iterative and empirical; model performance can only be evaluated after training, and the results of one stage can often necessitate reconsidering decisions from the previous one. A rigid linear approach would not work here. The Object-Oriented Development method was chosen because the program needed an extensive analysis of its architecture, data interrelation, and user interactions before writing any code, making sure that all the database schema, access control logic, and components were designed on purpose. The CRISP-DM methodology was chosen because it provides a structured and systematic approach for developing ML applications, it emphasizes on data understanding and preparation which is essential when working with medical datasets.

3.6 Method of Data Collection

The dataset used in this study was the "Prostate Cancer Prediction Dataset" obtained from Kaggle, it is a publicly accessible database containing clinically oriented data tables corresponding to patients who had undergone prostate biopsy. It contains 27,945 records comprising demographic, clinical, and lifestyle variables.

Key attributes utilized include:

1. **Numerical:** Age, PSA_Level, BMI, Screening_Age, Prostate_Volume.
2. **Categorical:** Family_History, Race_African_Ancestry, DRE_Result, Difficulty_Urinating, Weak_Urine_Flow, Blood_in_Urine, Pelvic_Pain, Back_Pain, Erectile_Dysfunction, Exercise_Regularly, Healthy_Diet, Smoking_History, Alcohol_Consumption, Hypertension, Diabetes, Cholesterol_Level, Genetic_Risk_Factors, Previous_Cancer_History.
3. **Target Variable:** Biopsy_Result (Benign or Malignant).

Inasmuch as the current study is purely a conceptual one, the data was not collected during this work nor was any direct contact established between the author and either the patients or the clinical staff. The decision to utilize a publicly accessible database in the course of this research was based on specific conditions, namely the lack of data sharing protocol with the Nigerian health care facility and the difficulty in obtaining annotations for the biopsy data sets.

The data set was downloaded and saved on Google Drive so it could be used it when training the model in Google Colab. Before doing any form of data processing and modeling, the dataset was checked for its structure, including checking for missing values, duplicates, and consistency of

columns. There were no missing values that needed to be filled. Then the dataset was put through the process of leakage removal, followed by splitting it into train and test data sets, this is described further in chapter 4

The point should be made here that the utilization of a synthetic dataset that is not based on the Nigerian setting creates limitations when it comes to the generalizability of the trained model to the population at hand. These limitations will be elaborated further on in Chapter Five. However, the approach towards data collection undertaken in this study is suitable for a proof-of-concept system, in which case the value of the research is more about the system itself.

3.7 System Modeling Using UML

3.7.1 Use Case Analysis

Use case analysis shows the main actors interacting with ProstaPredict AI along with their actions and relationships between use cases. There are two main actors involved in the system, namely Administrator, who manages the entire process along with the clinical users, and Doctor, who is the main clinical user and is supposed to manage the patients and predict their status. Another important actor, XGBoost ML Model, represents the external part of the process.

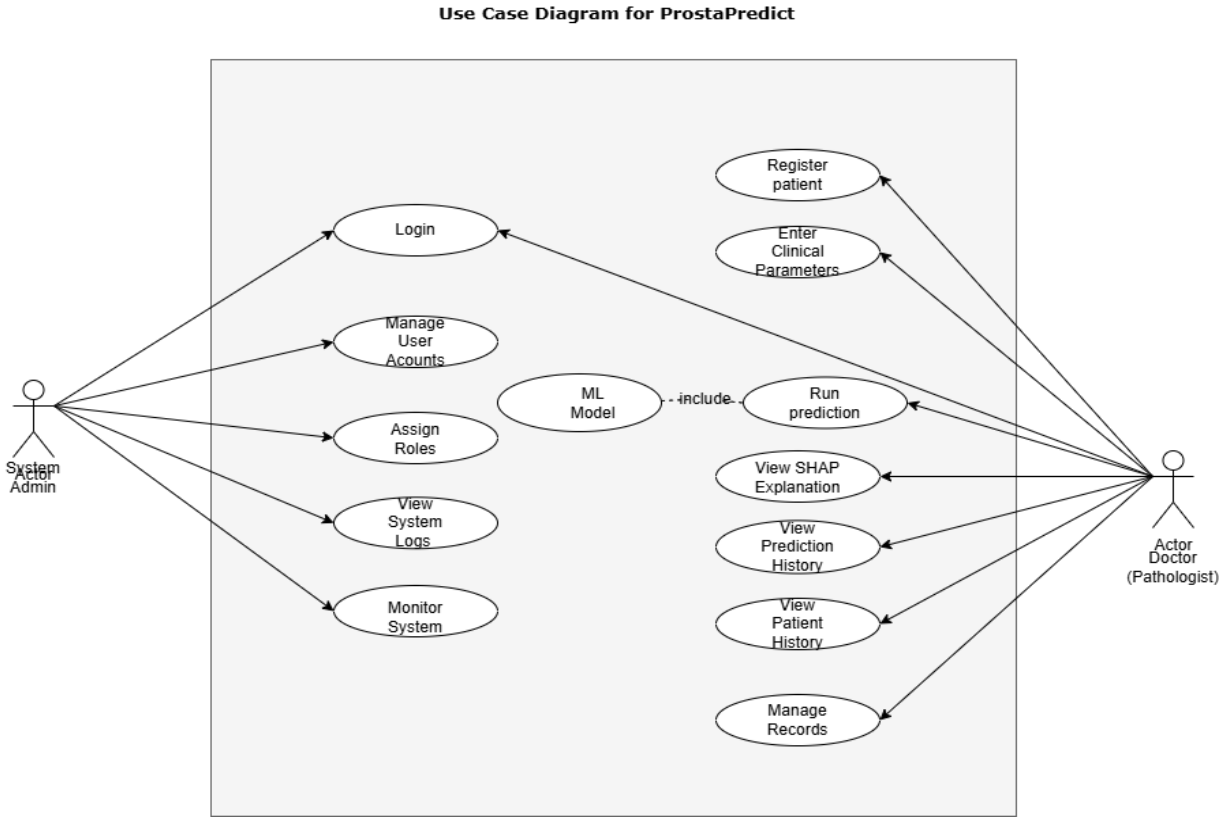


Figure 3.1: Use case diagram for ProstaPredict

Use case summary

Use Case	Actor	Description	Precondition
Login to System	Admin / Doctor	User provides username and password; system authenticates and redirects to role dashboard	User has a registered account
Register Doctor	Administrator	Admin creates a new doctor account with credentials and assigns doctor role	Admin is logged in
Delete Doctor	Administrator	Admin removes a doctor account and all associated patients and predictions	Admin is logged in; target is a doctor account
View System Statistics	Administrator	Admin views aggregate counts of doctors, patients, and predictions	Admin is logged in
Register Patient	Doctor	Doctor registers a new patient with demographic data linked to their account	Doctor is logged in
Enter Clinical Parameters	Doctor	Doctor fills in the 23-field prediction form for a registered patient	Patient exists; doctor is logged in
Run Biopsy Prediction	Doctor / ML Model	System preprocesses input, invokes XGBoost, applies threshold, returns result	Clinical form fully completed
View SHAP Explanation	Doctor	Doctor views SHAP feature importance bar chart and beeswarm plot on result screen	Prediction has been run
View Prediction History	Doctor / Admin	User views tabular log of all predictions; admin sees all, doctor sees own	User is logged in
View Patient History	Doctor	Doctor views longitudinal prediction history for a specific patient	Patient belongs to doctor; doctor is logged in
Delete Patient Record	Doctor	Doctor removes a patient and all linked predictions from the system	Patient belongs to doctor; doctor is logged in

Table 3.4: Use Case Descriptions

3.8 System Modeling Using UML

3.8.1 Context Diagram

In the context diagram, the system is depicted in an abstract form, where ProstaPredict AI is represented as a single process, and all other external systems interacting with the ProstaPredict AI are identified along with data flows.

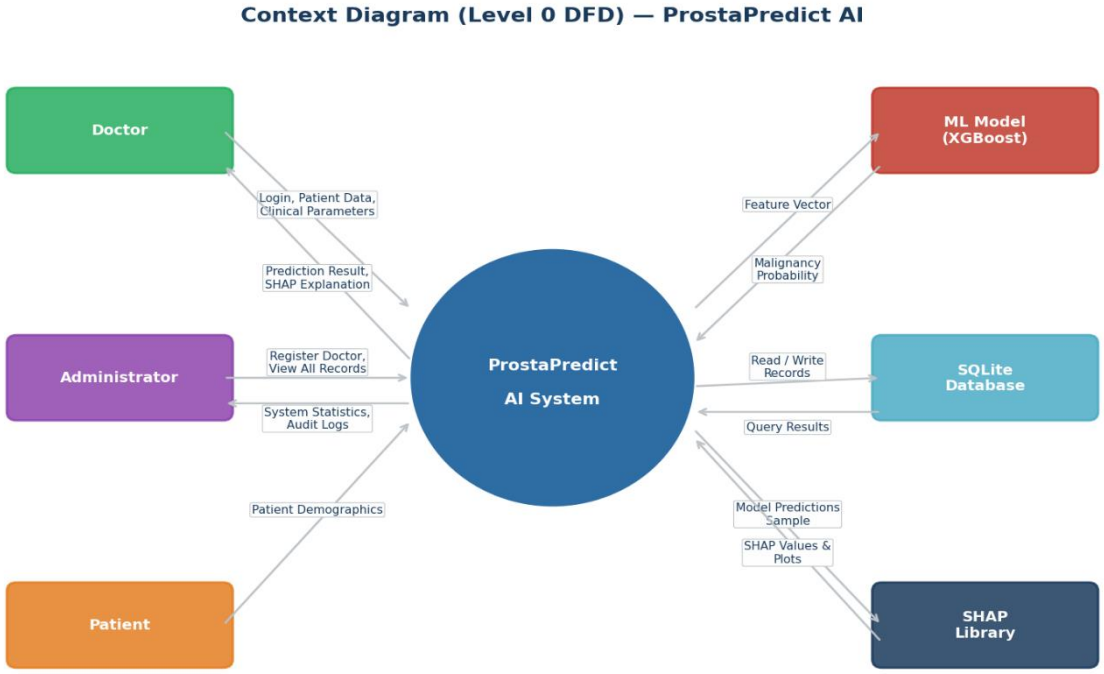


Figure 3.2: Context Diagram for ProstaPredict AI

The diagram highlights the six external entities involved, which include the following: The Doctor who inputs login details and clinical information and gets prediction outputs and explanations, the Administrator who inputs management information and gets system statistics, the Patient whose demographic information is recorded by the doctor, the XGBoost Machine Learning Model that inputs the preprocessed feature vector and outputs a malignancy probability, the SQLite Database that inputs record-write processes and outputs queries, and finally the SHAP Library that inputs a sample of model outputs and outputs explainability visualization data.

3.8.2 Entity Relationship Diagram

The entity relationship diagram models the data structures that persist in the SQLite database and the relationships between them. The system contains three entities: Users, Patients, and Predictions.

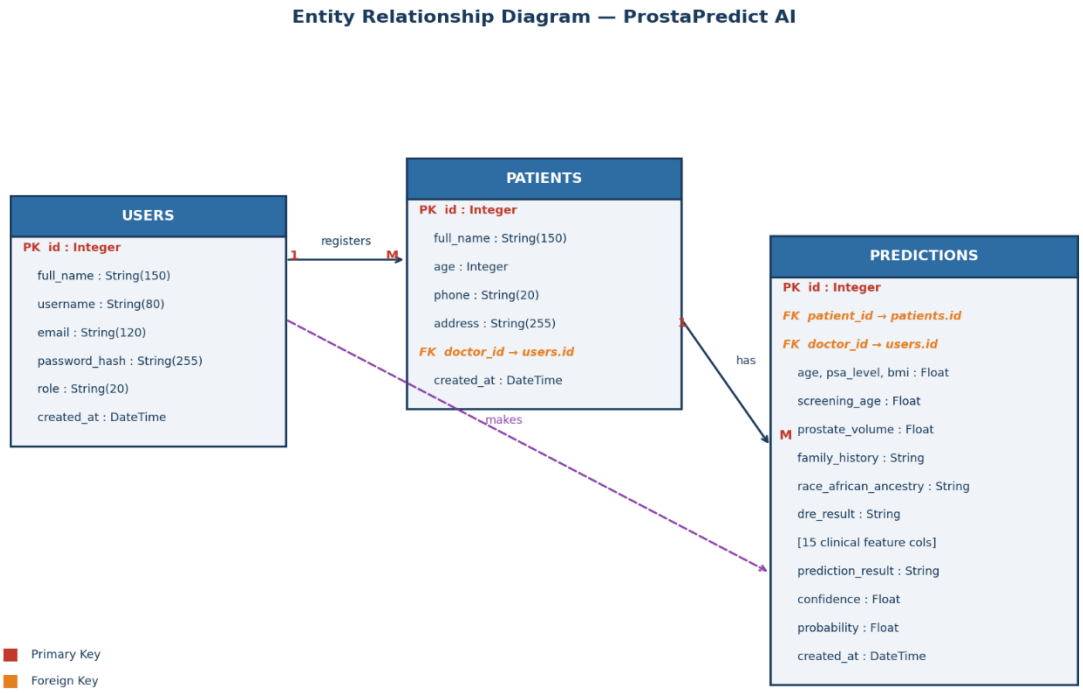


Figure 3.3: Entity Relationship Diagram for ProstaPredict AI

All the accounts in the system are stored in the Users entity, which is associated with Patients and Predictions in one-to-many relations, meaning that one physician can have several patients and predict their conditions several times. In the same way, the Patients entity is associated with the Predictions entity in a one-to-many relation for long-term tracking of patients across multiple sessions. All the data related to each prediction event are stored in the Predictions entity, including 23 clinical inputs, output values, and timestamp.

3.8.3 Activity Diagram

This shows the detailed process flow of the core use case in which a physician signs in to the system, chooses a patient, inputs relevant parameters, and receives the predicted value. The decision steps here include user login authentication and form validation.

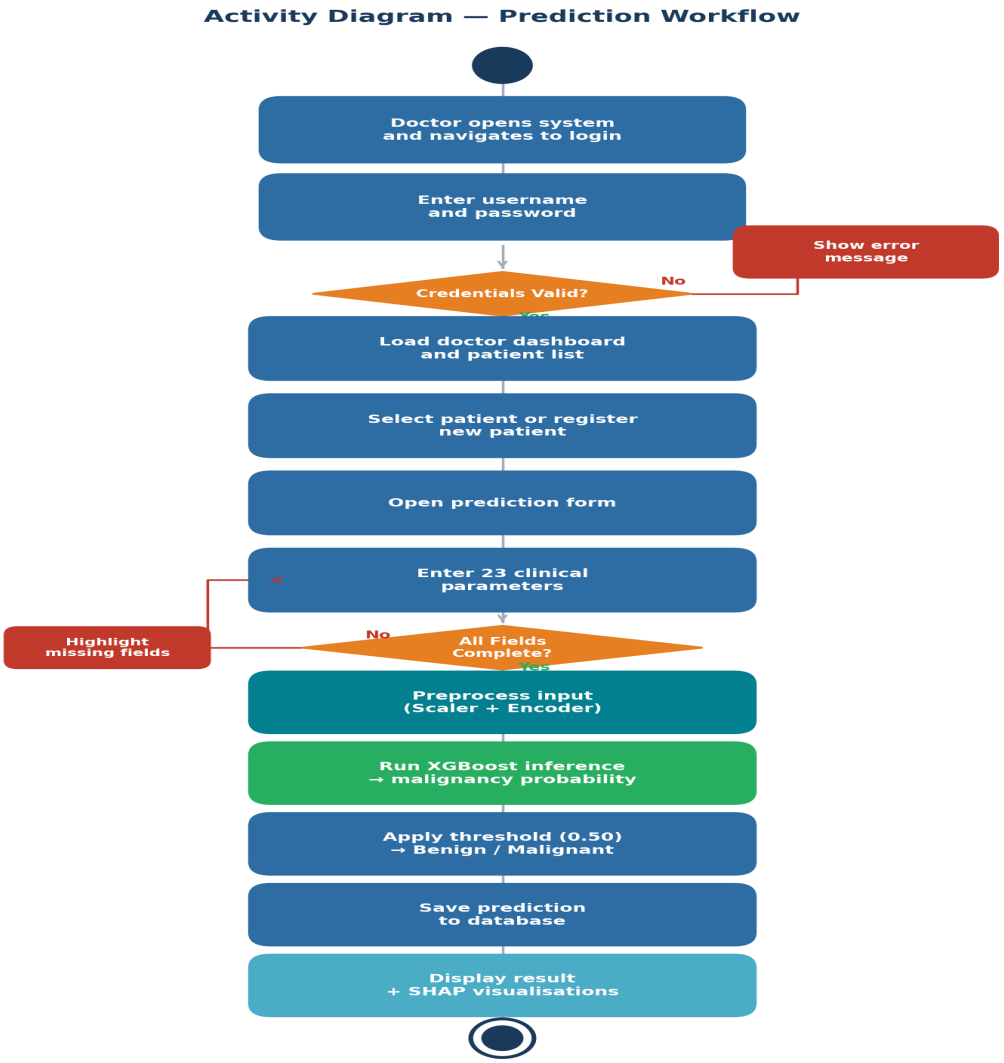


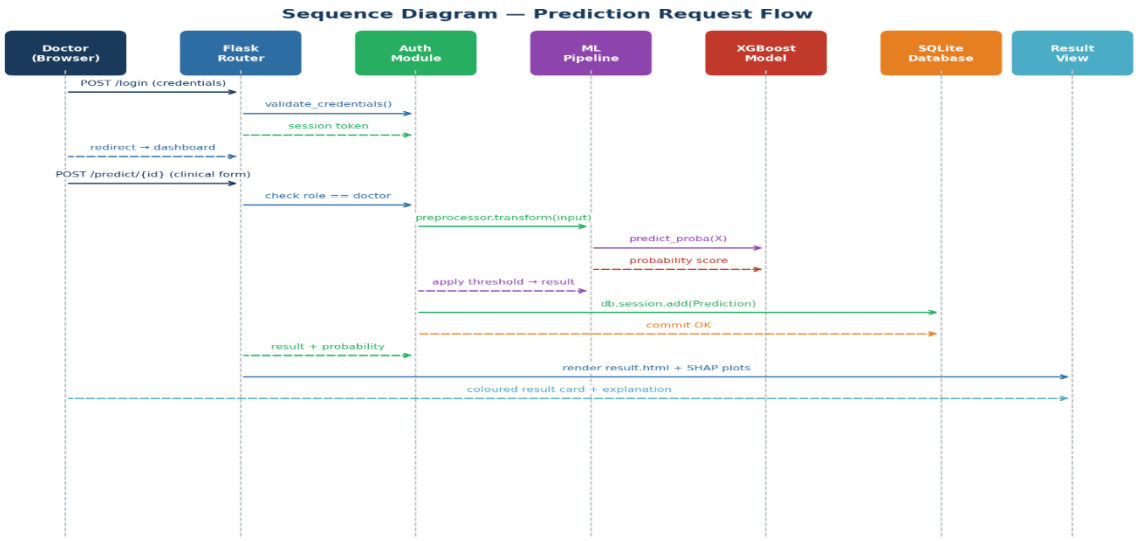
Figure 3.4: Activity Diagram for Prediction Workflow

The process starts when the doctor tries to log in. If he does not succeed, then an error message pops up, and the flow goes back to the login page. In case of successful login, the doctor either selects a patient or creates one, opens the prediction form, and fills in all the necessary parameters. In case some fields are empty, the system will mark them out. When the user completes the form, pre-processing occurs, followed by XGBoost prediction, decision-making based on threshold, prediction database insertion, and output of result visualization.

3.8.4 Sequence Diagram

Figure 1: Temporal order of the messages sent between system components throughout the course of a prediction request, from the time the doctor submits the form until the result is displayed. The system components used in this case are: the Doctor (Browser), Authentication Module, ML Pipeline, XGBoost Model, SQLite Database, and Result View template.

Figure 3.5: Sequence Diagram; Prediction Request Flow



The workflow starts when the doctor makes a POST request with the prediction form. The router confirms the user's permission, the ML pipeline processes the form data to prepare features, and passes them to the XGBoost model to receive probability estimates of malignancy. After applying the decision threshold, the router saves the complete prediction record to the database and creates an HTML page with predictions, probabilities, confidence levels, and SHAP plots.

3.9 System Architecture Overview

The ProstaPredict AI software is constructed on a tiered web application design, comprising presentation, application logic, and data persistence layers. The presentation layer comprises of HTML web pages with the help of the templating engine called Jinja2 along with the use of the Bootstrap 5 framework that makes the site accessible to be run on any general computer or mobile devices without requiring specialized equipment. Application logic layer includes the use of the Flask framework for Python programming language, which will take care of HTTP routing,

sessions management, and calling of the trained machine learning model. Data Persistence layer is done via the use of a relational SQLite database along with SQLAlchemy ORM.

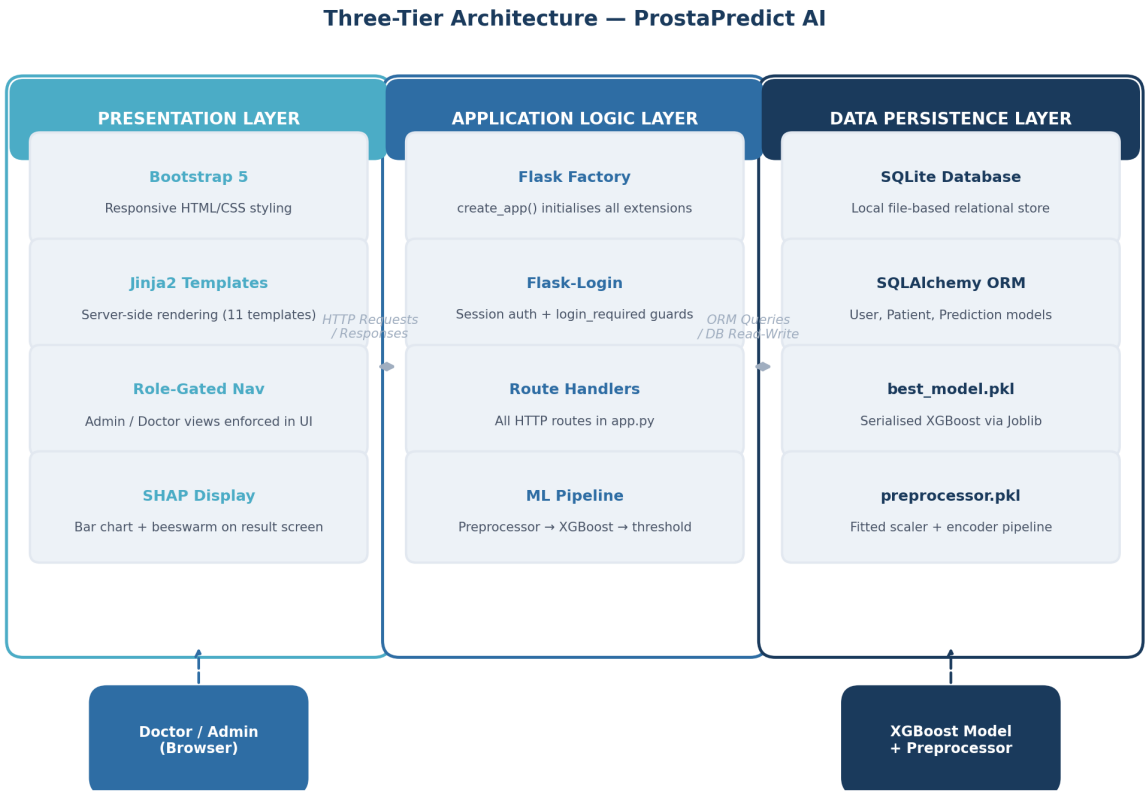


Figure 3.6: ProstaPredict System Architecture

The system architecture follows the application factory design of Flask, with create_app() method that initializes all extensions such as SQLAlchemy and Flask-Login as well as registering all route methods before deployment. The trained model and the pipeline used for preprocessing are both loaded only once upon starting using Joblib to save time from frequent access to disk. The architecture was purposely designed to be lightweight; all software stacks of the application could run even on less powerful machines.

CHAPTER FOUR

SYSTEM IMPLEMENTATION AND TESTING

4.1 Introduction

This chapter describes in detail the complete implementation, testing, and evaluation of ProstaPredict AI. This includes an explanation of the development environment, data preparation and preprocessing processes, machine learning model architecture and training, the modular design of the web application, and the testing and evaluation processes employed for the system. Finally, this chapter, present the results obtained through the use of the system.

4.2 Development Environment and Tools

4.2.1 Programming Language and Libraries

The entire system was coded in Python 3.10, selected based on its dominance within the fields of both machine learning and web development. The machine learning pipeline was initially tested on a Jupyter Notebook platform (Google Colab), providing real-time execution, visualization, and integration with Google Drive for data/model storage. The final web app was then built on a local environment using the Flask framework. Table 4.1 shows all libraries and frameworks employed in this project.

Library / Framework	Version	Role
Flask	3.1.3	Web application framework — routing, templating, session management
Flask-SQLAlchemy	3.1.1	ORM for SQLite database interaction
Flask-Login	0.6.3	Session-based user authentication and role management
scikit-learn	1.6.1	Preprocessing pipeline (StandardScaler, OneHotEncoder, ColumnTransformer), model evaluation
XGBoost	3.2.0	Gradient boosted tree classifier — deployed prediction model
imbalanced-learn	0.14.1	SMOTE for class imbalance oversampling (evaluated, not deployed)
SHAP	0.49.1	SHapley Additive exPlanations — global feature importance and local feature impact
Pandas	2.3.3	Data loading, cleaning, and manipulation
NumPy	2.2.6	Numerical operations and array handling
Matplotlib / Seaborn	latest	Visualisation of training metrics and SHAP plots
Joblib	1.5.3	Serialisation of trained model and preprocessing pipeline
Bootstrap 5	5 (CDN)	Responsive front-end styling
Jinja2	Bundled	Server-side HTML templating

Werkzeug	Bundled	Password hashing for secure credential storage

Table 4.1: Libraries and Frameworks Used in System Development

4.2.2 Development Platform and Hardware Requirements

The training process was done using Google Colab, providing access to a python environment running in the cloud, where GPUs were available, hence overcoming any limitations posed by local hardware while training the model. Flask was coded and tested using a regular Windows 11 laptop with 8 GB RAM and a 64-bit processor. Inference did not require any GPU, as XGBoost works with CPU prediction having very little to no latency.

In order to use this system, all that is needed is a computer with the capability to run Python 3.10+ along with a browser to access the web front end, with a total storage requirement of about 50MB for the software, the SQLite database, and the serialised models. Once the system is set up, an internet connection will not be necessary.

4.3 Dataset Description and Preprocessing

This dataset is an ordered set of tabulated clinical data of patients who have undergone a prostate biopsy. The target column is the output from the test which can be either benign (labelled 0) or malignant (labelled 1). The raw data was initially imported as a CSV file and then preprocessed in the following manner.

Data Leakage Elimination: In the initial dataset, there were various columns containing data that will only become available after the patient undergoes biopsy diagnostics such as Cancer Stage, Treatment Suggested, Survival after Five Years, Need for Follow-Up, and Detection of Cancer. Using this data in training will make the model learn from the effects of biopsy diagnostics instead of pre-biopsy patient conditions, giving a biased estimate on the performance of the machine learning algorithm. These five data columns and Patient ID were eliminated before training. Further discussion can be found in Section 4.6.2 about the leakage experiment findings.

Feature classification: Once the process of leakage elimination is completed, there are 23 features present in the clean dataset. Five features are numerical in nature and include Age, PSA level, BMI, Screening age, and Prostate volume. Rest 18 features are categorical features that encompass Family History, African ancestry, DRE test results, symptoms, lifestyle, medical comorbidities, and Genetic Risk Factor.

Feature Group	Features
Numerical (5)	Age, PSA Level, BMI, Screening Age, Prostate Volume

Clinical / Risk (5)	Family History, Race (African Ancestry), DRE Result, Genetic Risk, Previous Cancer History
Urinary Symptoms (3)	Difficulty Urinating, Weak Urine Flow, Blood in Urine
Pain Symptoms (3)	Pelvic Pain, Back Pain, Erectile Dysfunction
Lifestyle Factors (4)	Exercise Regularly, Healthy Diet, Smoking History, Alcohol Consumption
Comorbidities (3)	Hypertension, Diabetes, Cholesterol Level

Table 4.2: Input Feature Groups After Leakage Removal

Preprocessing Pipeline: The preprocessing pipeline in scikit-learn uses the ColumnTransformer function to transform both numerical and categorical features using separate pipelines. The numerical features were transformed using the StandardScaler method, where the features are centered around zero mean and scaled to unit variance. This is important because otherwise, features that are on a higher scale than others, such as PSA Level, would have more influence on the model. Categorical features were transformed using the OneHotEncoder with the handle_unknown parameter set to “ignore”.

Splitting Training and Testing Sets: The data set was split into an 80% training set and a 20% testing set using a stratified method with a fixed random state of 42. This ensures that the distribution of classes is the same in both subsets.

Handling Class Imbalance: There was a case of class imbalance in the dataset with respect to the classes, that is, benign and malignant. Two approaches were considered. One is SMOTE (synthetic minority oversampling technique). SMOTE oversamples the training dataset with artificial malignant data. Cost sensitivity was another approach that considers the weights assigned in favor of some classes over others during the training process. The weights for XGBoost, for example, were scaled with scale_pos_weight, while for other algorithms such as logistic regression and random forest, they were set using the parameter class_weight="balanced". The approach of cost sensitivity was chosen for deployment.

4.4 Model Architecture and Training

A total of three supervised classification models were used for training in order to find the most effective one for implementation; these include Logistic Regression, Random Forest, and XGBoost. All three have been trained using the same leakage-free data set.

The Logistic Regression model was fit using a maximum number of iterations of 2,000 and setting class_weight equal to balanced. It is a simple linear model and can therefore be used as a reference for appreciating the gains made from the more complex models.

The Random Forest algorithm consisted of an ensemble of 300 decision trees, and the parameter class_weight was set as balanced for balancing classes. The Random Forest algorithm uses bagging, which means combining multiple models to improve the predictive power, and was considered more capable of capturing non-linear features.

XGBoost was configured as a gradient boosting machine with 400 trees, a learning rate of 0.05, tree depth limit of 4, and subsampling rate of 0.9 for both rows and columns. The parameter 'scale_pos_weight' was estimated using the distribution of classes in the training dataset in order to assign penalties to the misclassification of malignant samples. XGBoost was chosen as the final classifier due to its good balance of performance across multiple criteria. The serialized model and preprocessor were saved as best_model.pkl and preprocessor.pkl respectively and imported into the Flask application.Table 4.3 summarises the key hyperparameters for each model.

Model	Key Hyperparameters
Logistic Regression	max_iter=2000, class_weight="balanced", random_state=42
Random Forest	n_estimators=300, class_weight="balanced", random_state=42
XGBoost (Deployed)	n_estimators=400, learning_rate=0.05, max_depth=4, subsample=0.9, colsample_bytree=0.9, scale_pos_weight=neg/pos, eval_metric="logloss", random_state=42

Table 4.3: Model Hyperparameter Configuration

4.4.1 Explainable AI Integration

Another fundamental aspect of the model is the explainability of its outputs in terms that can be understood by the physicians. The SHAP (SHapley Additive exPlanations) technique is used to provide explainability of the results through a game theory approach where each of the inputs receives a SHAP value reflecting its effect on the output result.

In the case of XGBoost, the TreeExplainer class was applied to compute SHAP values efficiently for decision tree models, returning exact SHAP values without approximation. The SHAP values have been calculated for 100 randomly selected test samples to construct two global explainability plots: a bar plot depicting average SHAP value by feature (feature importance), and a summary plot illustrating the distribution of SHAP values (magnitude and direction of the feature influence). The SHAP plots are depicted together with the result of the prediction on the screen, thus enabling the doctor to determine the importance of the clinical features contributing to the prediction of the

current patient. If the SHAP value is positive, then the feature tends to push the prediction towards Malignant, otherwise towards Benign.

It is worth mentioning that SHAP visualizations being shown here are obtained using a representative sample for testing purposes and capture the overall behavior of the model globally. This decision was taken to make sure that the visualization will render reliably on the web page, without having to calculate SHAPs per each patient separately because of the high computational cost involved.

4.5 System Implementation and Modules

The ProstaPredict AI system uses Flask framework and was designed as a web application employing the application factory design pattern. It consists of the following core components:

app.py - Application Factory & Routes: The create_app() method initializes all Flask extensions (SQLAlchemy, Flask-Login), loads the serialized machine learning model and the preprocessor from the disk via Joblib, creates the database tables if missing, and seeds the default admin user. Route handlers for all HTTP requests are defined here.

models.py - Database Models: Contains the definition for three classes using the SQLAlchemy ORM (Object Relational Mapper) that correspond to three database tables; they include the User, Patient, and Prediction models. The Prediction class contains all the 23 features in addition to the result, probability, and confidence values.

config.py - Configuration: Contains the configuration settings, including the secret key used in Flask, URI for the database (which is the local SQLite database), and others.

extensions.py - Extension Instances: Keeps the instance of db which refers to SQLAlchemy and login_manager which refers to Flask-Login to prevent circular imports from app.py and models.py.

templates/ - Jinja2 HTML Templates: The eleven HTML templates use a common layout named base.html, which gives navigation bars and the Bootstrap style sheet. These are predict.html (clinical information form with 23 fields), result.html (prediction outcome along with SHAP graphs), history.html (audit log of predictions), and patient_history.html (patient's record history).

static/ - Static Files: Holds the Bootstrap bundle, custom CSS, hero image, and pre-rendered SHAP graph images such as shap_bar.png and shap_summary.png.

4.5.1 Algorithms and Techniques Used

The following algorithms and analytical techniques were employed across the data analysis and model development phases of this study:

Algorithm / Technique	Library	Purpose
StandardScaler	scikit-learn	Normalise numerical features to zero mean and unit variance
OneHotEncoder	scikit-learn	Encode categorical features as binary indicator columns
ColumnTransformer	scikit-learn	Apply different preprocessing steps to numerical and categorical columns simultaneously
SMOTE	imbalanced-learn	Synthetic oversampling of minority class (evaluated; not deployed in final model)
Logistic Regression	scikit-learn	Baseline linear classifier with class_weight=balanced for cost-sensitive training
Random Forest	scikit-learn	Ensemble classifier with class_weight=balanced for cost-sensitive training
XGBoost	xgboost	Gradient boosted tree classifier with scale_pos_weight; selected as deployed model
TreeExplainer (SHAP)	shap	Compute exact SHAP values for tree-based models; generate global feature importance plots
Joblib Serialisation	joblib	Persist trained model and preprocessing pipeline for deployment-time loading
Stratified Train-Test Split	scikit-learn	Split dataset preserving class distribution; 80% train / 20% test, seed=42
ROC-AUC Evaluation	scikit-learn	Threshold-independent performance metric; primary criterion for model comparison

Table 4.4: Algorithms and Techniques Used in System

4.6 Testing and Evaluation

4.6.1 Evaluation Metrics

The performance of each model was assessed using five standard metrics for binary classification problems. The Accuracy refers to the ratio of the total number of predictions that were right out of all the predictions. Precision is the ratio of all those cases that were predicted as Malignant which indeed belong to the class Malignant out of all the cases predicted as Malignant. Recall (or

Sensitivity) is the ratio of all those truly Malignant cases that are detected as malignant by the models. F1 Score is the harmonic mean of Precision and Recall. It helps in measuring both at once by combining both into one measure. ROC-AUC (Area Under Curve) refers to the area under the curve.

4.6.2 System Testing

Unit testing was done on the primary code which includes the authentication code (such as login, password hashing, and checking for roles), the prediction route code (checking whether the data submitted in the form is correctly processed and passed to the model), and code related to working with the database (checking whether the predictions are successfully stored in the database with correct field values). The integration test was run to confirm that the entire process works correctly when the Benign/Malignant predictions happen.

Boundary testing was performed on the prediction form in order to test that: (a) missing values generate a validation error but do not cause any crash; (b) the age input is properly inherited from the patient database and cannot be modified manually; and (c) the threshold value of 0.50 works as expected in classification. The role isolation testing was done by trying to use administrator or doctor routes when logged in using a different role, and the redirection worked as expected in all cases.

4.6.3 User Interface Testing and Walkthrough

Interface testing across the entire clinical workflow was carried out for both users. The full clinical process for a doctor is registration of a patient, filling of the 23-field prediction form with different parameter combinations, checking of the prediction results and SHAPs visualization, and checking whether the predictions were saved in the general history and also in patient's history tabs. Both delete functions for patient deletion and for prediction entry deletion were checked and cascaded deletion function was verified.

For the Administrator role, the creation of a new account for doctors and the removal of an existing one was also checked, as well as the cascading action on patients and predictions associated with the removed doctor account. The system statistics appearing on the administrator interface were compared with the number of records in the database. The walkthrough for all interface screens will be provided in screenshots in section 4.7.1.

4.7 Results Presentation and Analysis

4.7.1 Output Samples and Interface Screens

The below illustrations capture some of the main screens in the ProstaPredict AI system's interface, during a common clinical session. These screens are captured while the application was in use. Figure 4.1 shows the Home Screen, which acts as the public interface and describes what the application does, as well as providing an entry to log into the system.

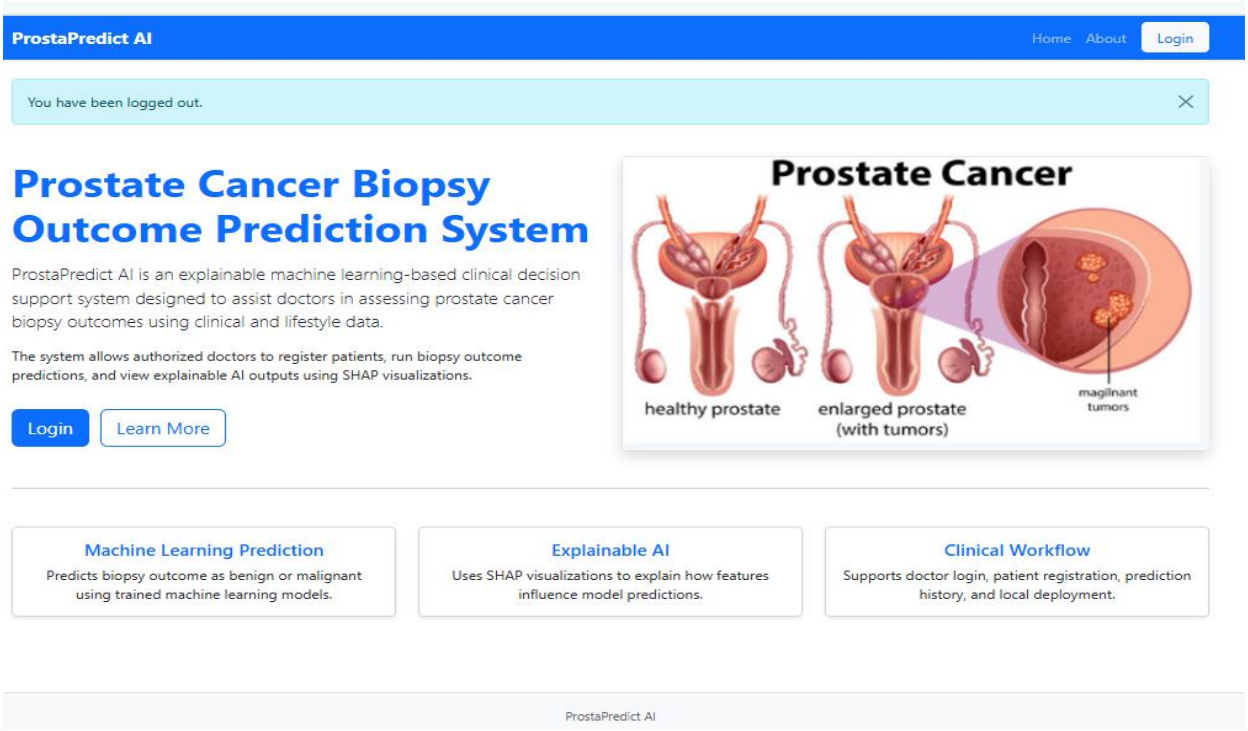


Figure 4.1: ProstaPredict AI Home Screen

Figure 4.2 shows the Login Screen, which provides role-based authentication for both Administrators and Doctors. The password field includes a show/hide toggle to reduce entry errors.

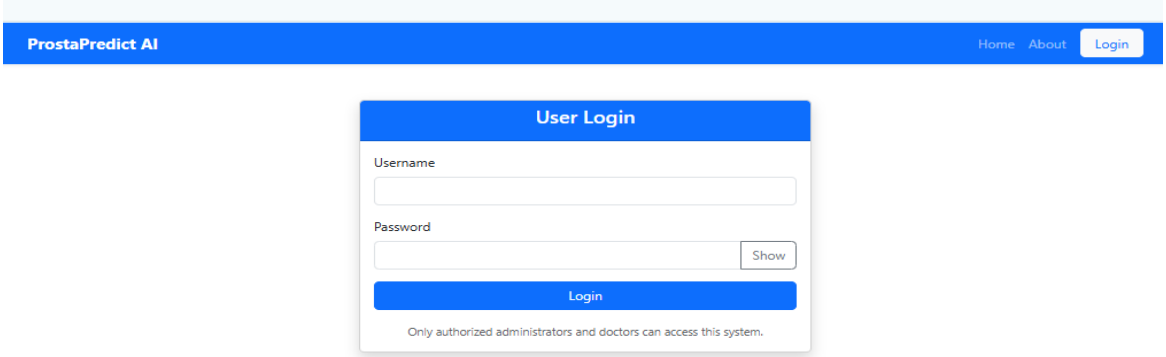


Figure 4.2: Login Screen with Role-Based Authentication

Figure 4.3 shows the Administrator Dashboard, which displays system-wide statistics and provides controls for registering and deleting doctor accounts.

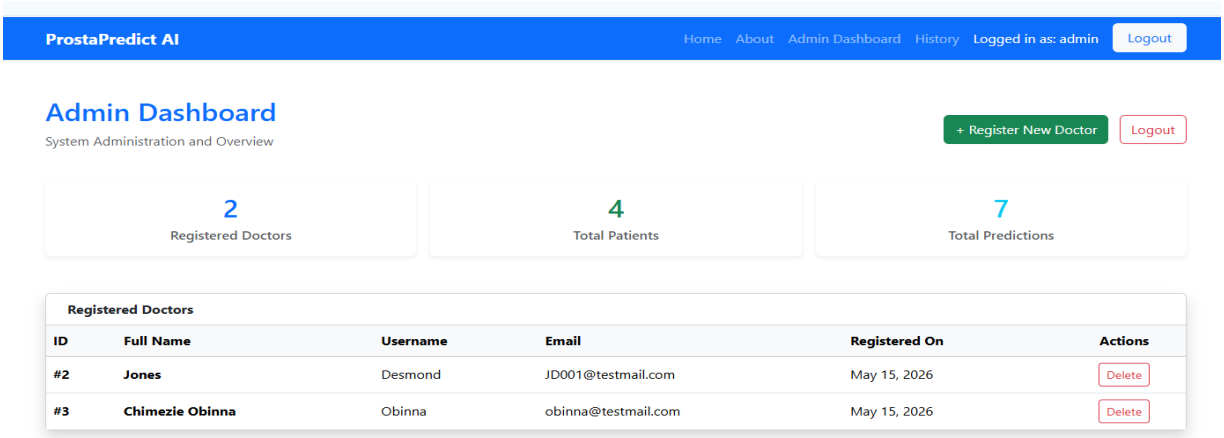


Figure 4.3: Administrator Dashboard — System Statistics and Doctor Management

Figure 4.4 shows the Doctor Dashboard, which lists all patients registered under the logged-in clinician, with action buttons to run a prediction, view history, or delete the patient record.

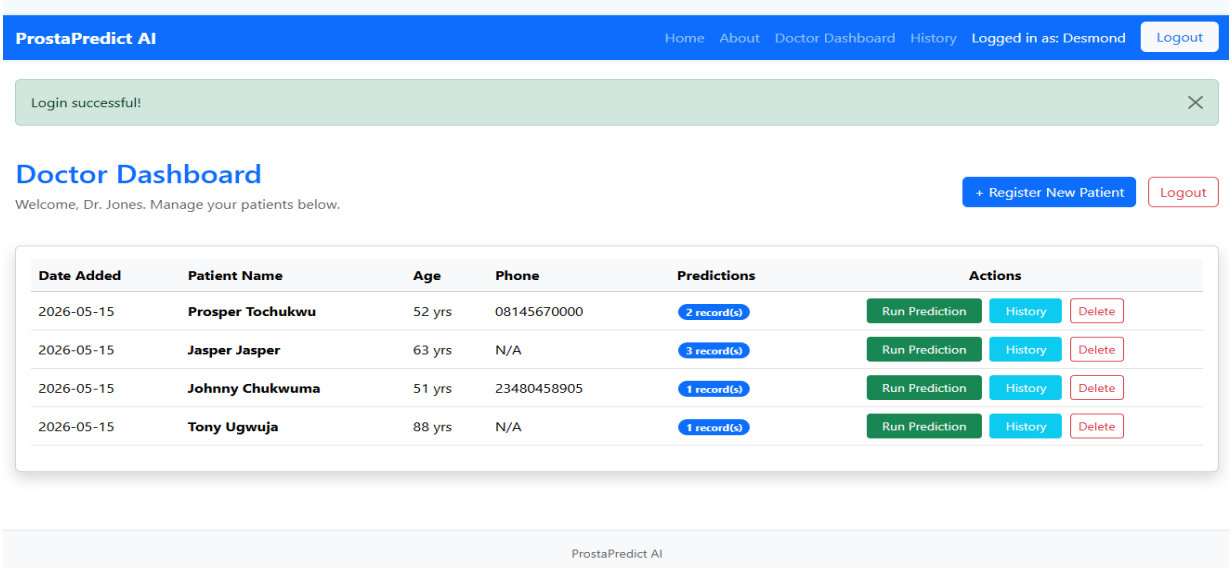


Figure 4.4: Doctor Dashboard — Patient Management Interface

Figure 4.5 shows the Patient Registration screen, through which a doctor enrols a new patient by entering their demographic details.

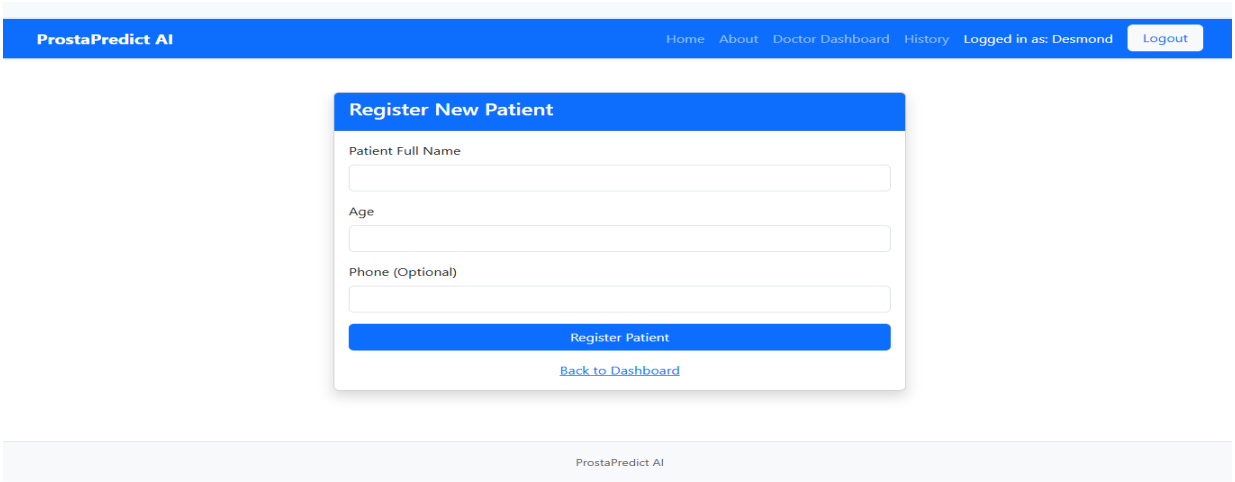


Figure 4.5: Patient Registration Screen

Figure 4.6 shows the Clinical Data Entry (Prediction) Form, which is the main tool for data entry. It includes a section for numerical and one for categorical variables, where the value for Age

variable is automatically inserted as read only.

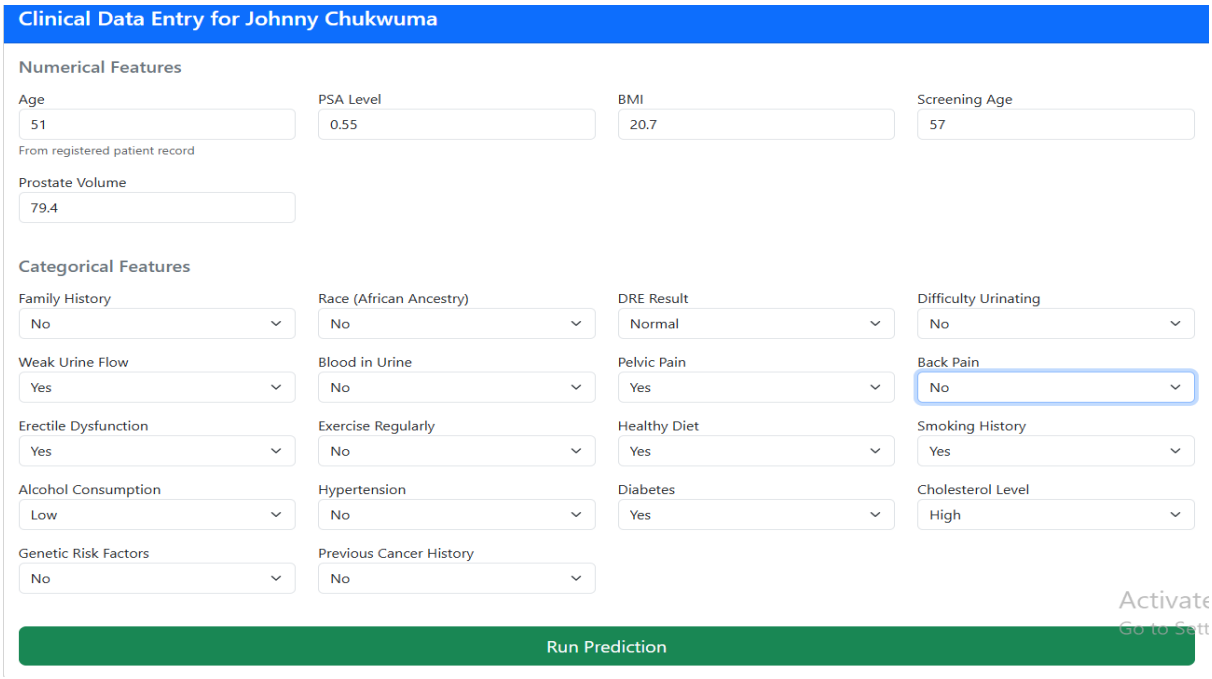


Figure 4.6: Clinical Data Entry Form — Numerical and Categorical Input Fields

The Prediction Result window for Malignant and Benign outputs are depicted in Figures 4.7 and 4.8 respectively. These outputs include the color-coded feedback, probability of malignancy, confidence score and two visualization results obtained from SHAP explainability models on the

right-hand side of the image.

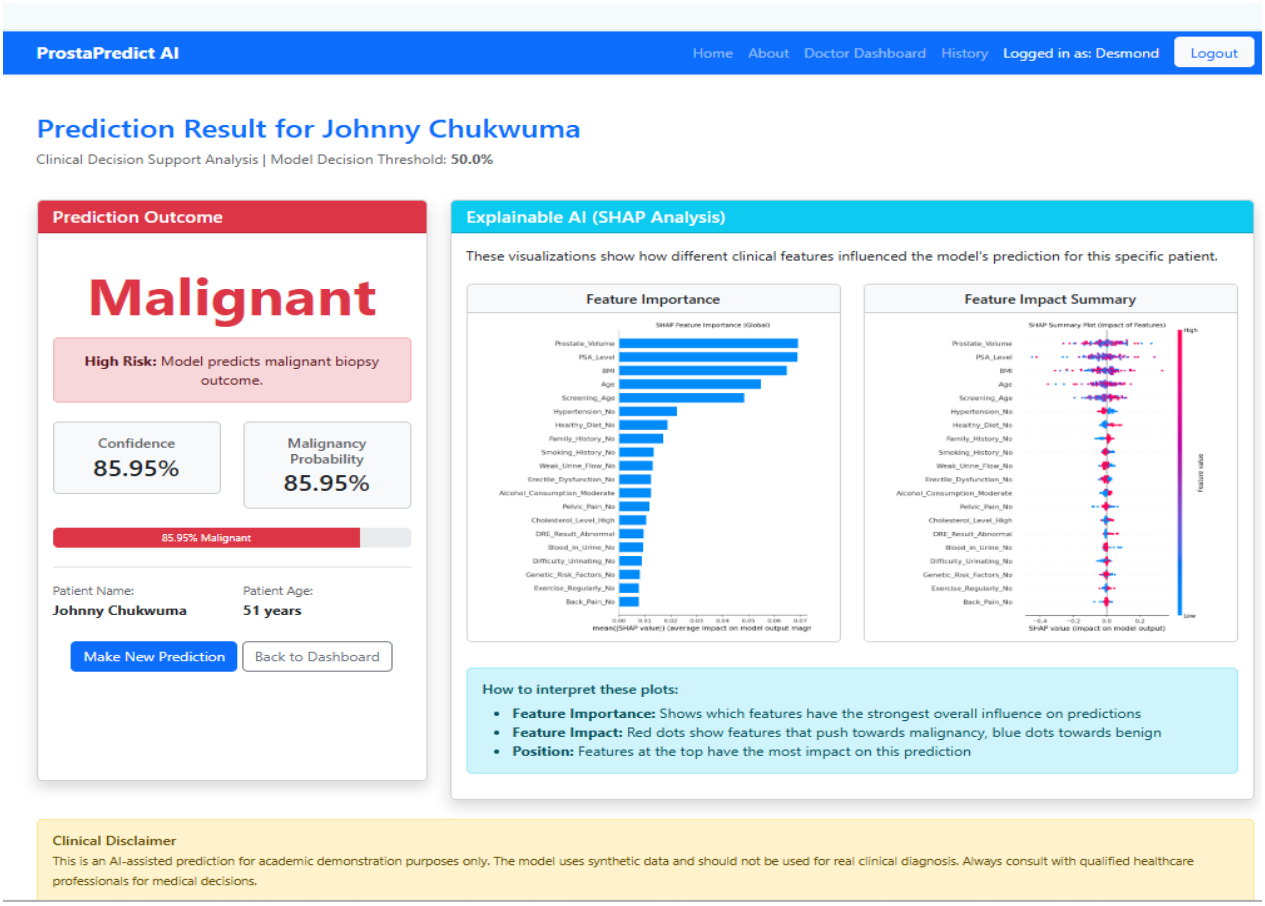


Figure 4.7: Prediction Result Screen — Malignant Outcome (Confidence: 85.95%)

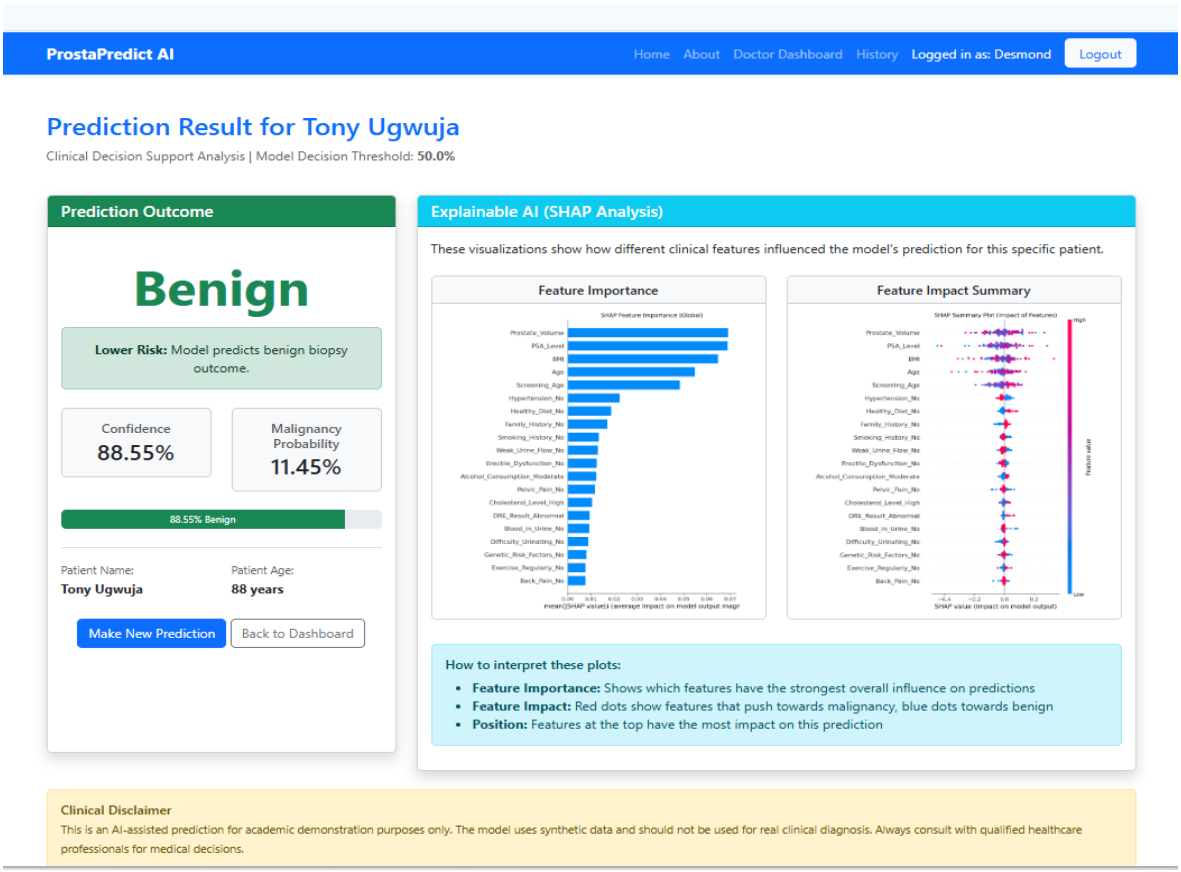


Figure 4.8: Prediction Result Screen — Benign Outcome (Confidence: 88.55%)

Figure 4.9 shows the Prediction History view, which provides a complete audit log of all prediction events with colour-coded outcome badges and confidence scores.

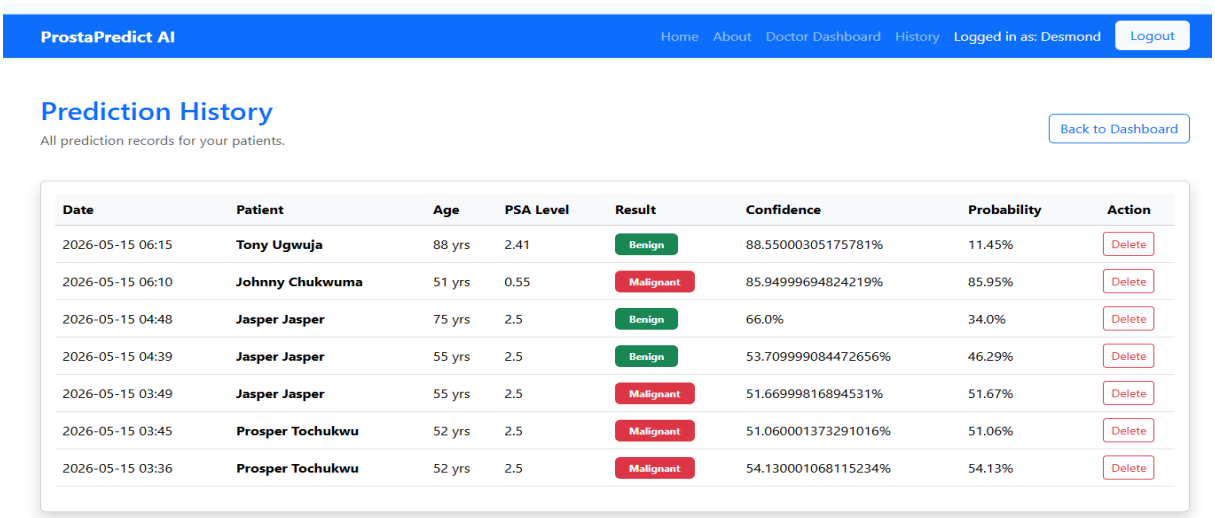


Figure 4.9: Prediction History — Full Audit Log

Figure 4.10 shows the Patient-Specific History view, which presents a longitudinal record of all predictions made for a single patient, supporting clinical monitoring over successive consultations.

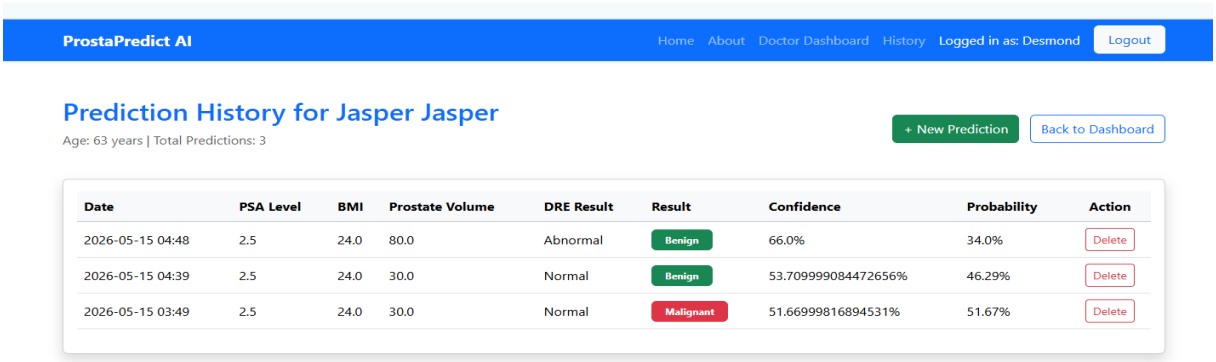


Figure 4.10: Patient-Specific Prediction History View

4.7.2 Discussion of Results and System Performance

Model Performance: Table 4.4 presents the full performance comparison of the three models trained under the leakage-aware, class-weighted condition.

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.5099	0.3204	0.4806	0.3846	0.5008
Random Forest	0.6987	0.5000	0.0018	0.0036	0.4958
XGBoost (Deployed)	0.5305	0.3450	0.4044	0.3723	0.4891
Always-Benign Baseline	~0.70	0.00	0.00	0.00	N/A

Table 4.5: Model Performance Comparison — Leakage-Aware, Class-Weighted Training

Figure 4.11 presents a grouped bar chart comparing all five metrics across the three models, providing a visual overview of the performance profile of each.

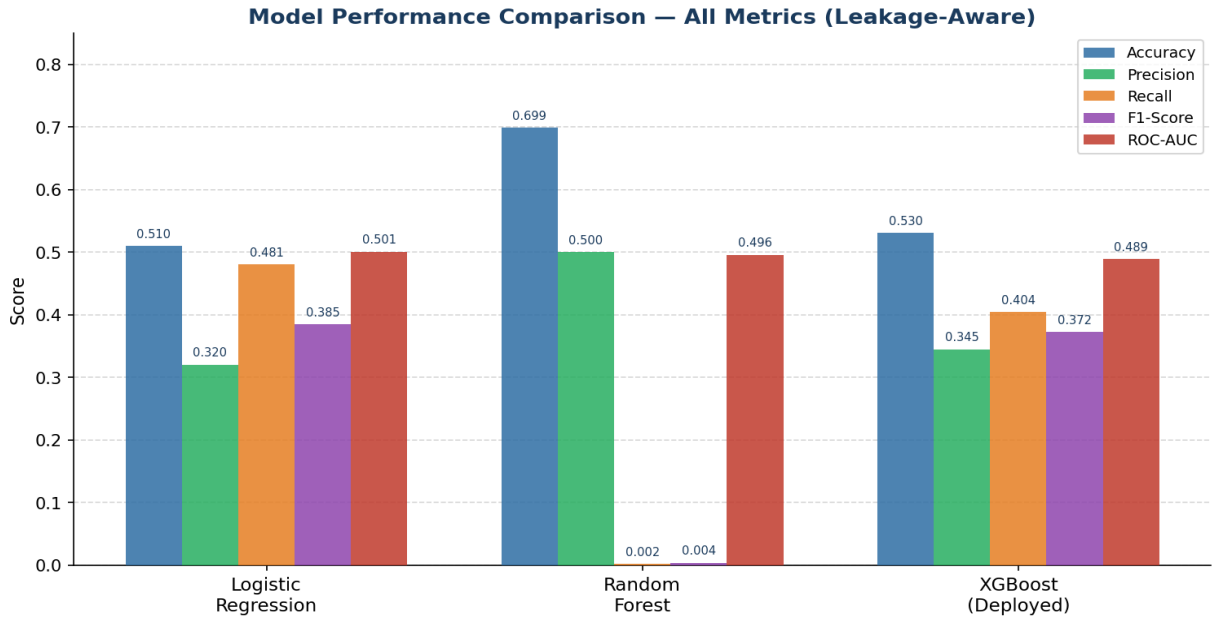


Figure 4.11: Model Performance Comparison — All Metrics (Leakage-Aware, Class-Weighted)

Confusion Matrixes: As illustrated in figure 4.12, the confusion matrixes are drawn out for all three models. This matrix classifies the output into True Positive (TP - Malignant prediction correct), True Negative (TN – Benign prediction correct), False Positive (FP – Benign classified as Malignant), and False Negative (FN – Malignant classified as Benign).

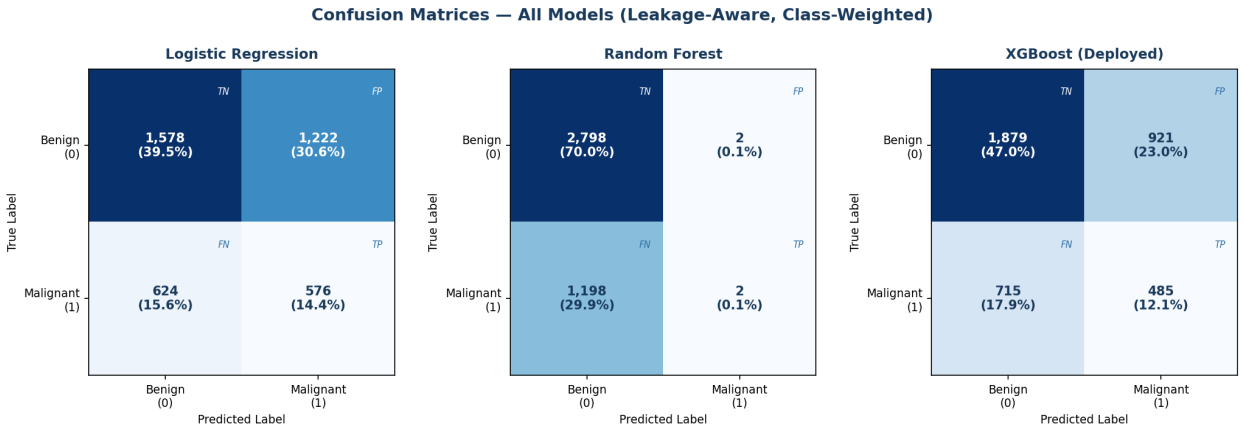


Figure 4.12: Confusion Matrices — All Three Models (Leakage-Aware, Class-Weighted)

Note that in the case of the Random Forest confusion matrix, the model has obtained high accuracy (0.6987) based on prediction of Benign class for the vast majority of the instances, resulting in 2 true positives. Such a failure mode – which would effectively make this tool useless in a clinical setting – is exactly what we wanted the model to catch during our evaluation. In terms of identifying malignant instances, Logistic Regression model has performed the best of all the three

models (Recall = 0.4806). On the other hand, the performance of XGBoost is a little more balanced between recall and precision, as it managed to correctly identify 485 malignant cases, but not produce as many false positives as Logistic Regression model. This is why XGBoost was selected for deployment.

ROC Curves: Figure 4.13 depicts the ROC curves of the three models with the random classifier line shown. The curves verify that the performance of all three models is very close to the random classifier performance (AUC = ~0.50).

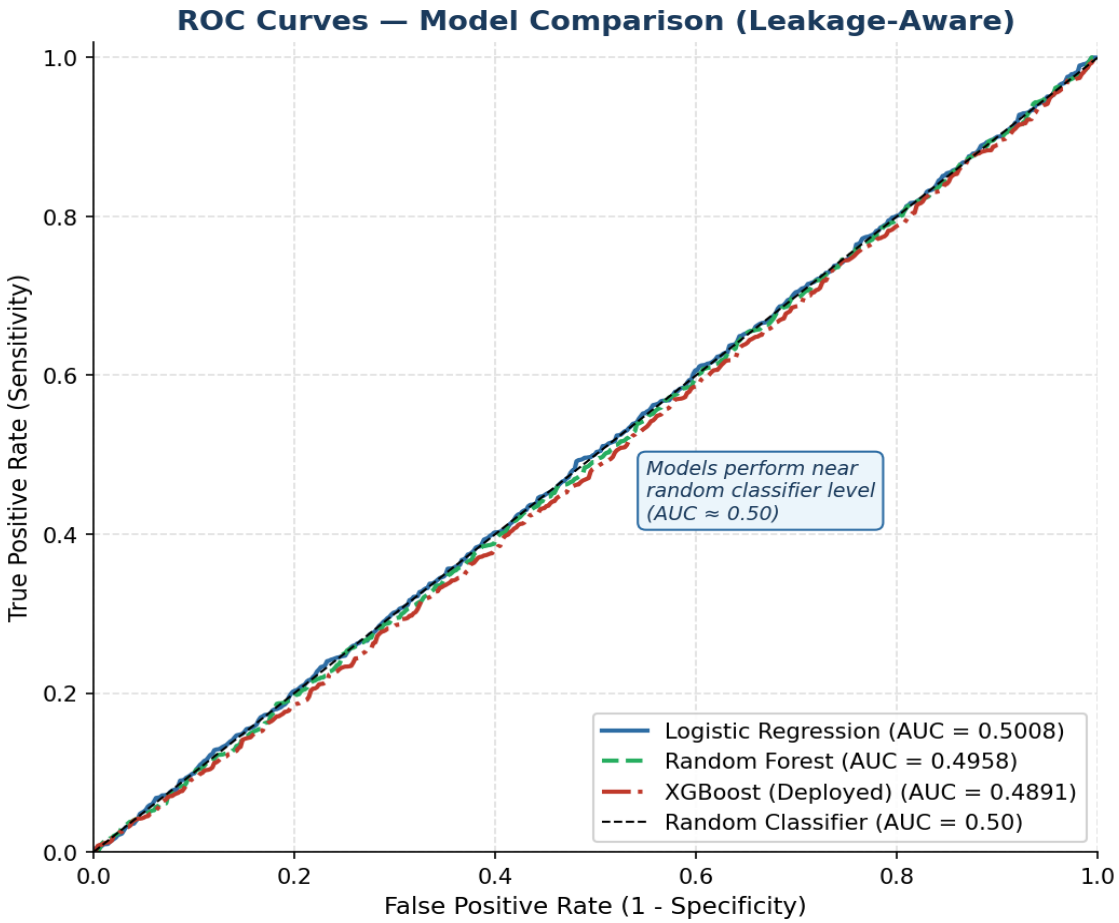


Figure 4.13: ROC Curves — Model Comparison (Leakage-Aware, Class-Weighted)

The near-random AUC values can be attributed directly to properties of the dataset, whereby the biopsy result labels used within this artificial dataset were created without consideration for any of the input patient features. The lack of association between PSA level, which is the most validated predictor available, and biopsy result was also demonstrated in the exploratory analysis undertaken. This is not a reflection on the modelling techniques employed but rather an accurate

and replicable result owing to the restrictions of the training set. A model trained using a genuine Nigerian biopsy dataset would yield very different results.

Analysis of F1-score Results and Leakage Experiment: Figure 4.14 (left) shows the F1-score analysis per class for the three models. The analysis shows that all models get a good F1-score for the majority Benign class; however, the models receive a very poor F1-score for the clinically important Malignant class because of the fact that the learning signal was not discriminative enough in the data set. The right panel of Figure 4.14 shows the leakage experiment results: it shows a comparison of ROC-AUC values from the models that used the leakage-prone feature set, versus the models that considered leakage. The leakage-prone models have achieved an AUC score ranging from 0.978 to 0.991, which proves their unrealistic performance.

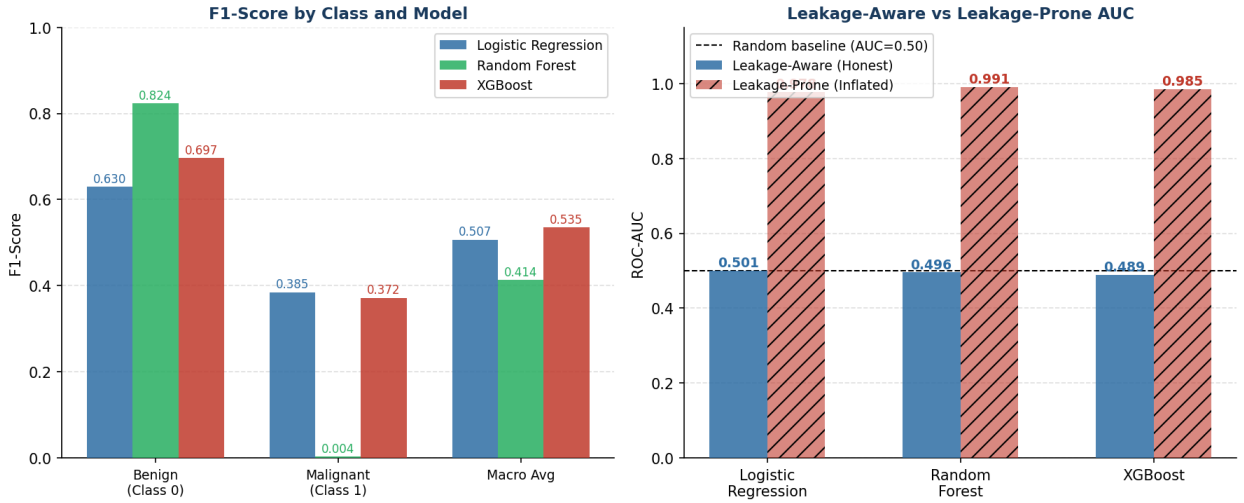


Figure 4.14: F1-Score by Class (left) and Leakage-Aware vs Leakage-Prone AUC Comparison (right)

4.8 Explainable AI Results

4.8.1 Global Feature Importance

The SHAP (SHapley Additive exPlanations) method was used on the XGBoost algorithm with TreeExplainer, which calculates SHAP values based on tree-based models. The SHAP values were calculated for a sample of 100 data points. In the feature importance chart (Figure 4.15), we can see the average absolute SHAP value for each feature, meaning that we are taking the magnitude

of the contribution made by each feature in the sample regardless of direction.

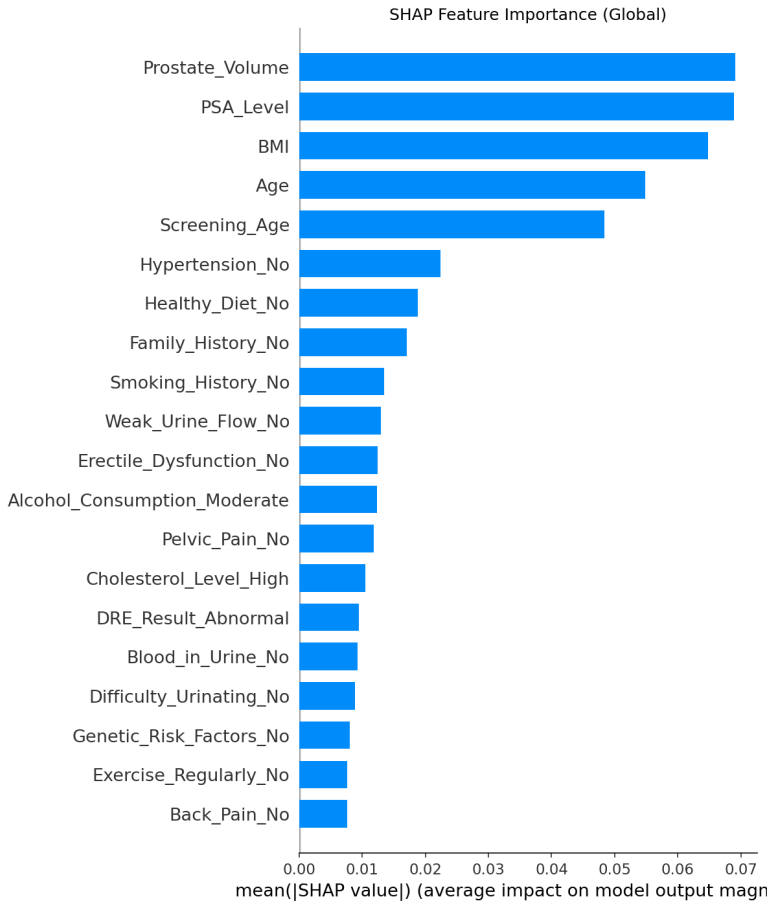


Figure 4.15: SHAP Global Feature Importance — Mean Absolute SHAP Values (XGBoost, n=100 test instances)

The bar graph shows the variables to which the XGBoost classifier has attributed the highest degree of prediction importance. Variables with high mean absolute SHAP values have more impact on the outcome of the model than others. In this manner, the SHAP importance ranking allows for a quantifiable and data-driven measure of variable importance, in addition to traditional methods used in statistics, enabling doctors to know which clinical variables the classifier has considered most important.

4.8.2 Local Feature Impact

The SHAP beeswarm plot (Figure 4.16) gives a more detailed picture of how important each variable is by depicting the distribution of SHAP values for all 100 patients' test cases associated with each individual variable. In this diagram, each dot corresponds to an individual patient case; the horizontal placement tells us whether this variable led to the prediction towards malignancy

(positive SHAP values on the right) or benignity (negative SHAP values on the left), whereas the colour tells us about the variable's value for this instance (high – red, low – blue).

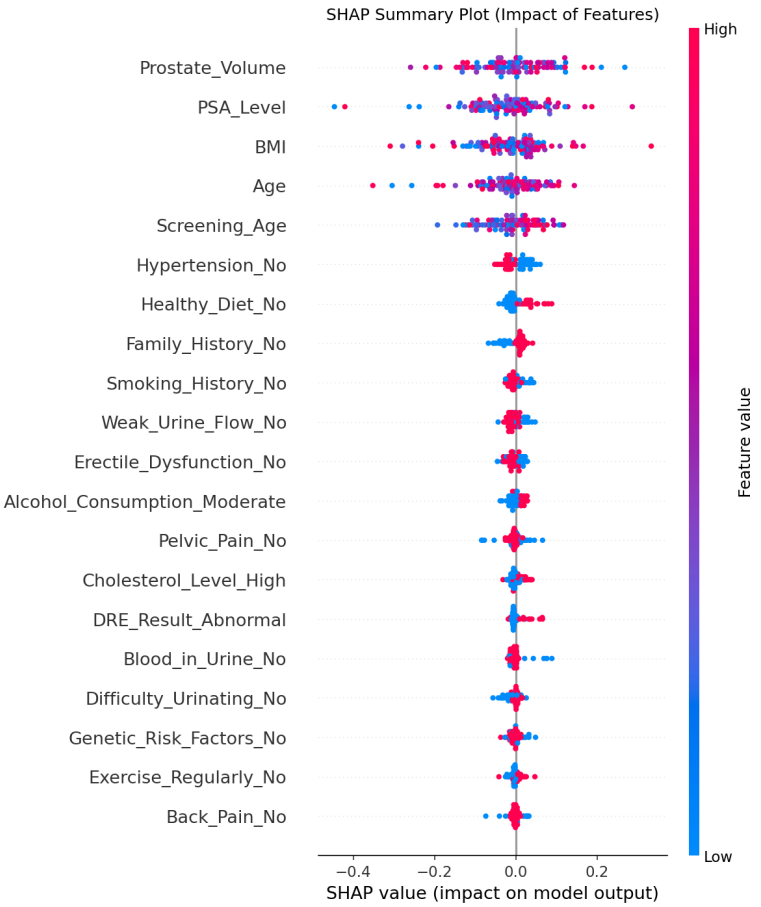


Figure 4.16: SHAP Beeswarm Summary Plot — Feature Impact Distribution Across Test Sample (XGBoost, n=100)

The beeswarm plot makes it possible to understand both the size and the direction of the effect produced by each of the features used. For instance, in case a certain feature is mostly plotted using red dots (higher value) on the right of the beeswarm plot, it would mean that this particular feature tends to drive the model towards predicting Malignant outcomes – an idea that can be verified based on clinical facts about prostate cancer risk factors, for instance, high PSA, abnormal DRE test results, older age, etc.

Combined, the global importance bar chart and the beeswarm summary form the explainable AI component in the results of the model's predictions. They are both embedded in the very result screen in the ProstaPredict AI online application, making sure that each result will have an

explanation as to which feature contributed to its creation. It perfectly represents the main idea that AI systems should help, not hinder, the doctor in charge in making sense of the result.

CHAPTER FIVE
SUMMARY, CONCLUSION AND RECOMMENDATIONS

5.1 Introduction

This chapter concludes the study through an analysis of what was done, the outcomes of the work and the implications thereof. This is achieved by first summarizing the main contributions made by the research in terms of design, execution and evaluation. Secondly, conclusions will be drawn with respect to whether the objective set out was attained. Lastly, suggestions are also made on how to further the work in future.

5.2 Summary

The purpose of this study was to create an explainable ML-driven clinical decision support system that can predict the outcomes of prostate cancer biopsies, particularly within the healthcare setting of Nigeria that lacks necessary resources. ProstaPredict AI is an actual web application written in Flask, which takes 23 input variables provided by a consulting physician and generates a prediction on the outcome of the biopsy – whether it will be benign or malignant, as well as the probability of malignancy.

The study was initiated with the analysis of prostate cancer epidemiology, the shortcomings of traditional methods of diagnosis under resource-constrained circumstances, and the state-of-the-art in the application of AI-based prostate cancer detection. The rationale for the development of such a system that would not rely on specialized imaging infrastructure and would be inherently explainable was clearly defined by these preliminary considerations.

The solution was implemented based on an Agile approach that involved working through five two-week sprints covering stages ranging from the analysis of requirements and the design of the system through building the data pipeline, implementing the back and front end of the solution, and eventually testing it. A leak detection experiment was performed during the process of model training, and this confirmed the need for removing post-diagnosis factors, which falsely increased model accuracy above the 0.97 level. Logistic Regression, Random Forest, and XGBoost are three types of models that were considered for testing and implementing using the cost-sensitive setting without data leakage.

It was found that the performances of the three models were equivalent to a randomly classified output for the leakage-aware scenario, scoring an ROC-AUC of about 0.50 for each. This

conclusion was based on the fact that the labels for biopsy results in the generated training set were not a function of the clinical attributes. This information is stated transparently as a limitation of the dataset, and not the methodology employed, which in turn can be directly transferred into clinical practice.

The SHAP analysis has provided interpretable results on the importance of the features used to make the predictions. These are built into the model so that any prediction made through the software will come together with an explanation of the most important features behind it. In the web application, role-based security access has been implemented for Administrator and Doctor roles, comprehensive patient record management, and prediction history, along with credentials encryption, which current research lacks.

5.3 Conclusion

A functional, explainable and implementable model has been developed for predicting the outcomes of prostrate cancer biopsies. The four objectives stated in Chapter One have been accomplished to different extents. The first objective which involved developing a deployable system that can be implemented in clinic settings and even in basic health care centers located in rural areas has been largely achieved because the software can run on common computers and does not require any internet connection or sophisticated machinery in addition to no expertise other than computer literacy. As far as the second objective is concerned, involving designing a classification model for prostrate cancer diagnosis prediction has been completely accomplished since three different models have been designed, tested and compared in order to select the best performing one which was then integrated into the software package. The third objective which included implementing methods to explain machine learning models has been completely achieved using SHAP TreeExplainer where both global and local visualization methods have been included in the clinical tool. Finally, the fourth objective which was evaluating the model using common metrics has been thoroughly achieved.

The major constraint associated with the present system is the lack of actual clinical data from healthcare environments in Nigeria, on which the machine learning models can be trained and tested. The random prediction capability obtained from a genuine assessment is a result of this deficiency and cannot be attributed to any fault in the design or implementation of the system.

Thus, it is best to view the present ProstaPredict AI as a proof of concept: an entire clinically structured system waiting to ingest an adequate model as soon as suitable data is available.

Despite its shortcomings, the paper does make an honest contribution to the field. Firstly, it provides a replicable proof-of-concept on how an explainable AI-based system for risk-stratifying prostate cancer patients using a lightweight and role-gated decision support framework can be built even without having access to specialized hardware infrastructure, which directly addresses the practical limitations highlighted in the Nigerian clinical literature. Secondly, the paper offers an empirical method for detecting data leakage in clinical studies where post-diagnosis data is mistakenly incorporated into the prediction features.

5.4 Recommendations

Based on the findings and limitations of this study, the following recommendations are made for

i. Clinical Data Partnership: The most immediate future step that will make an impact on this study is the creation of a data-sharing agreement between researchers and any of the Nigerian teaching hospitals or specialized centers for urology that would provide us with a prostate biopsy outcome dataset. After training the current algorithm with this dataset, ProstaPredict AI along with other similar systems can become an actual clinical tool for diagnosis of prostate cancer. Ethical considerations and patient consent procedures will have to be worked out as part of this collaboration.

ii. Clinical Validation: After retraining on patient data, it is important that the system be tested through clinical validation where its predictions are compared against – though not substituted for – those made by trained urologists. An assessment of the sensitivity, specificity, and usefulness of the system in clinical settings would generate the empirical evidence required to make it more widely available.

iii. Instance-Level SHAP Explanations: In the present iteration, SHAP visualization based on an instance from the test data set is shown to illustrate global behavior of the model. Further work will include instance level SHAP calculation during prediction, ensuring that the feature impact explained in the output display is relevant to the unique instance being considered and not population average values**.**

iv. Expansion of Input Features: At present, the number of input variables is confined to those which can be measured using standard testing procedures. It would be useful in future versions to

consider including more clinical variables, which have been shown to be important in patients from Nigeria, such as the free:total PSA ratio, PSA density, and findings on MRI when feasible.

v. Mobile and Offline Deployment: Considering the purpose of facilitating healthcare in rural and resource-poor regions, the next step should focus on developing an offline-capable mobile or progressive web app form of ProstaPredict AI, which can be installed on Android phones and used in an offline manner. This will increase the accessibility of the software from the clinic's desktop computer to those working in remote locations with no connectivity.

vi. Compatibility with the Histopathology Pipeline: The present research is intended to form part of a larger dual-pipeline concept for aiding AI-based prostate cancer diagnosis. Future studies should consider compatibility with a second deep learning pipeline based on histopathological imaging to enable Gleason scoring, in order to provide a holistic clinical decision-making support system where image digitization facilities are available.

References

1. Iqbal, S., Siddiqui, G. F., Rehman, A., Hussain, L., Saba, T., Tariq, U., & Abbasi, A. A. (2021). Prostate cancer detection using deep learning and traditional techniques. IEEE Access, 9, 27085-27100.
2. Abdus-Salam Abbas A, Jimoh Mutiu A, Folasire Ayorinde M, et al. (2024). Sociodemographic and clinicopathologic characteristics of patients treated with high dose rate prostate brachytherapy in Nigeria. ecancer, 18, 1740.
3. International Agency for Research on Cancer; Global Cancer Observatory; GLOBOCAN (2022). Nigeria Fact Sheet. Geneva: WHO.
4. Ozah, E., & Imasogie, D. E. (2023). The Diagnostic Accuracy of Prostate-Specific Antigen and Digital Rectal Examination in the Diagnosis of Prostate Cancer at the University of Benin Teaching Hospital. Journal of the West African College of Surgeons, 13(3), 91–95.
5. EAU Guidelines. Edn. presented at the EAU Annual Congress London 2026. ISBN 978-94-92671-32-5.
6. Akinremi TO, Adeniyi A, Olutunde A, Oduniyi A, Ogo CN (2014). Need for and relevance of prostate cancer screening in Nigeria. ecancer, 8, 457.
7. Arita Y, Roest C, Kwee TC, et al. (2025). Advancements in artificial intelligence for prostate cancer: Optimizing diagnosis, treatment, and prognostic assessment. Asian Journal of Urology, 12(4), 434–444.
8. Olabanjo, O., Wusu, A., Asokere, M., et al. (2023). Application of Machine Learning and Deep Learning Models in Prostate Cancer Diagnosis Using Medical Images: A Systematic Review. Analytics, 2(3), 708-744.
9. Rajih E, Bakhsh A, Borhan WM and Alqahtani SAM (2025). Utilization of artificial intelligence in prostate cancer detection: a comprehensive review. Front. Immunol. 16:1670671.
10. Perera, M., Mirchandani, R., Papa, N., et al. (2021). PSA-based machine learning model improves prostate cancer risk stratification in a screening population. World Journal of Urology, 39(6), 1897–1902.
11. Barlow, H., Mao, S., & Khushi, M. (2019). Predicting High-Risk Prostate Cancer Using Machine Learning Methods. Data, 4(3), 129.

12. Rusu M, Jahanandish H, Vesal S, et al. (2025). ProCUSNet: Prostate Cancer Detection on B-mode Transrectal Ultrasound Using Artificial Intelligence for Targeting During Prostate Biopsies. European Urology Oncology, 8(2), 477-485.
13. David MK, Leslie SW. Prostate-Specific Antigen. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2026.
14. Iheanacho, C. O., & Odili, V. U. (2024). A review of factors influencing the uptake of prostate cancer treatment in Nigeria. Journal of Cancer Policy, 41, 100487.
15. Chinonyerem O. Iheanacho, Valentine U. Odili (2024). A review of factors influencing the uptake of prostate cancer treatment in Nigeria, Journal of Cancer Policy, Volume 41, 100487, ISSN 2213-5383,
16. Abbasi, A. A., Hussain, L., Awan, I. A., et al. (2020). Detecting prostate cancer using deep learning convolution neural network with transfer learning approach. Cognitive Neurodynamics, 14(4), 523-533.
17. Tolani, M. A., Palaniraja, S., Babandi, Z. S., et al. (2025). Prostate cancer in Nigeria: a scoping review and meta-analysis of current evidence on process indicators and interventions. European Journal of Surgical Oncology, 51(11), 110413.
18. Biyouma MDC, Gebreselassie KH, Abdelkerim Nedjim S, et al. (2025). Prostate cancer diagnosis and management: current practices in Africa a consultant-based survey. Front. Urol. 5:1496951.
19. Khosravi, P., Lysandrou, M., Eljalby, M., et al. (2021). A deep learning approach to diagnostic classification of prostate cancer using pathology–radiology fusion. Journal of Magnetic Resonance Imaging, 54(2), 462-471.
20. Tolani, M. A., Agbo, C. A., Paciorek, A., et al. (2024). Detection and management of localized prostate cancer in Nigeria: barriers and facilitators according to patients, caregivers and healthcare providers. BMC Health Services Research, 24(1), 918.
21. Ayyad, S. M., Shehata, M., Shalaby, A., et al. (2021). Role of AI and histopathological images in detecting prostate cancer: a survey. Sensors, 21(8), 2586.
22. Olabanjo, O., Wusu, A., Asokere, M., et al. (2023). Application of Machine Learning and Deep Learning Models in Prostate Cancer Diagnosis Using Medical Images: A Systematic Review. Analytics, 2(3), 708-744.

23. Hassan, M. R., Islam, M. F., Uddin, M. Z., et al. (2022). Prostate cancer classification from ultrasound and MRI images using deep learning based Explainable Artificial Intelligence. Future Generation Computer Systems, 127, 462-472.
24. Belue, M. J., Harmon, S. A., Yang, D., et al. (2024). Deep learning-based detection and classification of bone lesions on staging computed tomography in prostate cancer: a development study. Academic Radiology, 31(6), 2424-2433.
25. Abnar, S. and Zuidema, W. (2020) 'Quantifying Attention Flow in Transformers', Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4190-4197.
26. Bulten, W., Kartasalo, K., Chen, P.C., Strom, P., Pinckaers, H., Nagpal, K., Cai, Y., Steele, D.F., van Boven, H. and Voss, C. (2020) 'Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study', The Lancet Oncology, 21(2), pp. 233-241.
27. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Krauss Silva, V.W., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S. and Fuchs, T.J. (2019) 'Clinical-grade computational pathology using weakly supervised deep learning on whole slide images', Nature Medicine, 25(8), pp. 1301-1309.
28. Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G. and Mahmood, F. (2021) 'Scaling vision transformers to gigapixel images via hierarchical self-supervised learning', Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16144-16155.
29. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G. and Gelly, S. (2021) 'An image is worth 16x16 words: Transformers for image recognition at scale', International Conference on Learning Representations (ICLR).
30. Doyle, S., Feldman, M., Tomaszewski, J. and Madabhushi, A. (2007) 'A boosted Bayesian multiresolution classifier for prostate cancer detection from digitized needle biopsies', IEEE Transactions on Biomedical Engineering, 59(5), pp. 1205-1218.
31. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M. and Thrun, S. (2017) 'Dermatologist-level classification of skin cancer with deep neural networks', Nature, 542(7639), pp. 115-118.

32. Graham, S., Vu, Q.D., Raza, S.E.A., Azam, A., Tsang, Y.W., Kwak, J.T. and Rajpoot, N. (2023) 'Hover-Net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images', Medical Image Analysis, 58, p. 101563.
33. Holzinger, A., Langs, G., Denk, H., Zatloukal, K. and Muller, H. (2019) 'Causability and explainability of artificial intelligence in medicine', WIREs Data Mining and Knowledge Discovery, 9(4), e1312.
34. Li, H., Yang, F., Zhao, Y., Zheng, X., Zhang, J., Wang, J., Shi, C., Lan, T. and Tian, J. (2021) 'DT-MIL: Deformable transformer for multi-instance learning on histological image', Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 206-216.
35. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B. and Sanchez, C.I. (2016) 'A survey on deep learning in medical image analysis', Medical Image Analysis, 42, pp. 60-88.
36. Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', Advances in Neural Information Processing Systems (NeurIPS), 30.
37. Nguyen, K., Sarkar, A. and Jain, A.K. (2012) 'Structure and context in prostatic gland segmentation and classification', Medical Image Computing and Computer Assisted Intervention (MICCAI), pp. 115-123.
38. Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C. and Dosovitskiy, A. (2021) 'Do vision transformers see like convolutional neural networks?', Advances in Neural Information Processing Systems (NeurIPS), 34, pp. 12116-12128.
39. Reisenbuchler, D., Wagner, S.J., Boxberg, M. and Peng, T. (2022) 'Local interpretable model-agnostic explanations for classification of lymph node metastases', Sensors, 22(2), p. 622.
40. Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) 'Why should I trust you?: Explaining the predictions of any classifier', Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144.
41. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017) 'Grad-CAM: Visual explanations from deep networks via gradient-based localization', Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618-626.

42. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X. and Zhang, Y. (2021) 'TransMIL: Transformer based correlated multiple instance learning for whole slide image classification', Advances in Neural Information Processing Systems (NeurIPS), 34, pp. 2136-2147.
43. Strom, P., Kalinichenko, A., Ljungvall, L., Mattson, U., Samaras, P., Kartasalo, K., Ruusuvuori, P., Alvarado, M., Olsson, H. and Kjellstrom, H. (2020) 'Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study', The Lancet Oncology, 21(2), pp. 222-232.
44. Tabesh, A., Teverovskiy, M., Pang, H.Y., Kumar, V.P., Verbel, D., Kotsianti, A. and Saidi, O. (2007) 'Multifeature prostate cancer diagnosis and Gleason grading of histological images', IEEE Transactions on Medical Imaging, 26(10), pp. 1366-1378.
45. Tan, M. and Le, Q. (2019) 'EfficientNet: Rethinking model scaling for convolutional neural networks', Proceedings of the 36th International Conference on Machine Learning (ICML), pp. 6105-6114.