

Cervical Cancer Diagnosis Prediction based on Risk Factor Analysis

Naomi Munyiri

122193

ICS 4A

Supervisor Name

Kevin Ochieng' Omondi

**Submitted in Partial Fulfillment of the Requirements of the Bachelor of Science in Informatics
and Computer Science at Strathmore University**

School of Computing and Engineering Science

Strathmore University

Nairobi, Kenya

October 2022

Declaration and Approval

I declare that this work has not been previously submitted and approved for the award of a degree by this or any other University. To the best of my knowledge and belief, the research documentation contains no material previously published or written by another person except where due reference is made in the research documentation itself.

Student Name: Naomi Munyiri

Admission Number: 122193

Student Signature: _____ Date: _____

The documentation of **Naomi Munyiri** has been reviewed and approved by **Kevin Ochieng' Omondi**

Supervisor Signature: _____ Date: _____

Acknowledgement

Primarily, I would like to thank God whose grace has guided me from the very inception to the completion of this project.

Apart from the efforts of me, the success of this project depended largely on the encouragement and guidelines of many others. I take this opportunity to express my gratitude to the people who have been instrumental in the successful completion of this project.

I give many thanks to the Strathmore University, School of Computing and Engineering Sciences, for giving me the opportunity to do this project.

My unalloyed appreciation goes to my ever-supportive supervisor, Kevin Omondi for his invaluable contributions and instructions throughout this journey. His encouragement to me during this project is highly appreciated. I am most grateful to him.

A special acknowledgement goes to Dr. Asaph Kinyanjui, Nairobi Hospice. I am in his intellectual debt for his constructive comments and suggestions towards the improvement of this project.

Finally, I want to thank my dear parents for clearing my doubts and giving me unfailing support and encouragement during the academic year and completion of this project.

Abstract

Cervical Cancer is the second common cancer in women in Kenya. Despite this, it ranks first, with the highest number of deaths due to late detection and diagnosis. Cancers are considered silent diseases because pain comes at advanced stages. Since pain is what generally takes one to hospital, one never knows until it is too late.

In the early stages of cervical cancer, a person may have no symptoms at all, making detection difficult. Regular screening is the only way to detect cancer early. Some of the approved screening modalities approved by World Health Organization and the Ministry of Health include HPV DNA testing and pap smear. Although women are strongly encouraged to take these tests, it is not the norm in Kenya and women who do take these tests describe them as intrusive. Regardless, early-stage test results are frequently referred to as false positives. This leaves the situation at a stalemate.

The solution is to develop a machine learning model that analyzes risk factors; including age and habits, in order to predict one's cervical cancer diagnosis. Unlike the HPV DNA and pap-smear tests, this model is a non-intrusive procedure that aims to work as a preventive and predictive measure.

Table of Contents

Declaration and Approval	ii
Acknowledgement	iii
Abstract	iv
List of Figures	ix
List of Tables	x
List of Equations	xi
List of Abbreviations	xii
Chapter 1: Introduction.....	1
1.1 Background Information	1
1.2 Problem Statement	2
1.3 Objectives.....	2
1.3.1 General Objective	2
1.3.2 Specific Objectives	3
1.4 Research Questions	3
1.5 Justification	3
1.6 Scope of the Study.....	4
1.7 Delimitations and Limitations	4
Chapter 2: Literature Review.....	5
2.1 Introduction	5
2.2 Cervical Cancer Screening in Kenya	5
2.2.1 Challenges in Cervical Cancer Screening.....	5
2.3 Related Works	6
2.3.1 Pap-Smear Test	6
2.3.2 Residual Neural Network Approach for Prediction of Cervical Cancer.....	8
2.3.3 Cervical Cancer Detection Using Causal Analysis and Machine Learning	9
2.4 Gaps in Related Works.....	10
2.5 Conceptual Framework	11

Chapter 3: Development Methodology	12
3.1 Introduction	12
3.2 Methodology	12
3.2.1 Requirements Gathering	13
3.2.2 Quick Design	13
3.2.3 Building Prototype	13
3.2.4 User Evaluation.....	13
3.2.5 Refining Prototype	13
3.2.6 Implement and Test.....	13
3.3 System Analysis	14
3.3.1 Database Schema	14
3.3.2 Use Case Diagram.....	14
3.3.3 Sequence Diagram	14
3.3.4 Entity Relationship Diagram.....	14
3.3.5 Class Diagram.....	14
3.3.6 Activity Diagram	14
3.4 System Design.....	15
3.4.1 Database Schema	15
3.4.2 Wireframes/Mockups.....	15
3.4.3 System Architecture.....	15
3.5 Deliverables.....	15
3.5.1 Documentation	15
3.5.2 Model and User Interface.....	15
3.6 Tools and Techniques.....	15
3.6.1 Google Colaboratory	15
3.6.2 Visual Studio Code	16
3.6.3 GitHub.....	16

3.6.4	Pandas	16
3.6.5	Matplotlib.....	16
3.6.6	XGBoost	16
Chapter 4: System Analysis and Design.....		17
4.1	Introduction	17
4.2	System Requirements	17
4.2.1	Functional Requirements	17
4.2.2	Non-Functional Requirements	17
4.3	System Analysis Diagrams	18
4.3.1	Use Case Diagram.....	18
4.3.2	Sequence Diagram	19
4.3.3	Class Diagram.....	20
4.3.4	Activity Diagram	21
4.4	System Design Diagrams	22
4.4.1	Database Schema	22
4.4.2	Wireframes/Mockups.....	23
4.4.3	System Architecture.....	26
Chapter 5: System Implementation and Testing.....		27
5.1	Introduction	27
5.2	Description of the Implementation Environment.....	27
5.2.1	Hardware Specifications	27
5.2.2	Software Specifications	27
5.3	Description of the Dataset	28
5.4	Description of Training	31
5.5	Description of Testing	32
5.5.1	Testing Paradigm	33
5.5.2	Testing Results.....	34

Chapter 6: Conclusions, Recommendations and Future Works	37
6.1 Conclusion.....	37
6.2 Recommendations	37
6.3 Future Works.....	37
References.....	38
Appendix.....	42

List of Figures

Figure 2.1 Pap test (Cervical Cancer Screening (PDQ®)–Patient Version - NCI, 2004)	7
Figure 2.2 a) Normal cervical smear image; b) Abnormal cervical smear image showing increased size of cell A’s nucleus and deep stain (Rajasekharan et al., 2015).	7
Figure 2.3 Structure of Residual Neural Network (Priyanka, 2021)	8
Figure 2.4 ResNet50 Architecture (Priyanka, 2021)	8
Figure 2.5 Conceptual Framework	11
Figure 3.1 Prototyping Diagram (Martin, 2020).....	12
Figure 4.1 Use Case Diagram	18
Figure 4.2 Sequence Diagram.....	19
Figure 4.3 Class Diagram	20
Figure 4.4 Activity Diagram	21
Figure 4.5 Database Schema.....	22
Figure 4.6 Homepage Wireframe	23
Figure 4.7 Login Wireframe	24
Figure 4.8 Register Wireframe	24
Figure 4.9 Form Wireframe	25
Figure 4.10 Results Wireframe	25
Figure 4.11 System Architecture	26
Figure 5.1 Risk Factors in the Dataset	29
Figure 5.2 Missing Data.....	30
Figure 5.3 Initial Train and Validation AUC as number of trees increase	31
Figure 5.4 Final Model Train and Validation AUC.....	32
Figure 5.5 Saving the model	32
Figure 5.6 Confusion matrix	33
Figure 5.7 Classification Report	33
Figure 5.8 White Box Testing.....	33
Figure 5.9 Black Box Testing	34
Figure 5.10 Testing results.....	36

List of Tables

Table 5.1 Hardware Requirements	27
Table 5.2 Software Requirements.....	28

List of Equations

Equation 2.1 Decision Tree Equation (Lilhore et al., 2022)	9
Equation 2.2 Random Forest Information Gain Equation (Lilhore et al., 2022)	9

List of Abbreviations

AUC – Area Under the Curve

CNNs – Convolutional Neural Networks

HPV – Human Papillomavirus

IARC – International Agency for Research on Cancer

MoH – Ministry of Health

WHO – World Health Organization

XGBoost – eXtreme Gradient Boosting

Chapter 1: Introduction

1.1 Background Information

The human body consists of trillions of cells. Normally, cells divide and form new cells when the cells become older. These cells then die, and those dead cells are replaced by the new cells. The uncontrollable count of new cell generation leads to cancer in the human body (What Is Cancer?, 2007). This production of unnecessary new cells leads to tumors (Priyanka, 2021). Cancers can be malignant or benign; with malignant meaning they spread to other tissues in the body. Cervical cancer takes place when malignant tumour cells grow in the cervix which is located in the lower part of the uterus, the female's reproductive system (The Application of Machine Learning in Cervical Cancer Prediction | 2021 6th International Conference on Machine Learning Technologies, 2021).

In Kenya, according to the global cancer observatory through the International Agency for Research on Cancer (IARC), cervical cancer ranked second common across women of all ages. Worldwide, it is considered the fourth most common female malignancy with around 270,000 deaths in 2015. Approximately 90% of these deaths occurred in underdeveloped countries ([PDF] Cervical Cancer: Machine Learning Techniques for Detection, Risk Factors and Prevention Measures | Semantic Scholar, 2020). The World Health Organization (WHO) also points out that the mortality rates when cervical cancer is concerned have risen mostly due to the absence of screening and treatment programs.

Although cervical cancer sounds prevalent, it can be easily mitigated with regular screening tests. Screening tests include cervical cytology- also called Pap smear, testing for the Human Papilloma Virus or both (Cervical Cancer Screening, 2021). However, on the patients' side, going for screenings may not be a cultural norm. This may be caused by a variety of reasons, be it the time taken, the costs for the test and hospital visit, and the intrusive nature of the test. In addition to that, the Pap smear test and HPV DNA tests, cells are taken from the cervix and vagina and examined under a microscope, are highly dependent on the doctors' experience and are always liable to human inaccuracies.

A web-based machine learning model created to not only be non-intrusive but also act as a preventive measure in the face of cervical cancer could potentially be accessible, affordable, and culturally acceptable to women (Reaching 2030 Cervical Cancer Elimination Targets - New WHO Recommendations for Screening and Treatment of Cervical Pre-Cancer, 2021). The model did this by analyzing the primary risk factors for cervical cancer for example HPV,

poor menstruation sanitation, adolescent pregnancy, tobacco use, alcohol consumption, multiple sexual partners, and oral prevention methods, just to name a few (Lilhore et al., 2022). The algorithm then forecasted the diagnosis of a person based on these factors.

1.2 Problem Statement

Cervical cancer is one of the most common types of cancer, affecting women over the world, across all ages. This is heavily influenced by several risk factors that increase the likelihood of developing cervical cancer. If it is not detected in the early stages, the patient's mortality rate skyrockets. As a result, early screening is recommended (Cervical Cancer Screening, 2017).

Early is relative in this case because, as long as cervical cancer is in its early stages with no symptoms, all tests obtained are deemed false positives or are postponed. This undermines the goal of early detection. Moreover, it is not a cultural norm for Kenyans to go for medical screenings on a regular basis. This is due to a variety of factors, including unaffordability, inaccessibility, and cultural unacceptability.

The cost of testing is far beyond the means of the majority of Kenyans, who earn less than \$2 per day (Poverty in Kenya, 2018). Aside from the free public screenings, pap smears are still prohibitively expensive. Furthermore, health care facilities that provide screening tests are very rare. As a result, the high cost of transportation makes access to health care facilities difficult. Even so, access to health facilities may not result in access to services because test kits are often unavailable. Finally, cultural unacceptability is the most common reason people avoid screening tests. This is because the tests are perceived as intrusive and may cause discomfort (Kivuti-Bitok et al., 2013).

Bearing all these in mind, there was a need to develop a model that analyzes primary risk factors known to lead to cervical cancer; thus, predicting the patient's diagnosis result. With these results, one can seek medical attention as soon as possible, if necessary, in an affordable, accessible, and culturally acceptable manner.

1.3 Objectives

Below are my objectives, both general and specific that guided the development and design of my research.

1.3.1 General Objective

To develop a machine learning model and deploy it on the web that will analyze risk factors to improve cervical cancer diagnosis prediction.

1.3.2 Specific Objectives

- i. To study and identify how cervical cancer screening tests are currently done, read, and analyzed in Kenya.
- ii. To evaluate the difficulties in cervical cancer screening.
- iii. To study and review how related works have solved the problems faced in taking and analyzing cervical cancer screening tests.
- iv. To develop and design a diagnosis prediction model that analyzes risk factors to improve cervical cancer biopsy prediction.
- v. To test and validate the developed diagnosis prediction model.

1.4 Research Questions

- i. How are cervical cancer screening tests done, read, and analyzed in Kenya?
- ii. What are the challenges in cervical cancer screening?
- iii. How have related works solved the problem faced in taking and analyzing cervical cancer screening tests?
- iv. How will the diagnosis prediction model be designed and developed?
- v. How will the diagnosis prediction model be tested and validated?

1.5 Justification

The importance of early detection of cancer cannot be overlooked. Cancer detection at late stages, which is common in developing countries such as Kenya, reduces the chances of successful treatment. This leads to cancer deaths that could have been avoided (WHO, 2018).

Prediction is common in oncology: innumerable decisions by patients, family members, oncologists and other care providers depend on assessing the likelihood of future events. This is seen across to screening. Screening is recommended for those at elevated risks of cancer; whether due to age, for example, older individuals are advised to consider colonoscopy or risk factors for example, lung imaging in smokers (Prediction Models in Cancer Care - Vickers - 2011 - CA: A Cancer Journal for Clinicians - Wiley Online Library, 2011). Despite this, when it comes to cervical cancer, pap-smear tests remain out of reach for many women.

This is where the developed model came in. The model aimed to predict the patient's diagnosis by assessing factors ranging from age to habits, allowing for early treatment if necessary and reducing the number of cancer-related deaths due to late detection. This model also significantly contributed to the WHO cervical cancer elimination initiative 2030 by removing access, affordability, and cultural acceptability barriers that women currently face (Reaching

2030 Cervical Cancer Elimination Targets - New WHO Recommendations for Screening and Treatment of Cervical Pre-Cancer, 2021).

1.6 Scope of the Study

The model primarily focuses on cervical cancer diagnosis prediction based on risk factor analysis. It functions as a non-intrusive screening procedure that allows women to know their status whenever and wherever is convenient for them. These findings will then be communicated to the inquirer, allowing them to make an informed decision about whether or not to seek further medical attention.

1.7 Delimitations and Limitations

This study is limited to a prediction system that acts as a preventive measure in the face of cervical cancer and should not be used as a diagnostic tool.

The anticipated limitations include a lack of training power and missing values from the dataset, which jeopardize the data's sufficiency. These missing values exist because some patients refused to answer certain questions out of concern for their privacy. Furthermore, the model may not be a humane way of receiving bad news.

Chapter 2: Literature Review

2.1 Introduction

As cervical cancer is one of the most common types of cancer affecting women all over the world, the regular screening agenda has been pushed by both the media and healthcare workers to promote early detection in order to maximize patients' survival rates. This chapter discusses the current methods used in the screening of the cervix in Kenya, the existing gaps in the current method and how existing works have addressed those problems.

2.2 Cervical Cancer Screening in Kenya

According to Fontham et al. (2020), cervical cytological testing has been the cornerstone of cervical cancer screening for more than 50 years, first with the Pap test and more recently with liquid-based cytology. In order to lower the incidence, mortality, and morbidity of cervical cancer, the main objective of cervical cancer screening is to identify curable abnormalities and precancers that are likely to develop into invasive cancer.

The Pap test and the HPV test are the two possible test kinds that could be requested (Cervical Cancer Screening, 2017). A Pap test involves examining cells taken from the cervix under a microscope for the presence of abnormal cells, whereas an HPV test involves detecting the DNA of high-risk types of HPV in a cervix sample. HPV is a sexually transmitted virus that is the cause of nearly all cervical cancers. Cell samples are obtained for both during speculum examination. A speculum examination is when a speculum is inserted into the vagina to widen it. A brush is then inserted into the vagina to collect cells from the cervix (Definition of Pelvic Exam - NCI Dictionary of Cancer Terms - NCI, 2011).

The tests obtained are analyzed manually. If the screening tests are abnormal, the doctor may do order more tests such as a biopsy.

2.2.1 Challenges in Cervical Cancer Screening

The primary health care center is the patient's first point of contact for healthcare and screening services. These centers also serve as a hub for referrals to specialized services. According to Rosser et al. (2015), inadequate infrastructure and a lack of basic supplies are among the challenges that healthcare workers face when it comes to cervical cancer screening. The study was supported by a tertiary level facility in Kigali that found that 33% of symptomatic cervical cancer patients did not have a speculum exam at the referring lower level facilities despite the suspicion of cervical cancer (Cervical Cancer Screening at a Tertiary Care Center in Rwanda -

ScienceDirect, 2017). These findings are concerning considering a basic speculum is the evaluation required for screening.

Because of the complexities of the results, screening test analysis necessitates expertise and experience. The analysis of these tests is the responsibility of the health workers. Staffing has been identified as a barrier to cervical cancer screening. As per Chirenje et al. (2001), while there may be an adequate number of personnel for cervical cancer screening services, these personnel may lack the necessary screening skills. As a result, the patient is misdiagnosed, resulting in unnecessary follow-up tests.

Kivuti-Bitok et al. (2013) states that the Kenyatta National Hospital charges Kshs 550 (approximately \$7) for a pap smear. However, the patient must also pay Kshs 500 (approximately \$6.5) for a file or identification card. The patient must then wait two weeks for the pathologist to return the pap results. The high cost of screening is directly related to the cost of disposable speculums. This is not an affordable option in an economy where the average person earns less than \$2 per day (Poverty in Kenya, 2018).

Following a study carried out by Kivuti-Bitok et al. (2013), many health care workers reported that many patients find the screening procedure to be too invasive, embarrassing, and against African culture. It is seen as culturally unacceptable for some women to have young nurses and doctors see their private parts. Others felt that allowing a male to see their private parts was unacceptable. Even among health-care workers, screening is still regarded as an unpleasant procedure associated with the risk of infection. This is one of the primary reasons that most women avoid cervical screening tests.

2.3 Related Works

Some of the related works include:

2.3.1 Pap-Smear Test

A Pap smear, also known as a Papanicolaou smear, is a microscopic examination of cells scraped from the cervix used to detect cancerous or pre-cancerous cervix conditions. Screening should begin at the age of 21 or within three years of the onset of sexual activity and can be stopped at the age of 70 years if no abnormal Pap test results have been obtained in the previous ten years (“Recommendations on Screening for Cervical Cancer,” 2013). The patient is placed in lithotomy position, which involves facing up with arms to the side but separated and supported legs. A speculum is then used to examine the cervix. Scraping is completed and evenly distributed onto a glass slide, which is immediately fixed with 95 percent ethyl alcohol

and ether to avoid air drying artifacts (Mehta et al., 2009). The Figure 2.1 below shows a Pap test.

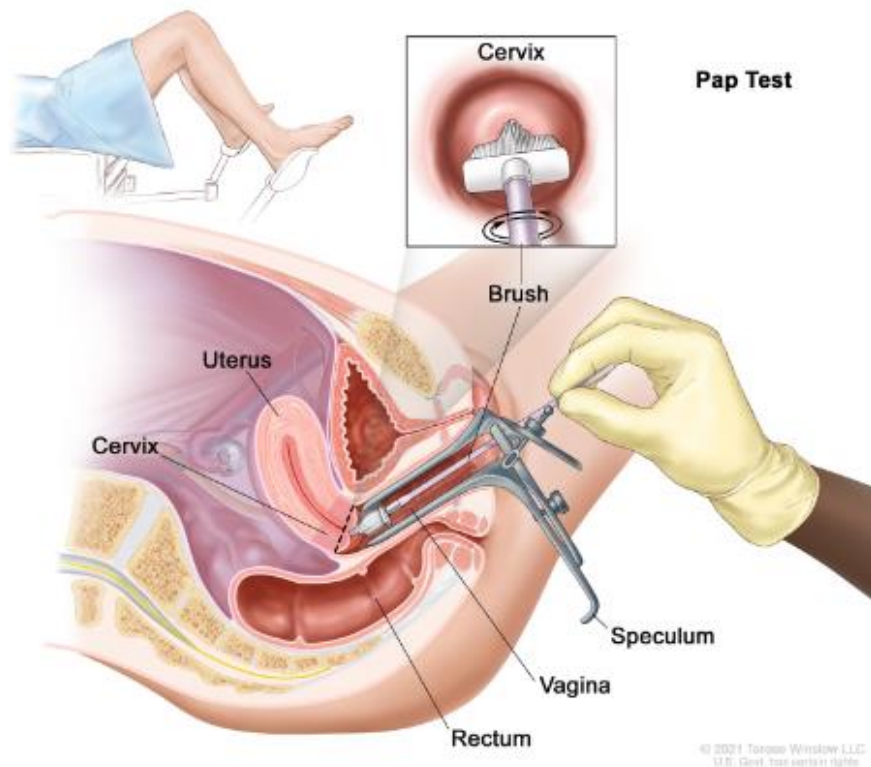


Figure 2.1 Pap test (Cervical Cancer Screening (PDQ®)–Patient Version - NCI, 2004)

The result can either show a normal cervical smear result or an abnormal cervical smear result as seen below in Figure 2.2.

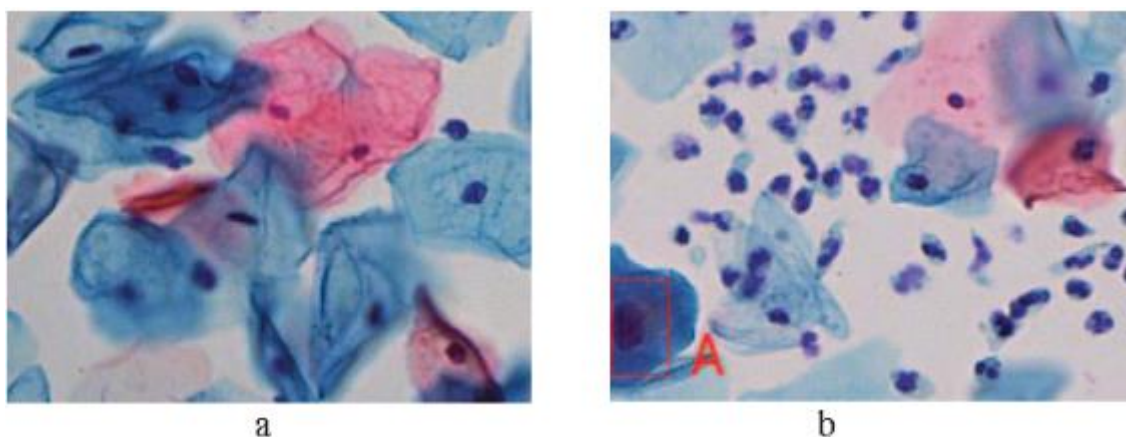


Figure 2.2 a) Normal cervical smear image; b) Abnormal cervical smear image showing increased size of cell A's nucleus and deep stain (Rajasekharan et al., 2015).

This procedure not only ensures that the result is normal, but it also aids in the prevention of cervical cancer by detecting cell changes in the body that would progress to cancer if left

untreated. Pap tests are critical for detecting cervical cancer before it spreads, which means less treatment and less time spent recuperating (Lee, 2014).

2.3.2 Residual Neural Network Approach for Prediction of Cervical Cancer

A study conducted by (Priyanka, 2021) combines pap-smear images with Deep Learning techniques to predict cancerous cells. Images from a database with seven different classes and twenty different features are collected. The dataset's seven cell classes are divided into two categories: normal and abnormal.

Convolutional Neural Networks (CNNs) are hidden layers that are used to process images. Residual Neural Networks (ResNet) are being studied for their ability to predict cervical cancer cells. The addition of layers one after the other in a CNN architecture causes degradation, which can be solved by introducing ResNet, which solves the problem by introducing residual blocks that generate residual functions, allowing us to adjust the input features to upgrade the high-level features(Priyanka, 2021).

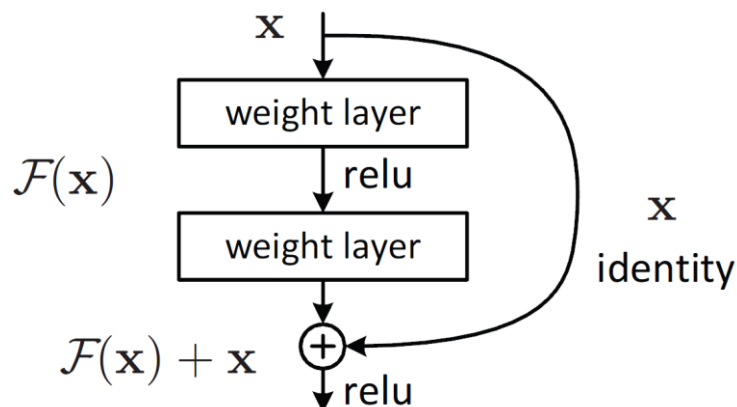


Figure 2.3 Structure of Residual Neural Network (Priyanka, 2021)

ResNet has many variants in it. The study uses ResNet50 for the prediction.

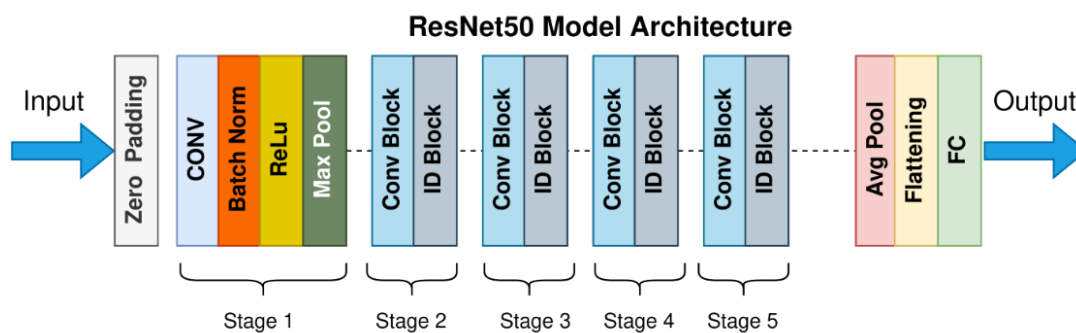


Figure 2.4 ResNet50 Architecture (Priyanka, 2021)

The prediction process was done only by selecting the image manually from the taken dataset. The image is then resized to specific size because it is easier to extract features from images when they are all the same size. The colored images are then converted to gray scale images in the following step. The given dataset is divided into training and testing sets, and the manually selected image is converted into pixels using filters from the pre-trained model, and the image is identified. As a result, the image and class to which it belongs are displayed.

2.3.3 Cervical Cancer Detection Using Causal Analysis and Machine Learning

A mathematical machine learning-based model for analyzing the possibility of cervical cancer was developed in the study by Lilhore et al. (2022). The prediction was investigated using a total of eight body factors. The study looked at cervical cancer and the various risk factors that contribute to its development. To create classification models, this study used random forest and decision tree.

A decision tree is a non-supervised learning technique that is commonly used to solve regression and classification problems. The goal is to use standard decision rules and advanced analytic features to expand a predictive model of the prediction error. An entropy E can be represented as the equation below. E represents entropy, s means samples, P_y represents the probability of yes, P_n represents no, and n represents the number of samples.

$$E(s) = \sum_{k=0}^n \binom{n}{k} - P_y * \log_2 P_n$$

Equation 2.1 Decision Tree Equation (Lilhore et al., 2022)

Random Forest is a regression and classification tree-based ensemble learning algorithm. A bootstrap specimen size is used to train each tree. The information gain for random forest can be calculated in Equation 2.2 below, where T represents the target variable, X represents the feature set to be split, and $\text{Gain}(T, X)$ represents the entropy value for dividing the data feature set X .

$$\text{Gain}(T, X) = \text{Entropy}(T) - \text{Entropy}(T, X)$$

Equation 2.2 Random Forest Information Gain Equation (Lilhore et al., 2022)

The performance of the random forest methods was reasonably good with accuracy, precision, and other parameters for identifying cervical cancer risk.

2.4 Gaps in Related Works

The existing solutions from the related works above are various contributions to cervical cancer screening, and prediction. Although efficient, there are a number of concerns that have not been tackled by the solutions presented above.

The Pap smear test, as described in the related works section poses several challenges. The test has a significantly low uptake due to a lack of accessibility, affordability, and widespread cultural unacceptability. Referencing a study carried out by Tiruneh et al. (2017), only 19.40 percent of married Kenyan women who participated and knew about cervical cancer reported ever being tested for cervical cancer. There is a high risk of false positive results for women who are tested, especially if the test is performed too early. When a Pap test produces a false positive result, it can be stressful, and it is usually followed by additional tests and procedures such as a colposcopy, cryotherapy, or LEEP. These have been linked with an increased risk of long term effects on fertility, and pregnancy especially in younger women.

The pap-smear produces pap-smear images, which are discussed in the section that discusses the residual neural network approach. Although the solution may reduce the number of false positives (Priyanka, 2021), it is primarily a doctor-oriented model. The patient cannot effectively interact with the model due to the expertise required when selecting the samples.

In the final section that focuses on the hybrid model, the solution centers on the use of random forest and decision tree algorithms. Decision trees are simple to grasp and provide a clear visual to aid in decision making. Despite this, there are several drawbacks, including overfitting which is when the model becomes so good at making predictions for a specific dataset that it performs poorly on a different dataset, bias error; when too many restrictions are placed on target functions, and variance error which is how much a result will change based on changes to the training set (Glen, 2019). The random forest steps in to reduce some of the variance seen in decision trees. Despite this, there is a chance that most of the trees could have predicted with some random chance because each tree had its own circumstances, such as sample duplication (Gupta, 2021).

The developed solution bridged all of these gaps because: the model is deployed on the web, making it available at all times, affordable to anyone with internet access, and culturally acceptable due to the lack of intrusion into personal space. Furthermore, because the model uses first-hand information of various variables that a patient fills in, it is patient oriented, making the model more objective. Finally, the XGBoost algorithm was used in the developed

model. The purpose of the use of this algorithm was to ensure that predictions are not made by chance, but rather with a thorough understanding of the data.

2.5 Conceptual Framework

The user enters the necessary variables required into the website. The website is linked to a background with a machine learning model that will analyze the input, predict the patient's diagnosis, and display the results to the user using the XGBoost Algorithm. Based on the results, the user can then decide whether medical attention is required.

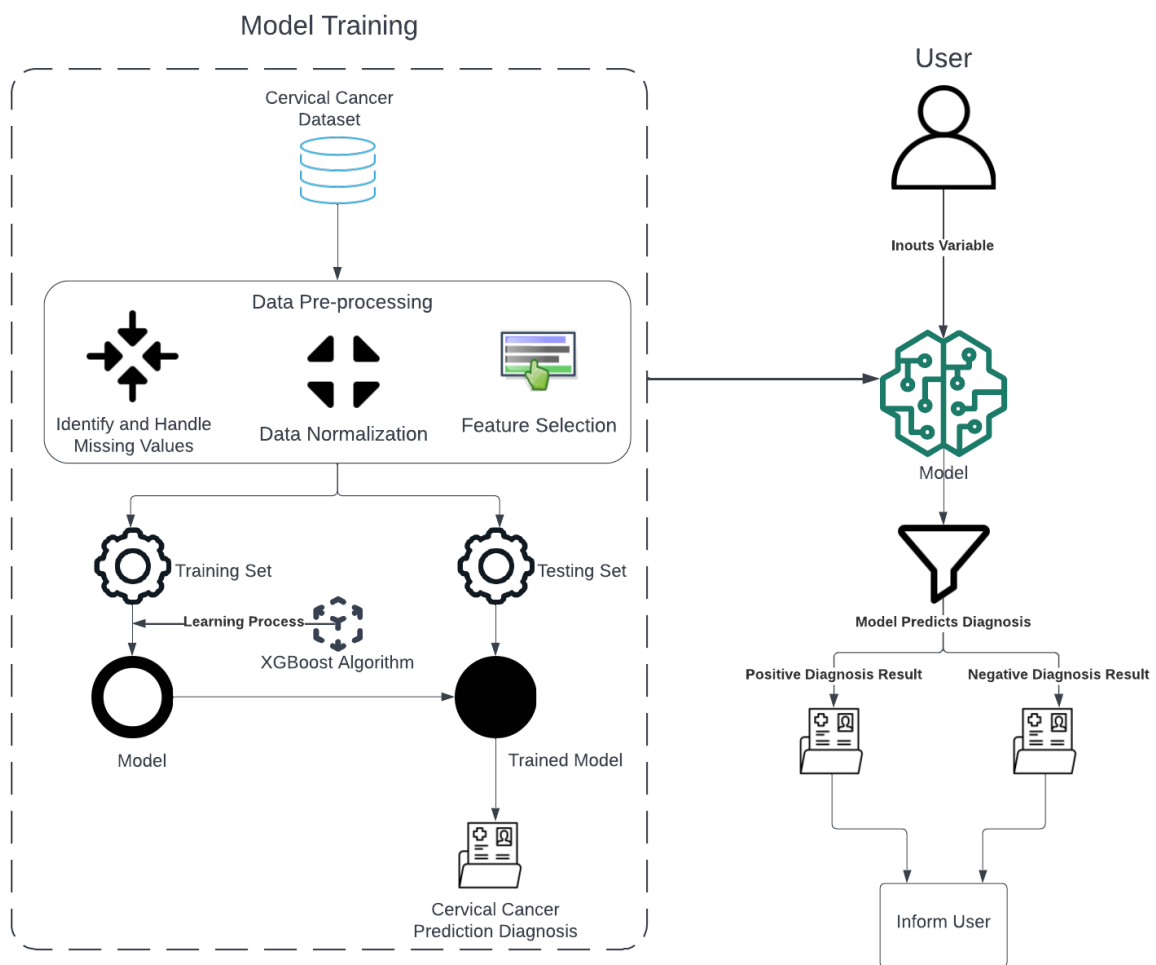


Figure 2.5 Conceptual Framework

Chapter 3: Development Methodology

3.1 Introduction

This chapter discusses the approach that was taken in the development of the system. It highlights the design paradigm, analysis and design approaches that were followed and finally the tools and techniques that were used.

The design paradigm that was used to develop the system was the Object-Oriented Analysis and Design (OOAD). This choice was mainly because of two reasons: the risk while using this technique is low, with its reusability high, and it allows for changes to be made easily according to user needs.

3.2 Methodology

The methodology that was applied in the development of this system was the prototype methodology. A prototype is a rudimentary working model of a product built as part of the development process.

Prototyping improves the quality of the requirements provided to customers. Most customers want to feel like they are involved with the intricate details of their project. Prototyping allows for this as it requires user involvement, allowing them to see and interact with working models of their project.

Systems, much like the one developed, which needs users to fill out forms before data is processed can use prototyping effectively to give the look and feel of the software before final development (SDLC - Software Prototype Model, 2020).

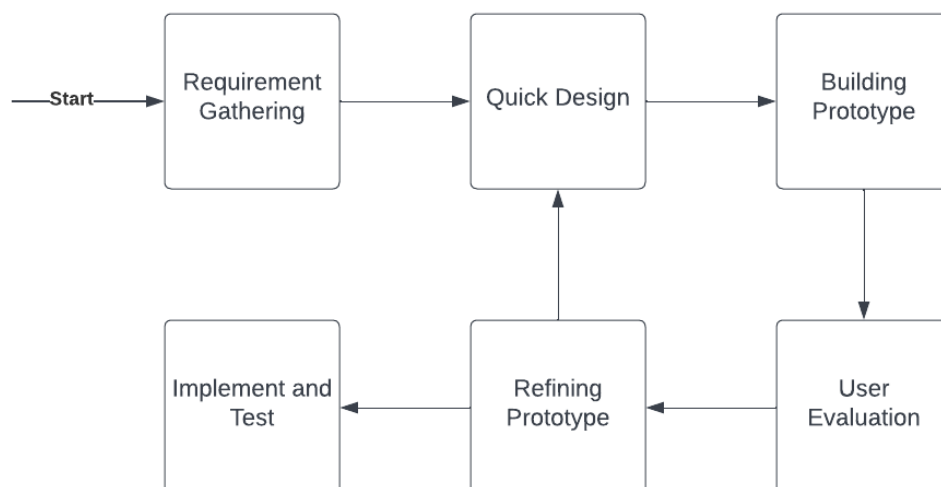


Figure 3.1 Prototyping Diagram (Martin, 2020)

3.2.1 Requirements Gathering

This initial stage involves understanding the very basic product requirements especially in terms of the user interface. The requirements needed for the system are gathered here from secondary sources. The UCI repository was used to provide a list of risk factors for cervical cancer leading to a biopsy examination (Cervical Cancer Risk Classification, 2018). These risk factors are the requirements of the system. Furthermore, journals on cervical cancer, and prediction and detection algorithms were used as points of reference.

3.2.2 Quick Design

The second stage was the design stage. The system was designed based on the requirements gathered. Analysis of the requirements of the system was established to assist with the revealing of important information on tools needed to develop and test the system, the need of stakeholders, and what was needed to make the application development process successful.

3.2.3 Building Prototype

The third stage entailed the building of the prototype. The prototype was a working model of the product containing basic features of the system. These features were given to the user to garner the look and feel of the final product.

3.2.4 User Evaluation

The prototype developed was then evaluated by the users in this stage. The testing paradigms that were used were the white box and black box testing.

White box testing was used to test the internal software architecture, check input flow, and improve usability of the software. The tester had access to the code. Black box testing was used to test the functionality of the software application without having knowledge of the internal code. It focused primarily on the requirements and specifications of the software.

The feedback that was collected was used to refine the product that was in development.

3.2.5 Refining Prototype

In the Refine Prototype stage, the feedback was discussed, and the accepted changes were incorporated into a new prototype. The cycle repeated until the system met the expected requirements.

3.2.6 Implement and Test

In this phase, the users tested the system one last time. No errors were detected, the system met all user needs and thus the final system was implemented.

3.3 System Analysis

In this phase, the classes and system requirements were determined. This technique aimed to improve the system by ensuring that all the components of the system were working as expected. Since Object-Oriented System Architecture and Design (OOAD) was used, the user requirements determined the design.

The system requirements were represented using the following diagrams:

3.3.1 Database Schema

A database schema is the structure that represents the logical view of the entire database. It defines exactly how the data is organized and the relations between them are associated. The database schema explained the relationships in the database.

3.3.2 Use Case Diagram

A Use Case diagram summarizes the details of system users (actors) and their interactions with the system. It may represent: Scenarios in which the system interacts with people, or external systems, Goals that the system helps the users achieve and the system scope.

3.3.3 Sequence Diagram

These diagrams detail how interactions are carried out. They visually show the order of the interaction using virtual axis to represent what messages are sent and when. The Sequence Diagram was used to map the processes involved in this system.

3.3.4 Entity Relationship Diagram

An Entity Relationship Diagram shows the relationship of entity sets (component of data) stored in a database. This was used to sketch out the design of the database.

3.3.5 Class Diagram

A class diagram is made up of a set of classes and a set of relationships between the classes. This diagram is a type of static structure diagram that was used to describe the structure of the system by showing the system's classes, their attributes, operations, and relationships.

3.3.6 Activity Diagram

An activity diagram is an advanced flowchart that models the flow from one activity to another. It was used to describe the dynamic aspects of the system.

3.4 System Design

The section below focuses on defining the architecture, interfaces, and data that the system required so as to satisfy the requirements set.

3.4.1 Database Schema

This is the skeleton structure that defines the entities in a system and their relationship.

3.4.2 Wireframes/Mockups

A wireframe is a simplified demonstration of the interface elements and how they will exist on a webpage. A mockup however is a more detailed demonstration of how the interface will look like. The wireframes gave a general view of the interfaces that were presented in the system. The mockups represented the ways in which the users will interact with the system.

3.4.3 System Architecture

The system architecture serves as a blueprint for the system. The system consists of a website which will be accessed online, after a user logs in and fills in the form. The prediction model will run in the background, and a prediction of the diagnosis will be sent back to the user.

3.5 Deliverables

Below are the components that constitute the developed system.

3.5.1 Documentation

A system documentation is a document that is used to define the objectives and requirements of a project. It comprehensively summarizes information including the problem and the strategies used to solve said problem.

3.5.2 Model and User Interface

A predictive model was developed, tested, and deployed on the web where users will be able to interact with the model.

3.6 Tools and Techniques

The developed system is a predictive model deployed on the web. The tools that were used in the development of the system are outlined and discussed briefly below.

3.6.1 Google Colaboratory

Google Colab is a product that allows anybody to write and execute arbitrary python through the browser.

3.6.2 Visual Studio Code

Visual Studio Code is a code editor redefined and optimized for building and debugging modern web and cloud applications.

3.6.3 GitHub

GitHub is a code hosting platform for version control and collaboration. It lets teams work together on projects from anywhere.

3.6.4 Pandas

Used to analyze and manipulate data, pandas is a fast, powerful, flexible, and easy to use open source tool built on top of the Python programming language.

3.6.5 Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

3.6.6 XGBoost

XGBoost is an optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. It implements machine learning algorithms under the Gradient Boosting Framework.

Chapter 4: System Analysis and Design

4.1 Introduction

This chapter is based on the analysis and design of the developed system. These two are extensively elaborated on: with System Analysis focusing on the requirements that the system fulfilled, and System Design involving the specific design of the systems' architecture and outlining all the requirements that were needed to implement this system. To do this, Object-Oriented Analysis and Design (OOAD) was used because of its low risk using the technique, and with its high reusability, allowing changes to be made easily according to the user needs.

4.2 System Requirements

Some of the system requirements reviewed in the project include:

4.2.1 Functional Requirements

Functional requirements are referred to as statements of services that the system should provide. These are the tangible parts of the system. They provide specification of how the system should react to input, and how the system should behave in situations. The functional requirements for the system included:

- (i) Authentication Module- The system must collect the new user information (Name, Email Address, Password) to allow users to register their accounts. The system must hash, using MD5 hashing, all user passwords before saving them into the database. Moreover, the system must allow the users to verify their accounts after registration by sending a link to the email addresses they inputted during registration. The system must allow the users to log into their account by entering their details (Email Address and Password).
- (ii) Form and Upload Module- This module enables the users to feed the form variables and upload this data to the system for diagnosis prediction.
- (iii) Machine Learning Module- This module is fed the form details that is uploaded by the user. It then analyses the data and predicts the diagnosis of the user.
- (iv) Notification Module- This module alerts the user once the prediction is completed.

4.2.2 Non-Functional Requirements

Non-functional requirements are referred to as constraints on the services that the system should provide. These are the qualities that a system does not necessarily need to function but are necessary to ensure efficiency. Some of the expectations of the system are as follows:

- (i) System Availability- This is done by ensuring that the system is hosted on a reliable platform i.e., the web.
- (ii) System Security- The user password is hashed using MD5 to ensure its inaccessibility when stored on the database.
- (iii) System Usability- The user interface of the system is easy to navigate while ensuring that the user expectations are met.
- (iv) System Performance- The accuracy of the model and the speed of the analyzing of the form details by the model ensures the performance of the system.

4.3 System Analysis Diagrams

Some of the system analysis diagrams considered are as follows:

4.3.1 Use Case Diagram

The Figure 4.1 represents a use case diagram consisting of two main actors: the user and the model. The use cases in the diagram include registering, login, updating of profiles, uploading of the risk form, the use of the diagnosis model to get a diagnosis and the viewing of the results. The relationships between the cases are also represented in the diagram.

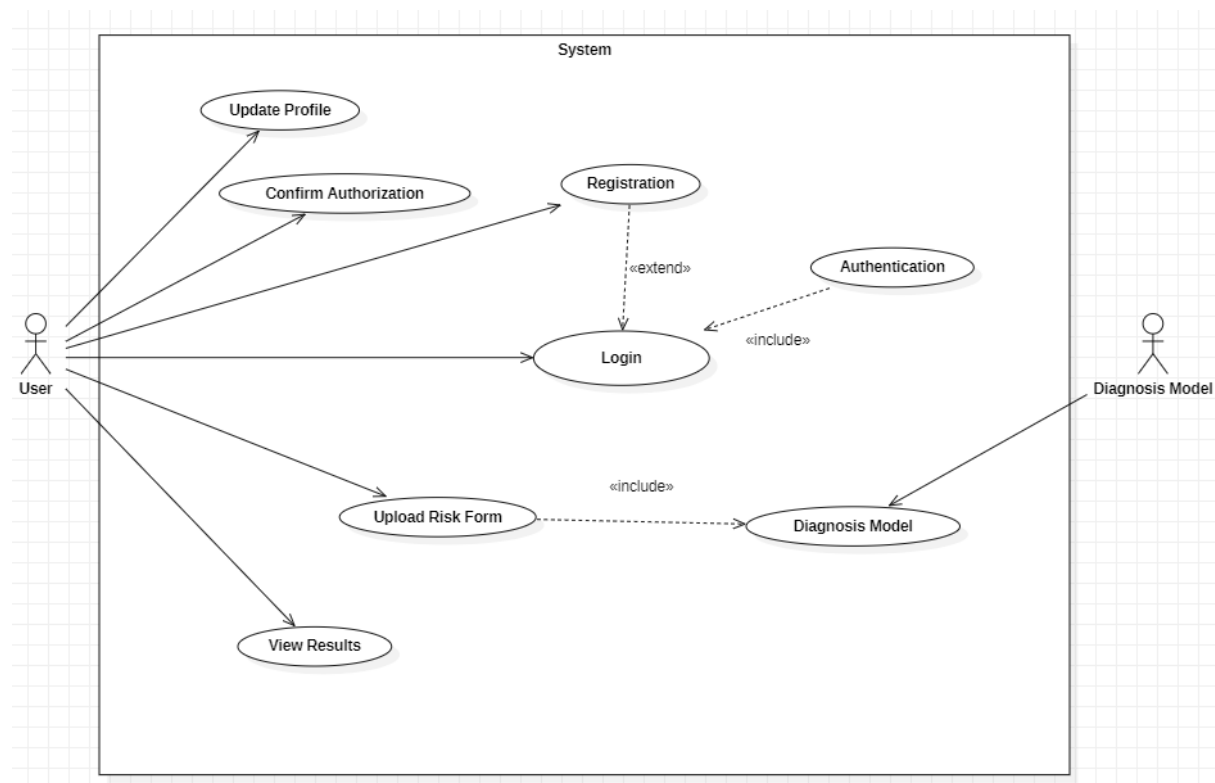


Figure 4.1 Use Case Diagram

4.3.2 Sequence Diagram

The Figure 4.2 below represents a sequence diagram. It illustrates the sequence of messages between objects in an interaction within the diagnosis prediction system. In this interaction, once a user is able to log in, they are able to fill in a form containing several risk factors. If all variables are filled correctly, it deems the form as valid. Once uploaded, the variables are processed by the machine learning module, the results are stored, and relayed back to the user.

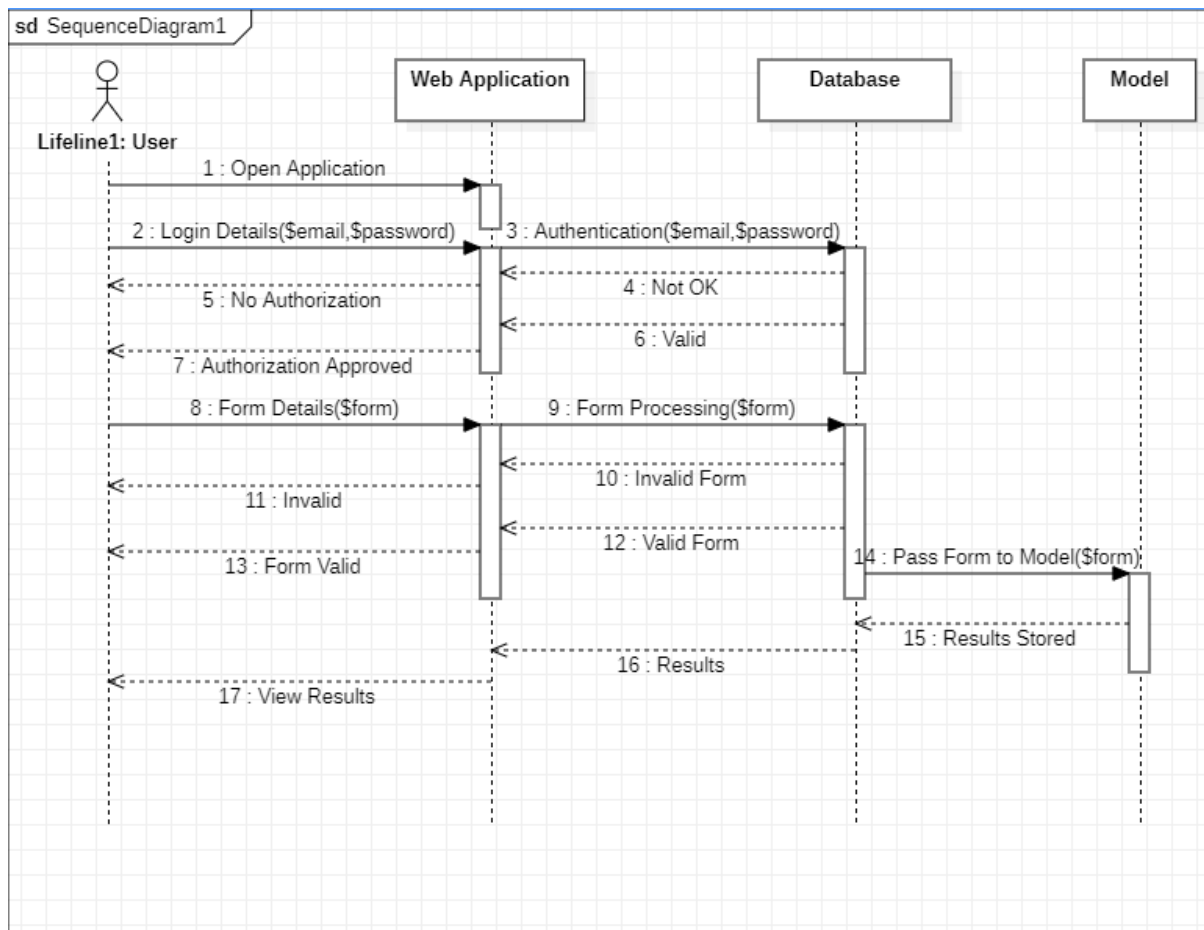


Figure 4.2 Sequence Diagram

4.3.3 Class Diagram

Figure 4.3 shows the interaction of all the system classes and attributes. It was used to describe the structure of the system by showing the system's classes, their attributes, operations, and relationships.

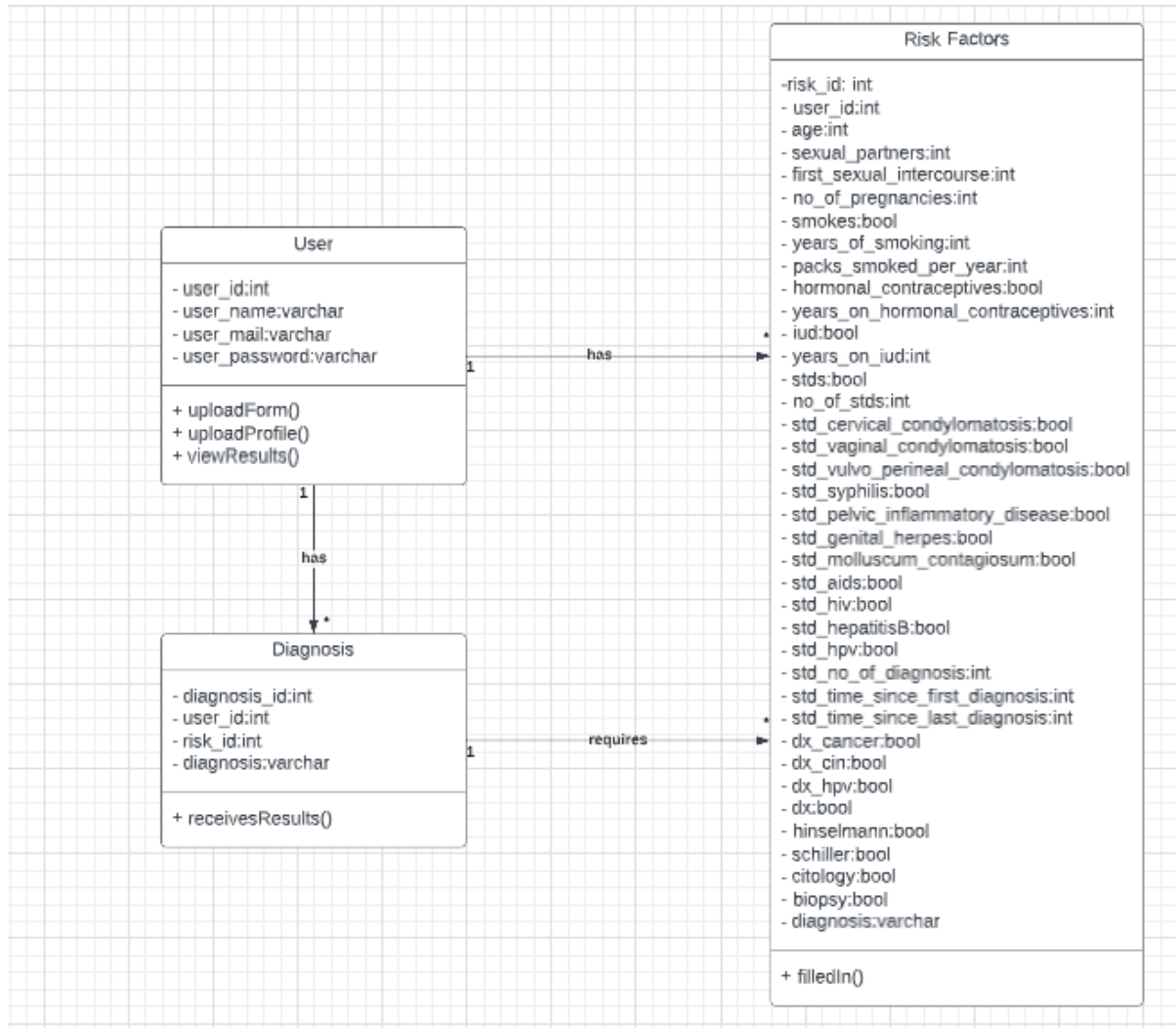


Figure 4.3 Class Diagram

4.3.4 Activity Diagram

Figure 4.4 describes the flow of activities from one to another. It is also described as the operation of a system. The diagram below shows the flow from when the user opens the application, through the use of the application where the model is concerned, all the way to the session logging out.

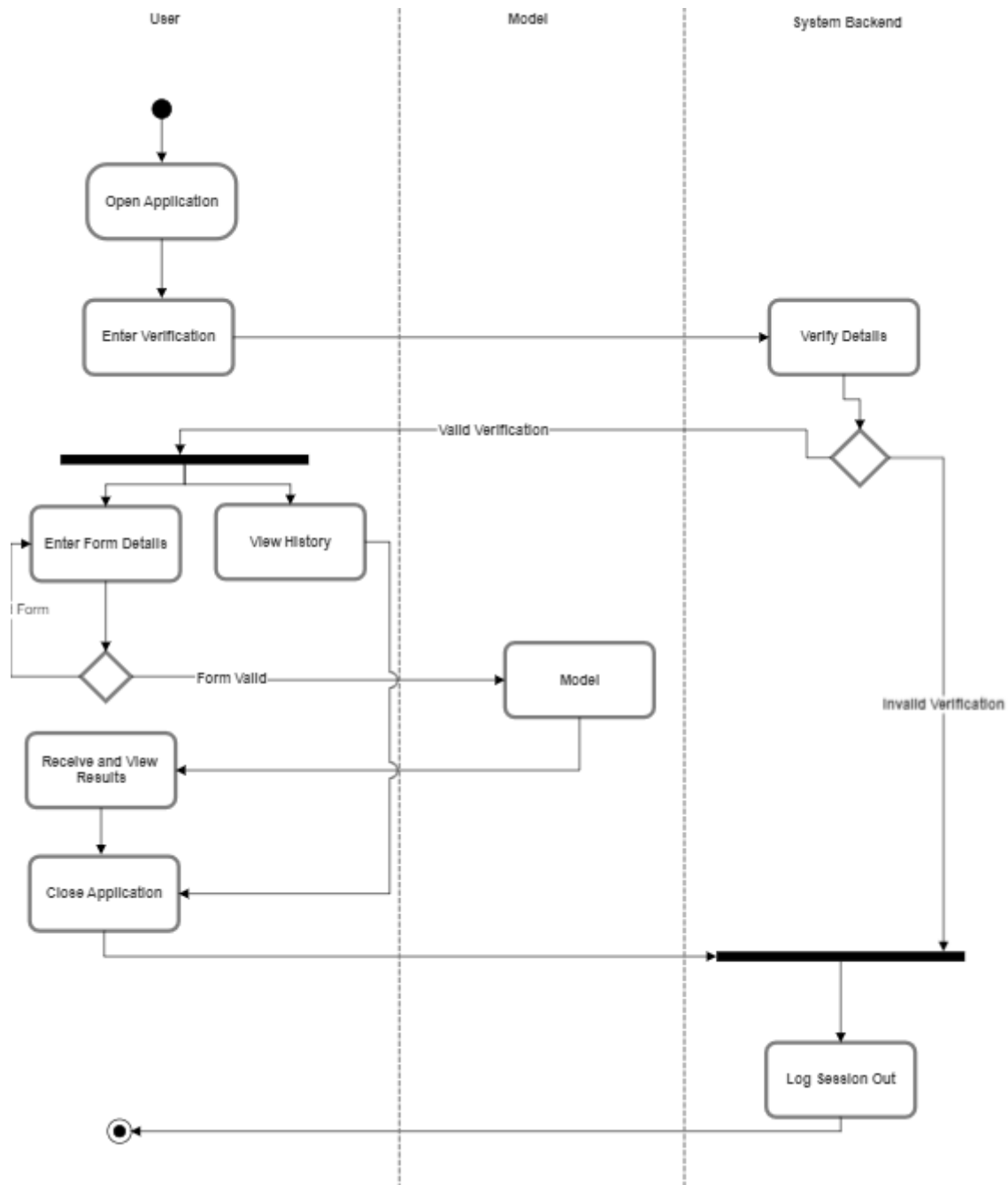


Figure 4.4 Activity Diagram

4.4 System Design Diagrams

Some of the system design diagrams considered are as follows:

4.4.1 Database Schema

Figure 4.5 represents the entities and the fields within the database as well as their primary, foreign keys and relationships. The 'Patients' table contains information about all the main users of the system. These users are uniquely identified by the 'user_id' which is the primary key of the table, their name, password stored in its hashed version, and their email. Each entry of a form of risk factors is stored in the 'Risk Factors' table which has the 'risk_id' as the primary key, 'user_id' and 'diagnosis' as the foreign keys. The prediction by the machine learning model is stored in the 'Diagnosis' table with 'diagnosis_id' as its primary key.

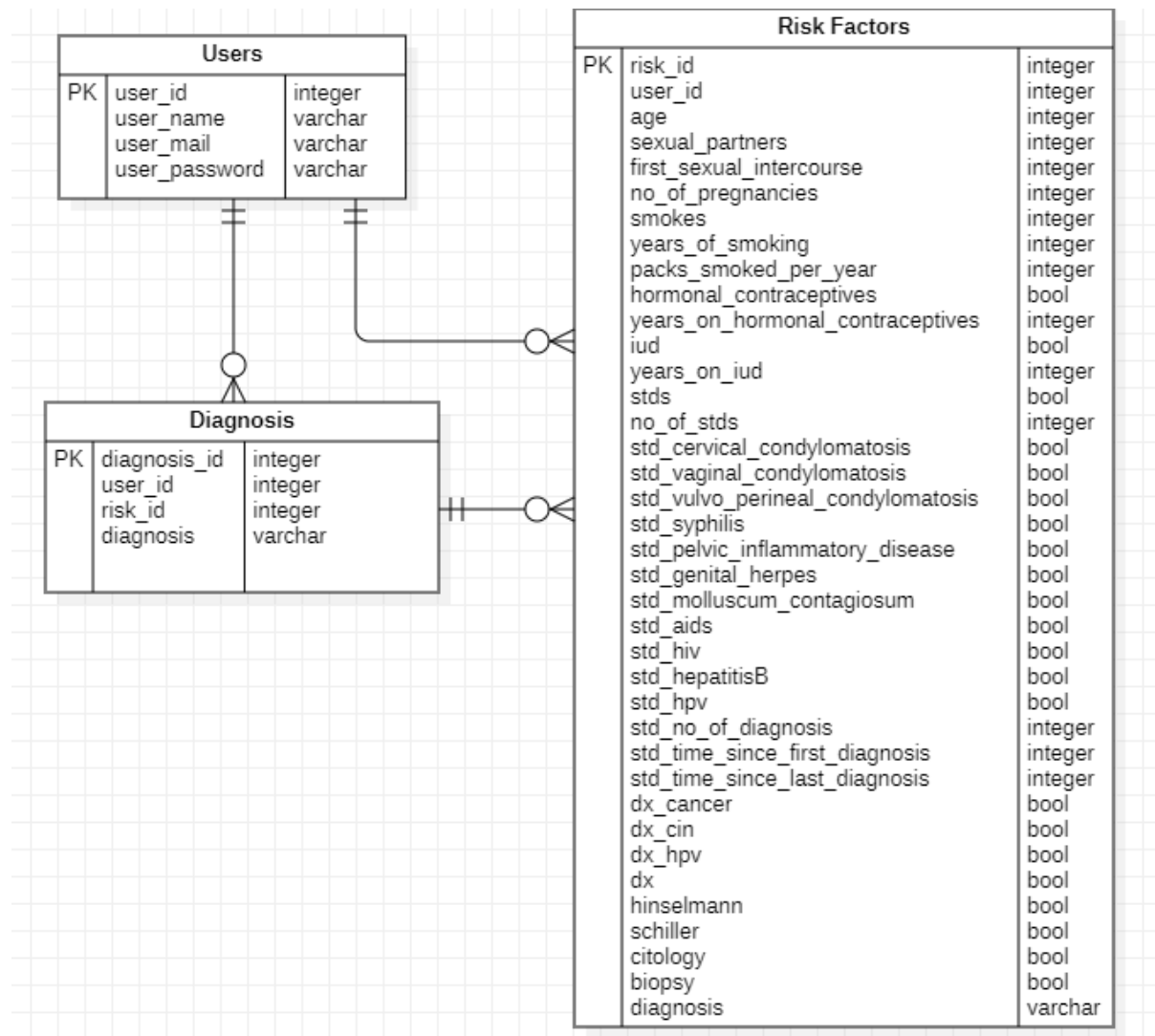


Figure 4.5 Database Schema

4.4.2 Wireframes/Mockups

The wireframes show how the user will interact with the system.

Figure 4.6 shows the homepage. This page shows when the user opens the application. From here, they are able to navigate to the login page.

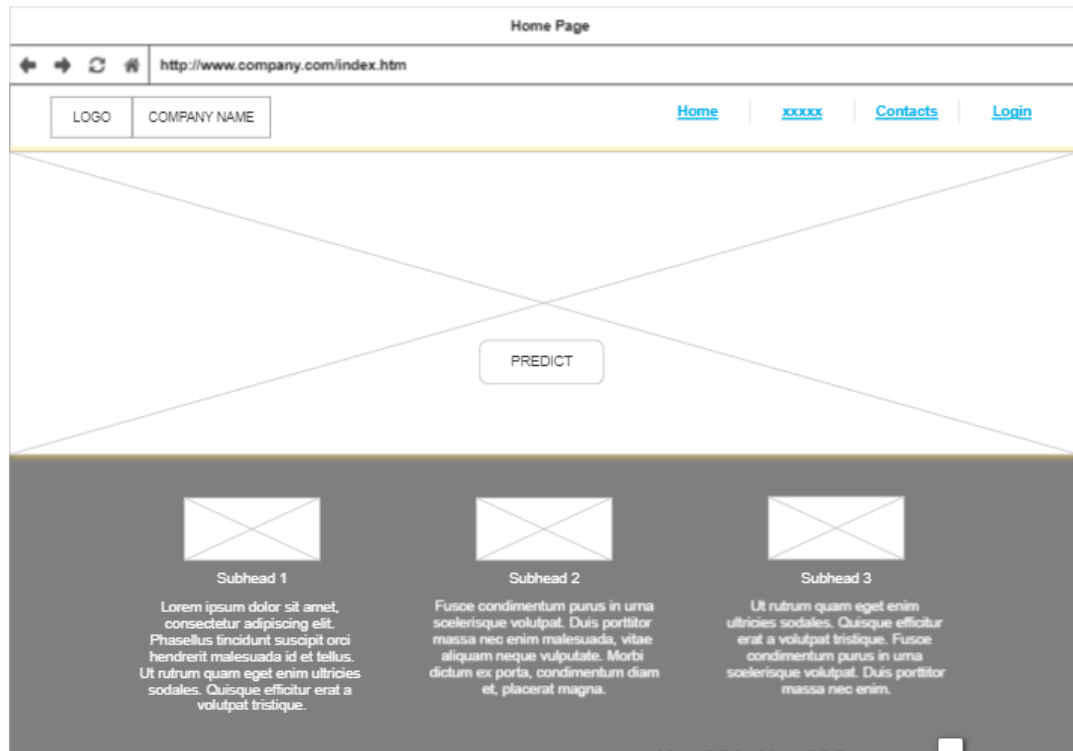
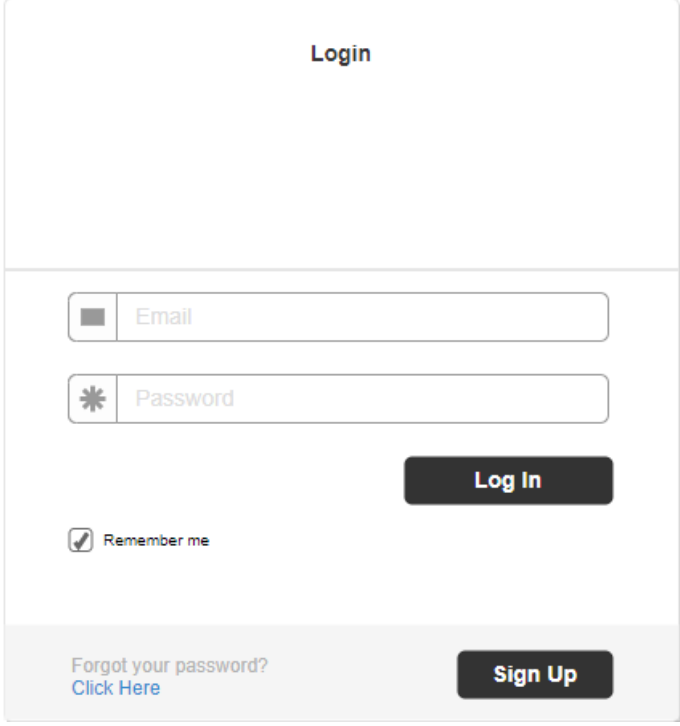


Figure 4.6 Homepage Wireframe

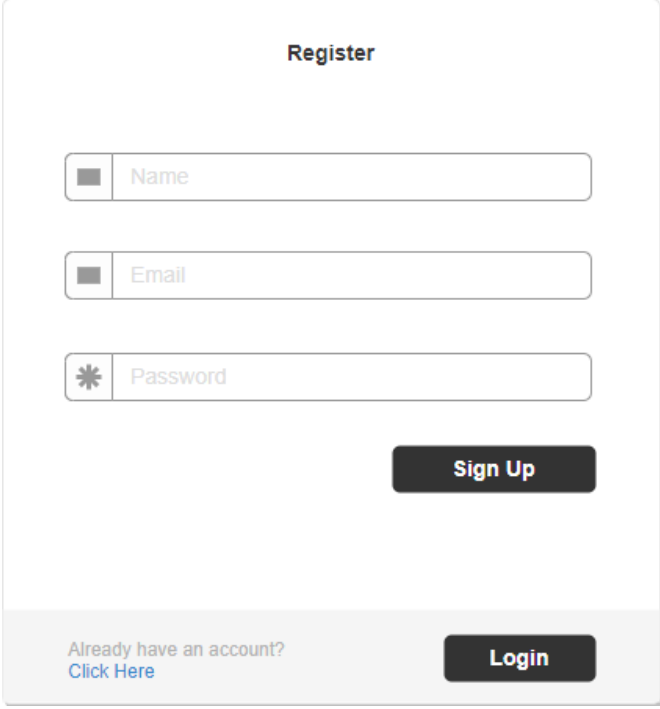
Figure 4.7 is the login wireframe.



The login wireframe is a rectangular box with a light gray background. At the top center, the word "Login" is displayed in a bold, black font. Below this, there are two input fields: the first is labeled "Email" and the second is labeled "Password". The "Password" field has a small asterisk icon to its left. To the right of the "Password" field is a dark gray button with the text "Log In" in white. Below the "Email" field, there is a checkbox with a checkmark icon and the text "Remember me". At the bottom of the wireframe, there is a light gray footer area. On the left side of this footer, the text "Forgot your password?" is followed by a blue link "Click Here". On the right side of the footer, there is a dark gray button with the text "Sign Up" in white.

Figure 4.7 Login Wireframe

Figure 4.8 is the register wireframe.



The register wireframe is a rectangular box with a light gray background. At the top center, the word "Register" is displayed in a bold, black font. Below this, there are three input fields: the first is labeled "Name", the second is labeled "Email", and the third is labeled "Password". The "Password" field has a small asterisk icon to its left. To the right of the "Password" field is a dark gray button with the text "Sign Up" in white. At the bottom of the wireframe, there is a light gray footer area. On the left side of this footer, the text "Already have an account?" is followed by a blue link "Click Here". On the right side of the footer, there is a dark gray button with the text "Login" in white.

Figure 4.8 Register Wireframe

Figure 4.9 is the form wireframe. Users input the variables of this form so as to predict their diagnosis.

The wireframe shows a web browser window titled "Form" with the URL "http://www.company.com/form.htm". The header contains a "LOGO" and a "COMPANY NAME" field, and a navigation bar with links: "Home", "xxxxxx", "Company", and "More". The main content area is titled "Please Fill in the Form Below" and contains ten questions arranged in two columns. Questions 1, 3, 5, 7, and 9 are on the left; Questions 2, 4, 6, 8, and 10 are on the right. Questions 1, 3, 5, 7, and 9 have radio button options (Option 1, Option 2). Questions 2, 4, 6, 8, and 10 have checkbox options (Option 1, Option 2). Questions 2, 4, 6, 8, and 10 also have a "Text Placeholder" input field. A "Predict" button is located at the bottom right of the form area.

Figure 4.9 Form Wireframe

Figure 4.10 is the prediction results wireframe. The result of the model is shown here.

The wireframe shows a web browser window titled "Results" with the URL "http://www.company.com/gateway.htm". The header contains a "LOGO" and a "COMPANY NAME" field, and a navigation bar with links: "Home", "xxxxx", "Company", and "Logout". The main content area is titled "Prediction Results" and contains a paragraph of placeholder text: "Lorem ipsum dolor sit amet, consectetur adipiscing elit. Phasellus tincidunt augue suscipit orci hendrerit Quisque efficitur erat a volutpat tristique. Fusce condimentum purus in urna scelerisque volutpat. Duis porttito." Below the text is a large white rectangular area with a black 'X' drawn across it, indicating a placeholder for a result or image. At the bottom right of the page, the coordinates "X= 946 Y= 333" are displayed.

Figure 4.10 Results Wireframe

4.4.3 System Architecture

The system architecture serves as a blueprint for the system. The model was integrated with Flask, deployed on a server and can be accessed using desktop computers.

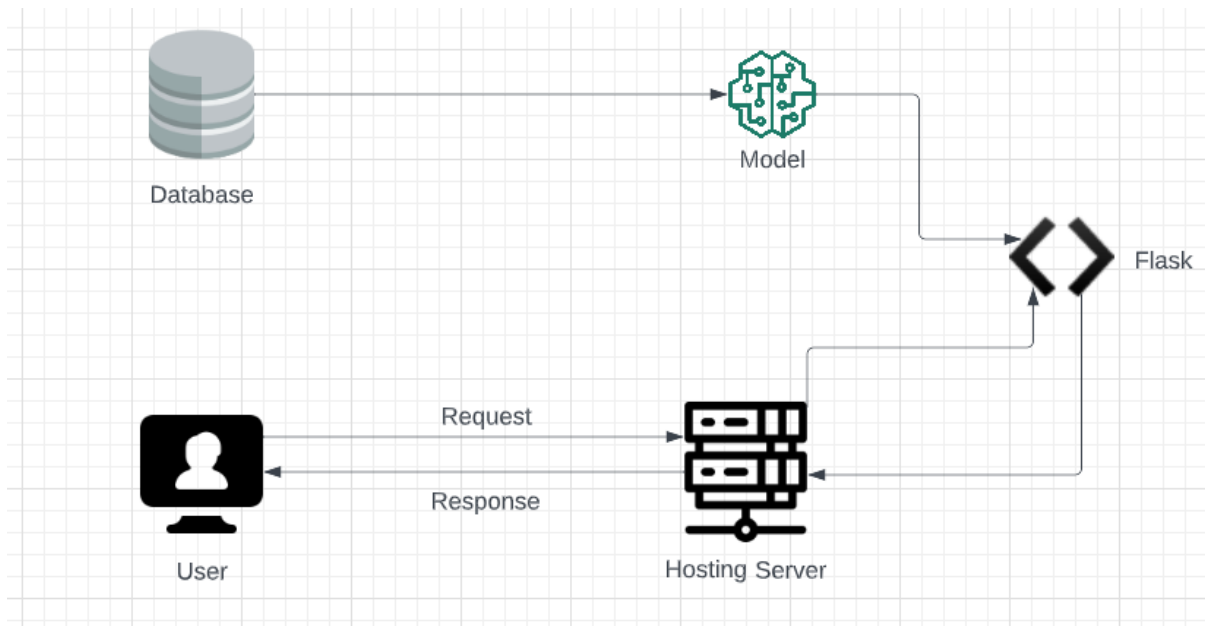


Figure 4.11 System Architecture

Chapter 5: System Implementation and Testing

5.1 Introduction

This chapter gives an in depth description of model and system implemented and tested for the developed solution. It defines the implementation environment: discussing the requirements needed for the system to run on a user machine. Furthermore, analysis of the model, from the dataset, its preparation, training, and finally testing is covered. The developed solution is the prediction model. The model is deployed on Flask and interacts with an SQLite database.

5.2 Description of the Implementation Environment

In this section the hardware and software requirements that were necessary for the development and implementation of the developed system are outlined. These are the optimum conditions for the system to be fully operational.

5.2.1 Hardware Specifications

To determine the hardware requirements, I examined what aspects must be considered to ensure that the system functions as intended. It is also paramount that the hardware used should properly accommodate the functionalities of a web server.

Table 5.1 Hardware Requirements

Item	Minimal Specifications	Justification
Processor	4 x 1.6 GHz CPU	A decent processor allows faster data loading.
RAM	12 GB RAM	A lower RAM will lead to frequent crashing of the program.
Hard Disk Storage	50 GB of free space on the drive is recommended	This is sufficient for the computer to work with the system comfortably.

5.2.2 Software Specifications

The software requirements a user needs to access the system on a web browser are as follows.

Table 5.2 Software Requirements

Item	Recommended Specifications	Justification
Operating System	Windows 7 and above	Any operating system below Windows 7 would not be supported well.
Web browser	Chrome or equivalent	The system would be accessed via the web.

5.3 Description of the Dataset

The dataset that was used in this analysis is the ‘Cervical Cancer Risk Classification’ dataset which is a dataset that was collected at ‘Hospital Universitario de Caracas’ in Caracas, Venezuela comprising of demographic information, habits, and medical records of 858 patients.

The dataset contains 36 features that represent cervical cancer risk. Four of the 36 features are categorical in nature. These values are the result of medical tests performed to confirm the diagnostic result on cervical cancer. The Hinselmann’s test is done to check if lesions are cancerous or not. With the Schiller’s test, a part of the body under observation is painted with a solution to investigate malignant nature of the body part. The Cytology test is done to help ascertain if there are cancerous fluids in the body part. Finally, a Biopsy test is done when most clinical options are exhausted and only a biopsy can reveal the persons state of health (Lilhore et al., 2022).

Figure 5.1 describes the variables that were used in constructing the cervical cancer diagnosis model.

Attribute Information:

(int) Age
(int) Number of sexual partners
(int) First sexual intercourse (age)
(int) Num of pregnancies
(bool) Smokes
(bool) Smokes (years)
(bool) Smokes (packs/year)
(bool) Hormonal Contraceptives
(int) Hormonal Contraceptives (years)
(bool) IUD
(int) IUD (years)
(bool) STDs
(int) STDs (number)
(bool) STDs:condylomatosis
(bool) STDs:cervical condylomatosis
(bool) STDs:vaginal condylomatosis
(bool) STDs:vulvo-perineal condylomatosis
(bool) STDs:syphilis
(bool) STDs:pelvic inflammatory disease
(bool) STDs:genital herpes
(bool) STDs:molluscum contagiosum
(bool) STDs:AIDS
(bool) STDs:HIV
(bool) STDs:Hepatitis B
(bool) STDs:HPV
(int) STDs: Number of diagnosis
(int) STDs: Time since first diagnosis
(int) STDs: Time since last diagnosis
(bool) Dx:Cancer
(bool) Dx:CIN
(bool) Dx:HPV
(bool) Dx
(bool) Hinselmann: target variable
(bool) Schiller: target variable
(bool) Cytology: target variable
(bool) Biopsy: target variable

Figure 5.1 Risk Factors in the Dataset

In summary, the dataset contains of information about lifestyle habits, sexual behaviors and last but not least, the outcome of the medical tests.

Using Python, NumPy, Pandas, Seaborn and Matplotlib, data visualization and analysis was carried out which revealed that most of the values in the set are Boolean in type, the dataset had a lot of empty values and that it was imbalanced.

Missing data occurred because several patients decided not to answer some questions because of privacy concerns. Figure 5.2 shows the extent of the missing data indicating that almost all variables had missing data.

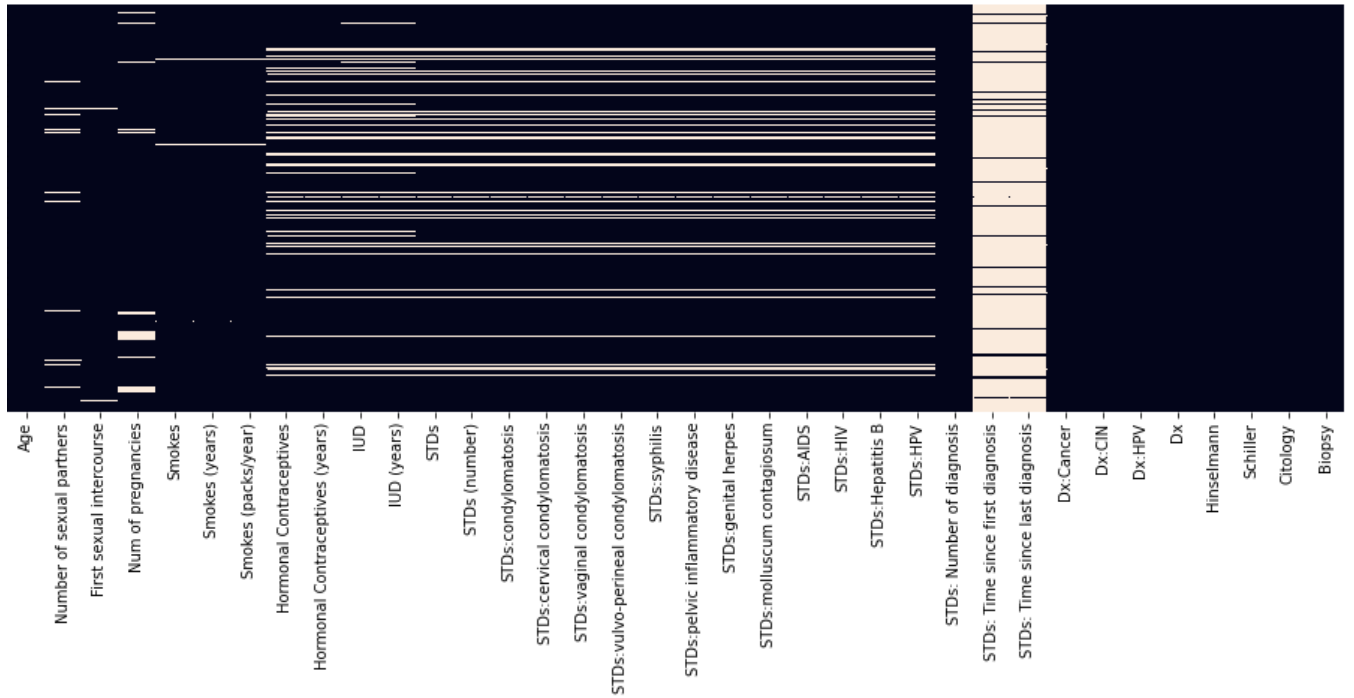


Figure 5.2 Missing Data

Variables that had too much missing data that is: ‘STDs: Time since first diagnosis’ and ‘STDs: Time since last diagnosis’ were dropped. Imputation techniques were used on the other variables with missing values. Imputation is the process of estimating a missing value based on other valid values (Anand-Kumar, 2015). The null values were replaced with their mean. Three other variables, ‘Smokes’, ‘Hormonal Contraceptives’ and ‘IUD’ were dropped. This is because their corresponding columns have a non-zero only if they are non-zero.

Having the final variables, they needed to be split into train, test, and validation sets. Stratified sampling was used. Unlike the Random Sampling method, this method ensured that the distribution of classes in each train, validation, and test set is preserved (How Stratified Random Sampling Works, with Examples, 2022). Because of this the machine learning model was trained and validated on the same data distribution. The dataset was split 70/30 into the train and test data. After, the test data was split 50/50 into test and validation data.

As stated earlier, the dataset is heavily unbalanced. A dataset is unbalanced when at least one class is represented by only a small number of training examples while the other classes make up the majority. This imbalance leads to the class imbalance problem which occurs when the majority class greatly outnumbers the minority class observation. In this case, the target variable to be used, ‘Biopsy’ has an imbalance of 803 negative observations to 55 positive observations. This imbalance was dealt with by using the SMOTE Technique (Synthetic

Minority Over-sampling TEchnique). SMOTE Technique selects examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line. This procedure is used to create as many synthetic examples in the minority class as required (Brownlee, 2020).

5.4 Description of Training

Following pre-processing steps, the building of the prediction model from the training set was carried out. I used eXtreme Gradient Boosting(XGBoost) algorithm to build the model. It is a decision tree ensemble Machine learning algorithm that uses a gradient boosting library (Morde, 2019). It is a type of gradient boosting that learns from previous iterations of itself. XGBoost has several tuning parameters. After the initial setting of parameters, and first training iteration, Figure 5.3 is the train and validation AUC as the number of trees increased.

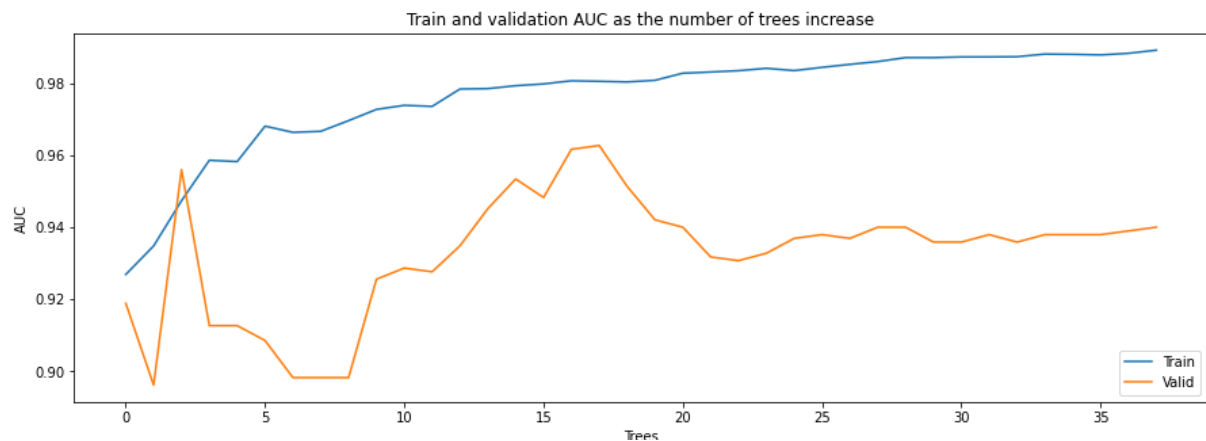


Figure 5.3 Initial Train and Validation AUC as number of trees increase

Area Under the Curve(AUC) is the valued metric used for evaluating the performance in classification models. The metric helps determine the capability of a model in distinguishing classes. The judging criteria is- the higher the AUC, the better the model (Dey, 2021). Keeping that in mind, there was a possibility that performance could increase if hyperparameters were tuned.

GridSearchCV was then used for hyperparameter tuning so as to attempt to increase performance. It tries all possible combinations of hyperparameters specified and gives result of model performance on validation set using cross validation technique (Sklearn.Model_selection.GridSearchCV—Scikit-Learn1.1.3Documentation,2016).

GridSearchCV helped identify the best set of parameters to use. The final parameters used were: learning_rate – weightage of every tree in xgboost classifier, which was set to 0.001, max_depth – max depth for each tree in the model, which was set to 20, n_estimators- the

maximum number of trees to be created, which was set to 1000, subsample - % of observations of trained dataset randomly selected for each individual tree in the model, which was set to 0.5, colsample_bytree – features used every time randomly when new tree is built, which was set to 0.3, and eval_metric- evaluation metric using area under the curve, which was set to ‘auc’.

After several iterations and training with these parameters, I achieved the final model. Figure 5.4 shows a significant improvement in the train and validation AUC as the number of trees increased.

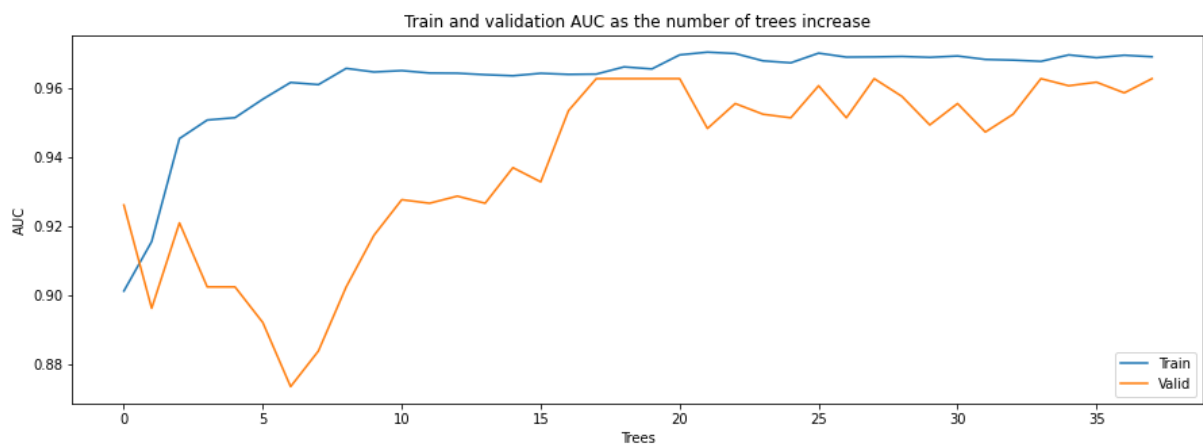


Figure 5.4 Final Model Train and Validation AUC

I saved the final model using Pickle. Pickle is a useful Python tool that allows saving of an ML model, to minimize lengthy re-training. To save the model, I passed the model object into the dump() function of Pickle. Figure 5.5 shows the code used to save the models.

```
import pickle
filename='cancermodel.sav'
pickle.dump(cervical_cancer_xgboost_final, open(filename,'wb'))
```

Figure 5.5 Saving the model

5.5 Description of Testing

Following the acquiring of the final model, I passed in the testing dataset containing risk factors to predict the cervical cancer diagnosis. The results of the predictions by the model were displayed in a confusion matrix and a classification report to determine its performance. Figure 5.7 below shows the confusion matrix and Figure 5.7 shows the classification report.

```
Confusion Matrix:
[[198  3]
 [ 3 11]]
```

Figure 5.6 Confusion matrix

```
Classification Report:
              precision    recall  f1-score   support

     0       0.99         0.99         0.99         201
     1       0.79         0.79         0.79          14

 accuracy          0.97         0.97         0.97         215
 macro avg       0.89         0.89         0.89         215
 weighted avg    0.97         0.97         0.97         215
```

Figure 5.7 Classification Report

The performance metrics used in the classification report were- precision: which attempts to determine what proportion of positive identifications were actually correct, recall: which attempts to determine the proportion of actual positives that were identified correctly, and F1 score: being the weighted average of precision and recall.

5.5.1 Testing Paradigm

The testing paradigms used were white box and black box testing.

White box testing is a testing technique in which software's internal structure, design, and coding are tested to verify input-output flow and improve design, usability, and security. In white box testing, the code is visible to testers (Hamilton, 2020). This was done during training of the model. The weights and performance metrics were used to improve the performance of the model.

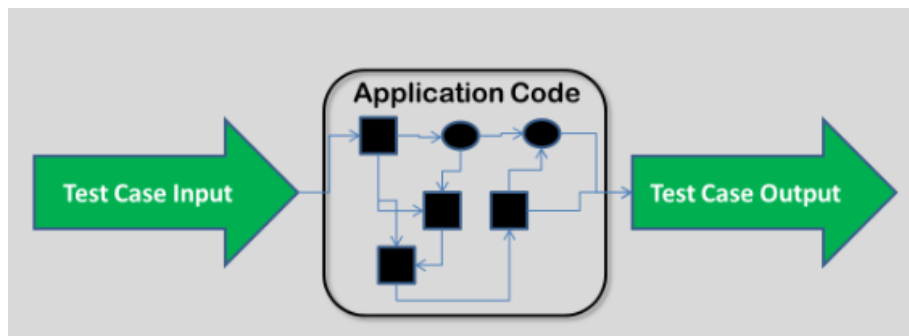


Figure 5.8 White Box Testing

Black box testing is a testing technique with no knowledge about the internal details of the model, such as the algorithm used to create it and its features. The main objective is to ensure the quality of the models (Functional Testing of Machine Learning Models, 2018). This was done in the prediction stage where the saved model was used to predict on the test dataset. The test dataset was fed to the model without any modification to the model. Moreover, black box testing was also used on the web application hosting the model.

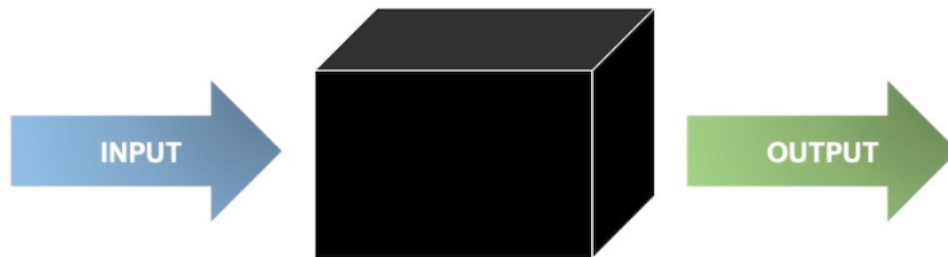


Figure 5.9 Black Box Testing

5.5.2 Testing Results

The following are the results of testing the model and core functionalities of the system.

Test Case	Description	Test Data	Result	Test Verdict
Raw data input	Raw data input prediction on model.	Age:38 Number of sexual partners:2 First sexual intercourse:15 Number of pregnancies:4 Years smoking:0 Packs per year:0 Hormonal Contraceptives:0 Years on IUD:16 STDs:0 Number of STDs:0 STDs- condylomatosis:0 STDs-vaginal condylomatosis:0 STDs-vulvo perineal:0 STDs-syphilis:0 STDs-pelvic inflammatory disease:0 STDs-genital herpes:0 STDs-molluscum contagiosum:0 STDs-HIV:0 STDs-Hepatitis B:0	Prediction is positive: 1	Pass.

		STDs-HPV:0 Number of STD diagnosis:0 Dx-Cancer:1 Dx-Cin:0 Dx-HPV:1 DX Test:0 Hinselmann Test:0 Schiller Test:1 Citology Test:0		
Raw data input	Raw data input prediction on model.	Age:18 Number of sexual partners:4 First sexual intercourse:15 Number of pregnancies:1 Years smoking:0 Packs per year:0 Hormonal Contraceptives:0 Years on IUD:0 STDs:0 Number of STDs:0 STDs- condylomatosis:0 STDs-vaginal condylomatosis:0 STDs-vulvo perineal:0 STDs-syphilis:0 STDs-pelvis inflammatory disease:0 STDs-genital herpes:0 STDs-molluscum contagiosum:0 STDs-HIV:0 STDs-Hepatitis B:0 STDs-HPV:0 Number of STD diagnosis:0 Dx-Cancer:0 Dx-Cin:0 Dx-HPV:0 DX Test:0 Hinselmann Test:0 Schiller Test:0 Citology Test:0	Prediction is negative: 0	Pass.
Registration providing all fields	A new user provides required fields	First Name-Jane Last Name-Doe Email-jane@gmail.com Password-111	Details of user are saved in the database.	Pass.
Registration without providing	A new user registers	First Name- Last Name-Doe Email-jane@gmail.com	Display of error message	Pass.

required details.	without filling all required fields	Password-111		
Login with correct credentials.	A user tries to login with the correct credentials.	Email-jane@gmail.com Password-111	The user logs in and is redirected to the home page.	Pass.
Filling risk form for prediction.	User fills in the variables for to get diagnosis prediction.	Filling in the form.	Form is uploaded and prediction displayed to the user.	Pass.
Viewing history of the system	The user tries to view his recent uploads and their predictions.	Email-jane@gmail.com Password-111	Prediction history of user is displayed	Pass.
Logging out.	The user tries to logout by pressing the logout button.	Logout button is clicked.	The user is logged out and redirected to the index page.	Pass

Figure 5.10 Testing results

Chapter 6: Conclusions, Recommendations and Future Works

6.1 Conclusion

Cervical cancer has a high mortality rate. This is due to its late detection associated with the lack of accessibility, affordability, and widespread cultural unacceptability of the pap smear test. The projects' main aim was to develop a web-based machine learning model that analyzes risk factors; including age and habits, in order to predict one's cervical cancer diagnosis.

The project presented a dataset to show whether one has cervical cancer or not. The data was analyzed, and a model developed using the XGBoost algorithm.

The problem has been solved in a better way because the solution is patient oriented unlike other solutions that are more doctor centered. Moreover, unlike the HPV DNA and pap-smear tests, the model is a non-intrusive procedure that aims to work as a preventive and predictive measure.

6.2 Recommendations

It is recommended that to fully implement and improve the precision, recall and F1 score of the developed solution, a larger balanced dataset, and a computer with a dedicated GPU will be required to further train the model.

6.3 Future Works

In case the developed solution is to be built upon; further experiments can be carried out. This includes the exclusion of Schillers, Hinselmann, and Cytology tests as features and having Biopsy as the sole target. This is because each of these variables are a result of a test carried out to determine the presence of cervical cancer which can be done only using Biopsy.

References

- Anand-Kumar, V. (2015). *Chapter 2 What is Imputation? / Applications of Machine Learning in Imputation*. https://bookdown.org/v_anandkumar88/docs2/what-is-imputation.html
- Brownlee, J. (2020, January 16). SMOTE for Imbalanced Classification with Python. *MachineLearningMastery.Com*. <https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/>
- Cervical Cancer Risk Classification*. (2018). <https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification>
- Cervical Cancer Screening*. (2017). [Text]. National Library of Medicine. <https://medlineplus.gov/cervicalcancerscreening.html>
- Cervical Cancer Screening*. (2021). <https://www.acog.org/en/womens-health/faqs/cervical-cancer-screening>
- Cervical cancer screening at a tertiary care center in Rwanda—ScienceDirect*. (2017). <https://www.sciencedirect.com/science/article/pii/S2352578917300565?via%3Dihub>
- Cervical Cancer Screening (PDQ®)—Patient Version—NCI* (nciglobal,ncienterprise). (2004, September 25). [PdqCancerInfoSummary]. <https://www.cancer.gov/types/cervical/patient/cervical-screening-pdq>
- Chirenje, Z. M., Rusakaniko, S., Kirumbi, L., Ngwalle, E. W., Makuta-Tlebere, P., Kaggwa, S., Mpanju-Shumbusho, W., & Makoe, L. (2001). Situation analysis for cervical cancer diagnosis and treatment in east, central and southern African countries. *Bulletin of the World Health Organization*, 79(2), 127–132.
- Definition of pelvic exam—NCI Dictionary of Cancer Terms—NCI* (nciglobal,ncienterprise). (2011, February 2). [NciAppModulePage]. <https://www.cancer.gov/publications/dictionaries/cancer-terms/def/pelvic-exam>
- Dey, V. (2021, September 5). *Understanding the AUC-ROC Curve in Machine Learning Classification*. Analytics India Magazine. <https://analyticsindiamag.com/understanding-the-auc-roc-curve-in-machine-learning-classification/>
- Fontham, E. T. H., Wolf, A. M. D., Church, T. R., Etzioni, R., Flowers, C. R., Herzig, A., Guerra, C. E., Oeffinger, K. C., Shih, Y.-C. T., Walter, L. C., Kim, J. J., Andrews, K.

- S., DeSantis, C. E., Fedewa, S. A., Manassaram-Baptiste, D., Saslow, D., Wender, R. C., & Smith, R. A. (2020). Cervical cancer screening for individuals at average risk: 2020 guideline update from the American Cancer Society. *CA: A Cancer Journal for Clinicians*, 70(5), 321–346. <https://doi.org/10.3322/caac.21628>
- Functional Testing of Machine Learning Models*. (2018). Qualcomm Developer Network. <https://developer.qualcomm.com/software/qualcomm-neural-processing-sdk/learning-resources/training-testing-machine-learning-models/functional-testing-machine-learning-models>
- Glen, S. (2019, July 28). *Decision Tree vs Random Forest vs Gradient Boosting Machines: Explained Simply* - *DataScienceCentral.com*. Data Science Central. <https://www.datasciencecentral.com/decision-tree-vs-random-forest-vs-boosted-trees-explained/>
- Gupta, A. (2021, June 1). XGBoost versus Random Forest. *Geek Culture*. <https://medium.com/geekculture/xgboost-versus-random-forest-898e42870f30>
- Hamilton, T. (2020, February 17). *White Box Testing – What is, Techniques, Example & Types*. <https://www.guru99.com/white-box-testing.html>
- How Stratified Random Sampling Works, with Examples*. (2022). Investopedia. https://www.investopedia.com/terms/stratified_random_sampling.asp
- Kivuti-Bitok, L. W., Pokhariyal, G. P., Abdul, R., & McDonnell, G. (2013). An exploration of opportunities and challenges facing cervical cancer managers in Kenya. *BMC Research Notes*, 6(1), 136. <https://doi.org/10.1186/1756-0500-6-136>
- Lee, S. (2014). *Benefits and limitations of screening for cervical cancer*. Canadian Cancer Society. <https://cancer.ca/en/cancer-information/find-cancer-early/get-screened-for-cervical-cancer/benefits-and-limitations-of-screening-for-cervical-cancer>
- Lilhore, U. K., Poongodi, M., Kaur, A., Simaiya, S., Algarni, A. D., Elmannai, H., Vijayakumar, V., Tunze, G. B., & Hamdi, M. (2022). Hybrid Model for Detection of Cervical Cancer Using Causal Analysis and Machine Learning Techniques. *Computational and Mathematical Methods in Medicine*, 2022, e4688327. <https://doi.org/10.1155/2022/4688327>

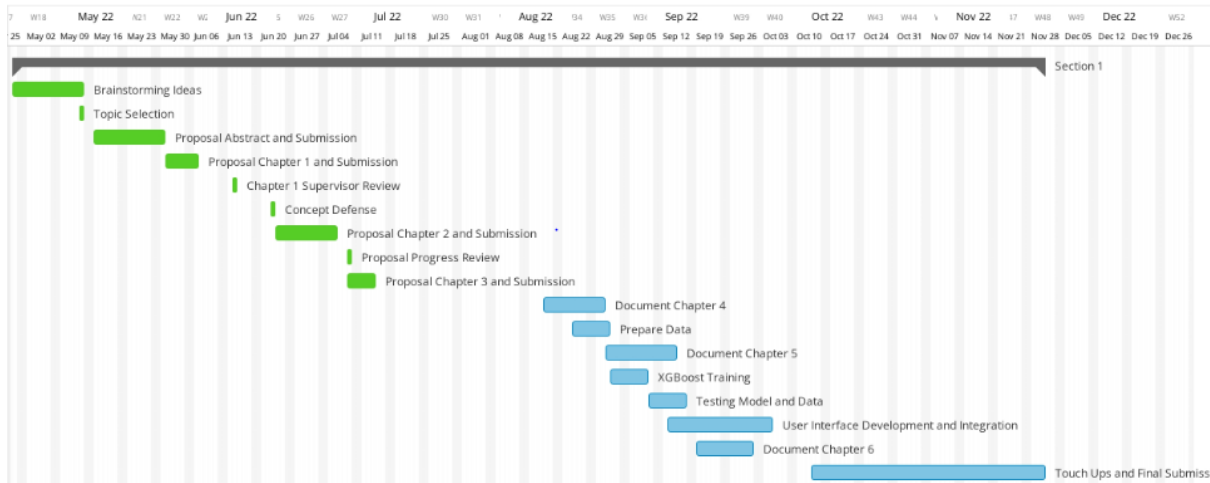
- Martin, M. (2020, February 20). *Prototype Model in Software Engineering*.
<https://www.guru99.com/software-engineering-prototyping-model.html>
- Mehta, V., Vasanth, V., & Balachandran, C. (2009). *Pap smear*. *Indian Journal of Dermatology, Venereology and Leprology*, 75(2), 214. <https://doi.org/10.4103/0378-6323.48686>
- Morde, V. (2019, April 8). *XGBoost Algorithm: Long May She Reign!* Medium.
<https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d>
- [PDF] *Cervical Cancer: Machine Learning Techniques for Detection, Risk Factors and Prevention Measures / Semantic Scholar*. (2020).
<https://www.semanticscholar.org/paper/Cervical-Cancer%3A-Machine-Learning-Techniques-for-Diaz-Ccopa/4bfcd073de6487f945fedeb39f2fac2b53d3aec>
- Poverty in Kenya*. (2018). The Borgen Project. <https://borgenproject.org/tag/poverty-in-kenya/>
- Prediction models in cancer care—Vickers—2011—CA: A Cancer Journal for Clinicians—Wiley Online Library*. (2011).
<https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.20118>
- Priyanka, B. J. (2021). *Machine Learning Approach for Prediction of Cervical Cancer*. 9.
- Rajasekharan, D., Kumar, R., Balakrishnan, B., Sharathkumar, P., Chandran, P., ss, S., Bengtsson, E., & Sujathan, K. (2015). Computer Assisted Pap Smear Analyser for Cervical Cancer Screening using Quantitative Microscopy. *J Cytol Histol*, 5:3.
<https://doi.org/10.4172/2157-7099.S3-010>
- Reaching 2030 cervical cancer elimination targets—New WHO recommendations for screening and treatment of cervical pre-cancer*. (2021). <https://www.who.int/news-room/events/detail/2021/07/06/default-calendar/reaching-2030-cervical-cancer-elimination-targets>
- Recommendations on screening for cervical cancer. (2013). *CMAJ: Canadian Medical Association Journal*, 185(1), 35–45. <https://doi.org/10.1503/cmaj.121505>
- Rosser, J. I., Hamisi, S., Njoroge, B., & Huchko, M. J. (2015). Barriers to Cervical Cancer Screening in Rural Kenya: Perspectives from a Provider Survey. *Journal of Community Health*, 40(4), 756–761. <https://doi.org/10.1007/s10900-015-9996-1>

- SDLC - Software Prototype Model.* (2020).
https://www.tutorialspoint.com/sdlc/sdlc_software_prototyping.htm
- Sklearn.model_selection.GridSearchCV — scikit-learn 1.1.3 documentation.* (2016).
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
- The Application of Machine Learning in Cervical Cancer Prediction | 2021 6th International Conference on Machine Learning Technologies.* (2021).
<https://dl.acm.org/doi/abs/10.1145/3468891.3468894>
- Tiruneh, F. N., Chuang, K.-Y., Ntenda, P. A. M., & Chuang, Y.-C. (2017). Individual-level and community-level determinants of cervical cancer screening among Kenyan women: A multilevel analysis of a Nationwide survey. *BMC Women's Health*, 17(1), 109. <https://doi.org/10.1186/s12905-017-0469-9>
- What Is Cancer? - NCI* (nciglobal,ncienterprise). (2007, September 17). [CgvArticle].
<https://www.cancer.gov/about-cancer/understanding/what-is-cancer>

Appendix

Appendix 1: Gantt Chart

The figure below shows a Gantt Chart displaying a list of activities that will be undertaken from the start to the end of the project.



Appendix 1 Gantt Chart