THE CO-OPERATIVE UNIVERSITY OF KENYA (CUK)

SCHOOL OF COMPUTING & MATHEMATICS PROJECT

PROJECT PROPOSAL

BREAST CANCER DIAGNOSIS PREDICTOR APPLICATION

BY

JOHNSON KANYI WAMWEYA

BCSC01/0019/2021

SUPERVISOR: *MS. EDNA CHEPKEMOI*

Project Proposal submitted in partial fulfillment of the requirements for the award of the
Bachelor of Science in Computer Science

©DECEMBER, 2024

DECLARATION AND APPROVAL

DECLARATION

I hereby declare that this Project Proposal is my own work and has, to the best of my knowledge, not been submitted to any other institution of higher learning for any award.

Student Name: Johnson Kanyi Wamweya   **Registration Number:** BCSC01/0019/2021

Signature: .... **Date:** ....

APPROVAL

This project Proposal has been submitted with my approval as the University supervisor.

Supervisor Name: Ms. Edna Chepkemoi

Signature: .... **Date:** ....

ABSTRACT

Breast cancer is the most frequent cancer in women, one in eight women worldwide, and continues to be the leading cause of cancer-related deaths among female patients. Despite the increase in diagnosis and treatment success rates, timely and efficient diagnosis remains a big concern because not all breast lesions are malignant, and a benign lesion does not always further develop into cancer. Traditional diagnostic approaches like physical examinations, mammography, and fine-needle aspiration cytology, though reliable, can be complemented by the use of machine learning technologies in improving diagnostic accuracy. This project deals with developing a web-based application for the prediction of breast cancer diagnosis using machine learning to classify breast masses as either benign or malignant based on diagnostic cell nuclei measurements from a breast tissue. The system will be implemented based on the publicly available, Breast Cancer Wisconsin (Diagnostic) Data Set from Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data). There are 569 samples, each with ten real-valued cell features: radius, texture, compactness etcetera. A logistic regression model is to be trained in predicting the malignancy likelihood for an optimal balance between precision and recall. The application, developed in Python and using the Streamlit framework, will provide a web-based interactive dashboard for medical professionals to enter the data collected from a breast sample tissue, to obtain predictions, accompanied by visualizations based on Plotly for intuitive understanding. The expected outcome is a very accurate, user-friendly tool that simplifies diagnosis. The project will demonstrate how ML can enhance diagnostic procedures, aligning with global health goals to reduce cancer mortality by supporting early detection and decision-making. By integrating technology into healthcare, this project bridges the gap between traditional diagnostics and modern AI-driven solutions, offering a scalable and accessible tool for clinical and educational use.

TABLE OF CONTENT

LIST OF ABBREVIATIONS

AI: Artificial Intelligence

BRCA1: Breast Cancer 1 Gene

BRCA2: Breast Cancer 2 Gene

CADx: Computer-Aided Diagnosis

CSS: Cascading Style Sheets

EDA: Exploratory Data Analysis

FNA: Fine Needle Aspiration

GDPR: General Data Protection Regulation

HIPAA: Health Insurance Portability and Accountability Act

IBM: International Business Machines

ML: Machine Learning

SDG: Sustainable Development Goals

SDLC: Software Development Life Cycle

UI: User Interface

UAT: User Acceptance Testing

UNSDGs: United Nations Sustainable Development Goals

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1- INTRODUCTION

1.1 Background to the Study

Breast cancer is one of the most common cancers causing death in women all over the world. Early detection and diagnosis of the disease is important in its management and improvement of survival rates among patients. Traditionally, diagnosis has been done by analysing a breast mass through clinical examination imaging and cytology, which can be very time-consuming and prone to human error. More recently, the emerging use of Artificial Intelligence and Machine Learning has made promising new tools available that will help medical professionals diagnose breast cancer more accurately and with greater efficiency. Machine learning models can analyse big datasets of medical information for patterns that might be too subtle for human observation. These models can then be trained to predict whether a breast mass is benign or malignant based on different cell measurements taken as input for diagnosis. The Breast Cancer Diagnosis Predictor Application, therefore, tries to put this technology into application to avail the tool with which medical professionals can effectively conduct early detection of cancerous cases, thereby upgrading the accuracy of diagnosis and decisiveness with speed.

1.2 Statement of the Problem

While medical technology is improving, the diagnosis of breast cancer remains a very challenging process often requiring time-consuming manual analysis of diagnostic data. Medical professionals are specifically tasked with reviewing complex cytology results to determine whether a breast mass is benign or malignant. The methods currently used are prone to delay, human error, and lack of uniformity in diagnosis, which can adversely affect patient care and treatment outcomes. What is needed is an intelligent system that can predict diagnoses of breast cancer with rapidity and accuracy to provide decision support to medical professionals.

1.3 Objectives

1.3.1 Main Objective

To develop a full-stack AI application powered by machine learning to aid medical professionals in diagnosing breast cancer.

1.3.2 Specific Objectives

i. Analyse various breast cell nuclei measurements from a dataset and use them to train the ML model using logistic regression.

ii. Develop and integrate a Streamlit application (UI/Frontend) with the backend(ML model)

iii. Implement and test the breast cancer diagnosis predictor to automatically identify whether a breast tissue is either benign or malignant of cancer.

iv. Design a full-stack AI application for the breast cancer diagnosis predictor

1.4. Significance of the Study

This study is significant because it addresses one of the most critical challenges to healthcare: early-and-accurate diagnosis of cases of breast cancer. The project introduces the use of a machine-learning-powered tool in assisting diagnosis for improvement in patient outcomes through early detection and more informed decision-making.

Alignment with United Nations Sustainable Development Goal (UNSDGs)

SDG 3: Good Health and Well-Being

The application is important to ascertain healthy lives, ensure and promote well-being to all hence supporting SDG 3 through better diagnostic quality care, early treatment of diseases,

detection and efficiency in the delivery of healthcare. This could also reduce mortality rates from cancer, hence leading to the best health outcomes.

SDG 5: Gender Equality

Breast cancer predominantly affects women and providing an advanced diagnostic tool that this application addresses, it promotes and advances gender equality in health. By improving early diagnosis and enhancing healthcare delivery for conditions that disproportionately affect women, this application contributes to SDG 5, by ensuring equal access to quality health care for all genders. Such a tool can help bridge gaps in healthcare for women and girls, especially in resource-poor settings where specialized diagnostics are limited.

SDG 9: Industry, Innovation and Infrastructure

The development of a diagnostic tool using machine learning techniques falls under SDG 9, since this endeavour fosters innovation within the healthcare sector. A representative example of this, show casing advanced technology for changing conventional diagnostic practices, the to be developed application is considered one more step toward modernizing infrastructure within healthcare. Using AI for better diagnostics in efficiency and quality will further ensure building resilient healthcare systems-innovative and responsive towards

1.5 Scope of the Study

The scope of the work will involve the design, development and deployment of a machine learning-based Breast Cancer Diagnosis Predictor Application that will take specific diagnostic cell measurements as input and make predictions (probability) about the malignancy of the breast mass. The work will be focused on the developing an easy-to-use interface by medical professionals for manual input, prediction and visualization. The scope of the project will not include the integration with external system such as cytology lab machines to automatically input data. In the future though, this may be considered in iterations of the application, but data input will be manual for this project.

1.6 Assumptions

The Breast Cancer Diagnosis Predictor Application will work under some crucial assumptions, forming the bedrock of the application's design, functionality and the expected outcomes. These assumptions will make the application achieve its objectives and create realistic expectations toward the development and use of the application in clinical environments. They include the following:

Compliance with Data Privacy and Security Standards

The application will assume compliance with medical data privacy regulations, such as HIPAA or GDPR to ensure that all patient data entered into the system is handled securely. This is crucial for the application's acceptance and ethical use within clinical settings.

Model Stability and Predictive Consistency

It is assumed that the machine learning model to be developed will continue to perform consistently over time and yield reliable predictions within the range of expected breast mass characteristics. If any of these distributions of input data change significantly (like new types of measurements or infrequent diagnostic patterns), the model may need retraining for predictive accuracy.

Continuous Model Improvement as New Data Becomes Available

It is hoped that over time, the model could be refined and re-trained with new data to enhance the diagnostic accuracy and relevance of diagnoses to modern medical findings. This continuous improvement process will allow the application to stay up-to-date with diagnostic advancements in breast cancer detection.

1.7 Limitations and Delimitations (Challenges and Counter Measures)

One of the main limitations of the study is the ***reliance on manual data input***, which can introduce human error and affect the accuracy of predictions. While the machine learning model is expected to be accurate, any inaccuracies in the input data could lead to less reliable

diagnoses. Future versions of the application could ***mitigate*** this limitation by ***integrating the system directly with laboratory equipment or machine for automatic data collection.***

Another limitation is the performance of the machine learning model across diverse patient populations. The ***initial training dataset*** may not capture all variations in breast cancer diagnostics, leading to ***potential biases*** in the prediction model. To ***counter*** this challenge, ***continuous model improvement*** and retraining with diverse datasets will be necessary.

Lastly, the application is only meant to be a ***decision support tool*** and not a replacement for medical expertise. Medical professionals must interpret the results in the context of a comprehensive diagnostic process

CHAPTER 2- LITERATURE REVIEW

2.1 Introduction

At this point in time, breast cancer remains one of the most common types of cancer in the world; thus, it requires a high level of diagnosis using advanced tools to assist in early detection and treatment planning. Machine learning and AI have shown tremendous potential in medical diagnostics, especially in the detection of cancer and has picked up considerable speed over the last decade. Machine learning algorithms form a basis for developing predictive models to assist healthcare professionals in diagnostic decision-making. This development has facilitated the creation of tools that can process huge datasets to identify patterns and trends that may not be immediately apparent to human analysts. This chapter highlights the current existing systems available for breast cancer prediction, their limitations and what improvements the proposed solution might be achieve.

2.2 Related Systems

Several systems have been developed to facilitate the early detection and diagnosis of breast cancer. Below are three prominent systems discussed in detail:

Breast Cancer Risk Assessment System Models

Breast cancer system risk assessment models such as the Gail Model and the Tyrer-Cuzick Model are statistical tools designed to estimate an individual's risk of developing breast cancer. They typically utilize data on genetic predispositions, lifestyle factors, and personal and family medical histories (Costantino et al., 1999).

The Gail Model for instance incorporates age, family history of breast cancer, reproductive history and previous biopsy results in order to produce a 5-year and lifetime percentage risk of cancer (Bondy et al., 2005). Likewise, the Tyrer-Cuzick Model further expands the aforementioned elements through the incorporation of information regarding hormonal variables, as well as genetic alterations such as BRCA1 and BRCA2 (Tyrer et al., 2004).

IBM Watson for Oncology

IBM Watson for Oncology uses AI algorithms to analyse medical literature and patient data, providing recommendations for cancer diagnosis and treatment. It considers patient-specific factors and aligns its suggestions with available clinical guidelines and best practices (Jiang et al., 2017).

Computer-Aided Diagnosis (CAD) Systems for Mammography

Computer-Aided Diagnosis (CAD) systems are designed to assist radiologists in the interpretation of mammographic images by highlighting areas that may indicate the presence of breast cancer. These systems use image recognition techniques to identify suspicious regions and highlight them for further analysis by medical professionals (Doi, 2007).

2.3 Limitations

Despite their advancements, the above systems have significant weaknesses that hinder their effectiveness in certain contexts:

Breast Cancer Risk Assessment System Models:

Generalization Issues

Most of the breast cancer risk assessment models like the Gail Model suffer from generalization problems on diverse populations since they are validated mostly on Caucasian datasets (Bondy et al., 2005).

Narrow Predictive Power

The risk models provide probabilities and do not diagnose a patient; therefore, diagnostic tests must be conducted to reach a conclusion (Costantino et al., 1999).

Genetic Risks Underestimation

Although the Tyrer-Cuzick Model incorporates genetic risk factors, it does not fully capture the complexities of polygenic risks or other unknown genetic factors (Tyrer et al., 2004).

Computer-Aided Diagnosis (CADx) Systems for Mammography:

High False Positive Rates

CADx systems are known to generate a high number of false positives, leading to unnecessary biopsies and patient anxiety (Kooi et al., 2017).

Limited to Imaging Data

CADx systems are primarily focused on mammographic imaging, not taking into consideration other diagnosis parameters such as cell feature characteristics from biopsy results or even the history of the patient; this is important for full diagnosis (Doi 2007; Lehman et al. 2015).

Effectiveness depends on radiologists.

While CADx systems offer value in terms of insights, their effectiveness can vary from case to case and is most pronounced when used as an assistive tool with the expertise of radiologists. (Taylor et al. 2019).

IBM Watson for Oncology:

Complexity and Cost

The high computational power and related infrastructure costs required for IBM Watson for Oncology make it an unsuitable approach for small healthcare settings found in low- and middle-income countries (Jiang et al., 2017).

Limited Adaptation to Local Guidelines

While the recommendations of IBM Watson will be based on global best practices, it does not adapt the recommendations to the local healthcare system and therefore, the constraints and often reduces its applicability (Lee et al., 2018).

Data Privacy Concerns

The fact that there is a lot of sensitive medical data being manipulated increases ethical and regulatory complexities, especially in jurisdictions that have tight data protection legislation (Amann et al., 2020).

2.4 How the proposed solution will handle these weaknesses

The Breast Cancer Diagnosis Predictor Application will address the identified limitations of the existing systems as follows:

Improved Generalizability

The application will employ the publicly available Breast Cancer Wisconsin (Diagnostic) Dataset from Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data) that includes varied features derived from FNA procedures offering better representation of characteristics related to breast cancer. In addition, machine learning allows for the model to be retrained or fine-tuned with the regional population dataset to solve generalization problems across different demographics.

Customizable for Local Contexts

While IBM Watson is bound to global standards, the modular architecture of the predictor application will allow customization based on region-specific datasets or a new ML model and design architecture in the future, hence being more relevant to the local healthcare systems making it a better fit for diverse healthcare environments.

Direct Diagnostic Classification

Unlike the traditional risk assessment models, the application will make use of a trained logistic regression model to directly classify cell clusters as either benign or malignant. This diagnostic capability eliminates the reliance solely on probabilistic risk scores and offers actionable insights, reducing the need for extensive confirmatory testing.

Feature-Based Complementary Insights

While the system does not explicitly incorporate genetic risks, its focus on FNA-derived features such as texture, concavity, symmetry and etcetera provides very useful predictive insights that complement genetic risk factors. The modular design allows for the integration of genetic data in the future if datasets with such features become available.

Optimized Predictions

The logistic regression model, trained and tested on well-curated data will reduce false positives by optimizing hyperparameters and performance metrics during the training process. Besides this, the predictions will be given with their probability scores, which allow the stakeholders to get an idea about the confidence level of the diagnosis and avoid unnecessary alarms.

Multimodal Diagnosis

The predictor application will use non-imaging FNA data as from the Kaggle dataset, focusing on critical cellular characteristics such as concavity and compactness. This will make it a complementary tool to imaging-focused CADx systems, broadening the diagnostic scope by incorporating biopsy-derived parameters and user-input clinical data.

Standalone Diagnostic Accessibility

Unlike CADx systems which rely heavily on radiologists for interpretation, the predictor application is designed as a stand-alone diagnostic aid accessible to any medical cancer professional without extensive expertise in mammographic analysis. A streamlined interface and interpretative output will enable users with minimal technical training to make informed decisions.

Cost-Effective Framework Deployment

The application will leverage lightweight frameworks, Streamlit and Python, which are computationally efficient and low-cost. The tech stack will be designed to be deployable on consumer-grade hardware to make it accessible for small clinics and hospitals in low- and middle-income regions.

Secure On-Premise Data Handling

Data privacy will be ensured by offering secure local data storage and optional anonymization of patient inputs. The app will be designed for on-premise deployment, sensitive patient data would remain with the healthcare institution reducing the probability of data breaches and full compliance with privacy regulations.

User-Friendly Interface

The application shall include a well-designed user interface to address usability enabling medical practitioners to input diagnostic data into the system manually, predict, and interpret the results through a visual aid called a radar chart. The inclusion of a radar charts will clearly visualize diagnostic data and help medical professionals to interpret the features influencing the prediction-a requirement that is often lacking in most existing current systems

CHAPTER 3- METHODOLOGY

3.1 Introduction

This chapter outlines the development methodology that will be followed for the Breast Cancer Diagnosis Predictor Application. It gives the actual roadmap on how project objectives will be achieved; right from the requirements gathering to the final deployment of the project. The project shall be developed using the Software Development Life Cycle (SDLC) methodology to make sure that each phase of the development process follows a structured and systematic approach.

Python will be used as the backend principal development language for machine learning and data processing. Development of the UI, that is the frontend including an interactive web-based dashboard will be done using Streamlit and CSS, which will enable the medical professional to input diagnostic data and get predictions. Additional data visualization libraries such as Plotly will be used in making an informative and interactive radar chart. The machine learning model to be used, logistic regression, will be trained using the Breast Cancer Wisconsin Diagnostic Dataset from Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data) and will serve as the basis for the predictive model to classify whether a cell cluster is either benign or malignant.

For deployment and hosting, this project will make use of Streamlit Cloud which provides easy deployment and maintenance of the application to keep it accessible via any web browser. The same platform also supports scalability so that the application is accessible across a range of devices without requiring any local installations.

Along with the development tool and framework, this chapter also designates the procedure for designing the system architecture, defining the system requirements, both functional and non-functional, and establishing data collection and analysis methods. The project schedule and budget are also provided to make sure of effective planning and resource allocation throughout the project lifecycle.

3.2 Project Design (Waterfall Method)

The project will follow the **Waterfall Model** of the Software Development Life Cycle (SDLC) methodology. The Waterfall model is a linear and sequential approach to software development, where each phase of the project flows downward, much like a waterfall, from one stage to the next. The project will only move to the next phase once the current one is completed ensuring that no important steps are overlooked.

Justification for the Waterfall Model:

Simplicity: The rigid structure of the waterfall model is easy to understand and implement; hence, it is easy to manage and thus suitable for this project.

Well-defined Milestones: Each project phase has clear objectives and deliverables, ensuring clarity in the progress tracking.

Focus on Documentation: The model emphasizes documentation at each stage, something that is essential for this project.

Minimized Risk: Testing falls in a different phase after development, thus minimizing unexpected problems at deployment.

The **Waterfall Model** will guarantee that the project is done systematically in a timely manner to ensure the creation of a high-quality, user-friendly, functional Breast Cancer Diagnosis Predictor Application.

3.3 Design Procedures

The design procedures of the *Breast Cancer Diagnosis Predictor Application* is structured to align with the Waterfall SDLC methodology model. Each step of the process is elaborated on in this section to ensure clarity and alignment with the project objectives and deliverables.The steps include the following:

1. Requirements Gathering and Analysis:

This phase focuses on understanding the needs of the stakeholders including the technical requirements needed for the application. The stakeholders include *patients*, *certified medical professionals* (radiologists, oncologists or lab technicians), *hospitals and clinics* and *myself (as the developer).* The primary objective is to develop a user-friendly application capable of:

Diagnosis Prediction: It should classify cell clusters as either benign or malignant using a logistic regression model.

Visualization: It should visualize predictions and insights through interactive visualization using a radar chart.

User Interaction: A smooth and usable interface for medical professionals to input diagnostic data.

Technical Requirements

The application should be *compatible* with the latest web browsers, make fast predictions, and be *responsive on all devices*. *Performance indicators* such as *low latency* in predictions and *seamless integration* of machine learning and visualization components will be *prioritized.*

2. System Architecture:

The application architecture will be designed in a *three-layered* structure:

Frontend: the frontend will be developed using Streamlit and a bit of CSS to provide an interactive dashboard where users input diagnostic data and view predictions along with their probability scores.

Backend: A Python-based machine learning model (logistic regression) will be implemented using machine learning libraries like Scikit-learn for predictive analysis etcetera.

Database: Since the application uses a pre-existing dataset (Breast Cancer Wisconsin Diagnostic Data from Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data)), there wont be a traditional database. The model will directly load and process this dataset. The data flow will involve receiving inputs from the user interface, passing them through the trained model and displaying predictions and related visualizations.

Visualization: The application will utilize Plotly to generate an interactive radar chart to enhance user comprehension of predictions.

3. Detailed Design:

The system shall be designed in a manner to maximize user experience and computational efficiency;

User Flow: The user's journey from inputting diagnostic data to presenting results shall be mapped. Each step in the interaction shall be intuitive thus ensuring that users can understand and trust the predictions provided.

Integration: The backend that is the machine learning model, will interact with the frontend to produce results in real time. The architecture will be modularly integrated to support updates and enhancements.

4. User Interface Design:

The UI will be designed with simplicity and functionality in mind as follows:

Input Fields: The user will input diagnostic metrics, such as radius, texture, perimeter, etcetera., via sliding bars which will be created for every cell feature characteristic.

Interactive Dashboard: The dashboard will present the predictions, the probability score and visualizations in a clear, orderly fashion.

Charts: Interactive visualizations via a radar chart using Plotly will be used to better understand the model predictions thus help in decision making.

The UI will also be tested for usability and responsiveness across different devices and browsers.

5. Data Design:

The project shall use the Breast Cancer Wisconsin Diagnostic Data Set, from Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data);

Data Format: The dataset will be pre-processed for handling missing values, data normalization and preparation of features for the logistic regression model (EDA).

Data Flow: User inputs will be passed from the UI to the backend for the trained logistic regression model to process the input data and predict whether a cell cluster is benign or malignant. Results including the probability score and visualizations will be displayed on the dashboard.

6. Implementation:

Backend Development

Train the logistic regression model using Scikit-learn and preprocess the dataset using Pandas and Numpy and develop a Python script to load the model and return predictions based on user inputs.

Use Pickle machine learning library to export the model to the frontend to prevent re-training of the model thus reducing latency and improving performance of the application.

Frontend Development

Bid on an interactive Streamlit application that captures user inputs and displays model predictions. Employ Plotly for creating the radar chart for visualization of the results.

Integration of all the components above will give us the full-stack AI application.

7. Testing and Validation:

Full Testing (System Testing) will make sure the application is seamlessly functional:

Functionality Testing
The logistic regression model's accuracy and classification report including the precision call etcetera will be validated by splitting the dataset into training and testing subsets, to test the correctness of predictions from the Logistic Regression model, ensuring it performs well on unseen data

Unit Testing
The system's individual components such as the machine learning model, user input forms, probability score and visualizations will be tested.

Integration Testing
Frontend and backend work in cohesion, and Streamlit allows seamless integration of CSS, Plotly, and the machine learning model.

System Testing
Test the system application as a whole.

User Acceptance Testing (UAT)
After the system is deployed, it will undergo UAT to ensure that it meets user expectations and works as intended.

8. Deployment

The application will be deployed on Streamlit Cloud, ensuring; ***public accessibility*** with no need for local installations, ***scalability*** to support multiple users and ***ease in maintenance*** in case of future updates and enhancements. Streamlit Cloud offers a free deployment solution that is efficient and hosts support to share Streamlit applications.

9. Maintenance

After deployment, the system will enter the phase of maintenance where:

Bug Fixes: Any bugs detected post-deployment are fixed.

Updates: Integrate users' feedback to enhance the working and usability of the application.

Enhancements: Visualizations will be enhanced; the predictive capabilities will be extended or new features added to the application based on changes in medical knowledge or users' requirements.

This structured design process will ensure that the Breast Cancer Diagnosis Predictor Application is built efficiently, is user-friendly and achieve its objectives of assisting medical professionals in diagnosing breast cancer.

3.4 System Requirements

Functional Requirements:

Data Input

The user (medical professional) should be able to enter the values for different features of the cell cluster, like radius, texture, perimeter and etcetera.

Prediction Output

The system should predict along with the probability score whether a cell cluster is benign or malignant scores based on the input data fed into it.

Visualization

The system should visualize the model's predictions through an interactive radar chart.

Model Accuracy

The machine learning model should have high accuracy in predicting the classification of breast cancer.

Non-Functional Requirements:

Usability

The web application should be user-friendly, ensuring that users with minimal technical knowledge can interact with it effectively.

Performance

The system should be responsive, having little to no delay in the display of the UI components including the dashboard, predictions and the radar visualization.

Security

The application must securely handle any sensitive user data; in this case, no personally identifiable information is used.

Scalability

The application should be able to handle large volume of user data and multiple concurrent live users.

Compatibility

The system should be able to function on multiple OS platforms and different device architectures for user accessibility.
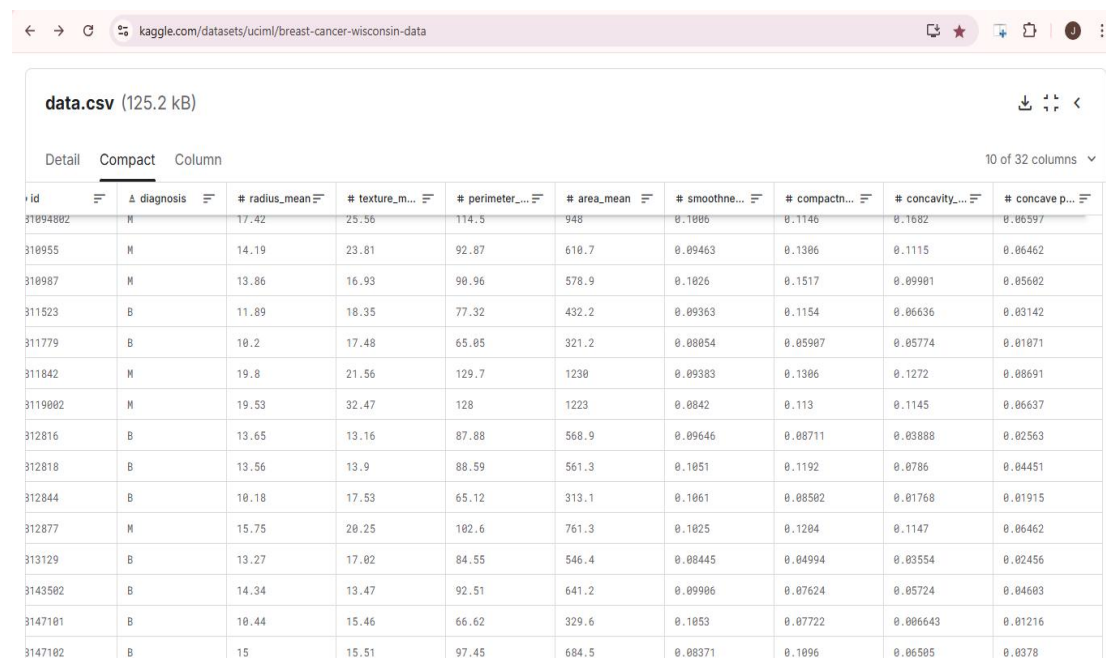
Reliability

The system should be reliable and readily available to users with minimal downtime and disruptions

3.5 Data Collection and Analysis (needs assessment) methods and tools

The dataset used for the project is the Breast Cancer Wisconsin (Diagnostic) Dataset which is freely available on Kaggle (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data). The dataset originates from the University of Wisconsin Hospitals, Madison compiled by Dr. William H. Wolberg (Wolberg, 1995). It has been quite popular in machine learning research for breast cancer prediction because of its well-organized structure and good quality features.

Figure 1

Dataset Characteristics

The dataset contains 569 rows and 32 columns. Each row represents a single instance of a diagnosis for a cluster of breast cells while the columns represent the features of that diagnosis.The columns include the following:

ID Column:

- ▪ *`id`:* A unique identifier for each sample.

Target Variable:

The dataset contains a ***diagnosis*** column which is the target variable that classifies each sample as either:

- **M** (Malignant): Representing malignant or cancerous cluster.
- **B** (Benign): Representing non-cancerous or benign cluster.

Feature Variables:

There are 30 numeric features; every single feature in the list originally generated from the fine-needle aspirate (FNA) of a tumour in a breast mass. They fall into three fundamental classes:

*Mean Values***:** Measurements averaged across the sample.

*Standard Errors***:** Variability of the sample measurements.

Worst Values: Most extreme measurements within the sample.

The columns describe specific characteristics for each of the 3 classes above of cell nuclei, derived from the FNA images:

Radius: The average distance from the center to points on the perimeter.

Texture: The standard deviation of gray-scale values.

Perimeter: The circumference of the cell nucleus.

Area: The total size of the cell nucleus.

Smoothness: The variation in the radius lengths.

Compactness: The ratio of the perimeter squared to the area (indicating circularity).

Concavity: The severity of concave portions of the cell nucleus.

Concave Points: The number of concave portions.

Symmetry: The degree of symmetry in the nucleus shape.

Fractal Dimension: A measure of complexity, describing changes in detail at different scales.

Each of the above 10 features is described as the mean, standard error and worst value thus forming 30 feature columns.

Dataset Properties

The dataset is well-balanced, with a slight predominance of benign cases, with 357 being benign and 212 malignant.

All the feature values are numeric making it an apt dataset for machine learning use.

There are no missing or mismatching values in the dataset reducing the time and complexity of preprocessing.

Importance of the Dataset (Why this Dataset)

The Breast Cancer Wisconsin (Diagnostic) Dataset is highly suitable for the project due to its:

Real-world Relevance: Since it is from medical research, it reflects real clinical data.

Completeness: The high-dimensional feature set enables the fine-grained characterization of cell clusters and hence it will improve the predictive capability of the machine learning model.

Quality: There are no missing or mismatching data points hence allowing a sound and reliable model training.

Data Analysis:

Preprocessing: The dataset will undergo necessary preprocessing steps such as normalization feature selection etcetera.

EDA: Using visualization and statistical analysis with the ML libraries; Pandas, Numpy etcetera in Python to understand the distribution of the data and identify patterns.

Model Training: Data will be split into both training and testing sets. In this case, the Logistic Regression model will be trained using Scikit-learn.

Model Evaluation: The model's performance will be evaluated using metrics such as accuracy, precision, recall, and F1-score.

Tools To Be Used:

Python: The primary language for the backend development and programming including model training and analysis.

Streamlit: Streamlit shall be employed for building up a web interactive dashboard UI for the app.

Scikit-learn: Will be use to train the logistic regression of the machine learning model.

Plotly: For an interactive visualizations of data input and predictions.

Pandas: For data Manipulation and pre-processing.

Numpy: For numerical computing and array operations to enhance data preprocessing and statistical computations.

Pickle: Will be employed to serialize and save the trained machine learning model for efficient performance, deployment and future use.

Altair: For creating declarative and simple statistical visualization to complement Plotly's interactive visualizations.

The above tools will collectively ensure a robust, efficient and visually appealing Breast Cancer Diagnosis Predictor application that integrates machine learning, data visualization and interactive user experiences.

REFERENCES

1. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Medical Informatics and Decision Making, 20(1), 310. https://doi.org/10.1186/s12911-020-01332-6

2. Bondy, M. L., et al. (2005). Race, ethnicity, and breast cancer: The debatable role of genetics. Oncologist, 10(1), 59–69.

3. Costantino, J. P., et al. (1999). Validation studies for models projecting the risk of invasive and total breast cancer incidence. Journal of the National Cancer Institute, 91(18), 1541–1548. https://doi.org/10.1093/jnci/91.18.1541

4. Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status, and future potential. Computerized Medical Imaging and Graphics, 31(4–5), 198–211. https://doi.org/10.1016/j.compmedimag.2007.02.002

5. Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: Past, present, and future. Stroke and Vascular Neurology, 2(4), 230–243. https://doi.org/10.1136/svn-2017-000101

6. Kooi, T., Litjens, G., van Ginneken, B., Gubern-Mérida, A., Sánchez, C. I., Mann, R., & den Heeten, A. (2017). Large scale deep learning for computer-aided detection of mammographic lesions. Medical Image Analysis, 35, 303–312.

7. Lee, C. H., Yoon, H. J., & Kim, D. H. (2018). Application of artificial intelligence in cancer diagnosis and predictive models. Cancer Informatics, 17, 1–7.

8. Lehman, C. D., Wellman, R. D., Buist, D. S. M., Kerlikowske, K., Tosteson, A. N. A., & Miglioretti, D. L. (2015). Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Internal Medicine, 175(11), 1828–1837. https://doi.org/10.1001/jamainternmed.2015.5231

9. Taylor, P., Potts, H. W. W., & Benson, R. (2019). Computer-aided detection in mammography: An overview and future outlook. Radiology Research and Practice, 2019, 5876543.

10. Tyrer, J., et al. (2004). A breast cancer prediction model incorporating familial and personal risk factors. Statistics in Medicine, 23(7), 1111–1130. https://doi.org/10.1002/sim.1668

11. Wolberg, W. H. (1995). Breast Cancer Wisconsin (Diagnostic) Dataset. University of Wisconsin Hospitals, Madison. Retrieved from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

APPENDICES

APPENDIX A: PROPOSED SCHEDULE

Table 1

Activities	Months			
	1	**2**	**3**	**4**
Project Conceptualization and Requirements Gathering	██			
Project Research and Literature Review		██		
System Methodology			██	
Project Proposal Writing and Submission		██	██	██
Corrections and Final Submission				██

APPENDIX B: PROPOSED BUDGET

Table 2

S. No	Items/Activities	Quantity	@Ksh	Amount in Ksh
1.	Paper Printing	6 dozen	60	360
2.	Spiral Binder	2	70	140
3.	Paper Punch	1	700	700
4.	Miscellaneous			500
		Total		**sh.1750**