Enhanced Medical Intelligent System for Cancer Disease Prediction, Using Case-Based Reasoning

Ike Mgbeafulike
Computer Science Department, Chukwuemeka Odumegwu Ojukwu University,
Anambra State
ike.mgbeafulike@gmail.com

Anyadiegwu, Happiness Onyinye
Computer Science Department, Nnamdi Azikiwe University Awka
Anambra State
ho.anyadiegwu@unizik.edu.ng

DOI: 10.56201/ijcsmt.vol.11.no4.2025.pg.23.45

Abstract

Breast cancer remains a leading cause of morbidity and mortality among women globally, necessitating timely and accurate diagnostic support systems. This research proposes a hybrid intelligent system that integrates Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Case-Based Reasoning (CBR) for early breast cancer prediction and diagnosis. The model was trained on a clinical breast cancer dataset, with feature attributes such as tumor size, symptom severity, hormone receptor status, HER2 status, and other relevant features. To ensure high-quality input, rigorous data preprocessing and feature engineering techniques were applied, including encoding of categorical variables, feature scaling, and exploratory analysis. The hybrid SVM-KNN model demonstrated high predictive performance, achieving an accuracy of 95.3%, along with strong precision, recall, and F1-score metrics. The system was deployed using Streamlit and integrated with an SQLite database, enabling real-time predictions, case-based retrieval, and diagnostic report generation.

Keywords: *Cancer prediction; Case-Based Reasoning (CBR); Machine Learning; Support Vector Machine (SVM); K-Nearest Neighbors (KNN)*

1.1 Introduction

Breast cancer remains a significant global health concern, ranking as the most commonly diagnosed cancer among women and accounting for approximately 30% of all female cancer cases worldwide (Wilkinson & Gathani, 2022). The breast, composed of glandular tissue, ducts, connective tissue, and adipose tissue, primarily functions in milk production and secretion for infant nourishment (Colleluori et al., 2021). While hormonal changes during puberty, pregnancy, and lactation influence breast development, pathological conditions such as breast cancer pose serious health challenges.

Cancer is a complex group of diseases characterized by the uncontrolled proliferation of abnormal cells, leading to tumor formation and potential metastasis (Singh & Roghini, 2023). Breast cancer originates in the cells of the breast, typically within the milk-producing ducts or lobules. Although it primarily affects women, men can also develop breast cancer, albeit at a lower incidence. Breast cancer is classified based on the affected cell type and its invasiveness, distinguishing between invasive and non-invasive forms. Early detection through screening techniques such as mammography has significantly improved survival rates by enabling timely interventions through surgery, radiation therapy, chemotherapy, and targeted therapies (Łukasiewicz et al., 2021).

Studies indicate that in Nigeria, as in many other parts of the world, breast cancer is a leading cause of morbidity and mortality among women. However, systemic challenges—including limited healthcare infrastructure, low awareness, and restricted access to advanced diagnostic tools—exacerbate delays in diagnosis and treatment (Pesapane et al., 2023). Consequently, late-stage diagnoses remain prevalent, reducing the likelihood of successful treatment and long-term survival.

Findings from various sources demonstrate that with rapid advancements in Artificial Intelligence (AI) and Machine Learning (ML), there is increasing interest in utilizing these technologies to enhance medical decision-making and predictive capabilities. Case-Based Reasoning (CBR), a form of AI that applies knowledge from previously encountered cases to solve new problems with similar characteristics, has demonstrated significant potential in healthcare applications, particularly in diagnostic decision-making, prognosis prediction, and treatment planning (Patel et al., 2023).

The development of an intelligent medical system utilizing CBR for breast cancer prediction presents a promising avenue for improving clinical decision-making. By analyzing historical patient data and identifying patterns from past cases. It is widely acknowledged that CBR systems can provide personalized predictions, assisting healthcare professionals in diagnosing, predicting, and managing breast cancer more effectively(Mustafa et al., 2023), (Gu et al., 2021). This research explores the design and implementation of an enhanced CBR-based medical system for breast cancer prediction, focusing on key components such as case representation, similarity measures, and knowledge acquisition processes. The study aims to highlight the potential of AI-driven predictive models in improving early detection, prognosis, and treatment outcomes, ultimately contributing to better healthcare delivery and patient management.

2.1 Related Works

In recent years, Case-Based Reasoning (CBR) has gained significant attention in various healthcare applications, particularly in disease prediction and diagnosis. Several studies have explored the integration of CBR with machine learning techniques to enhance predictive accuracy and improve clinical decision-making.

Gu et al. (2021) proposed a case-based ensemble learning system for explainable breast cancer recurrence prediction. The system combines ensemble learning, which enhances predictive accuracy by aggregating multiple models, with CBR, which allows predictions to be explained through historical cases. This hybrid approach provides a transparent decision-support system, enabling healthcare professionals to compare new patient cases with previously diagnosed cases. The study highlights the importance of explainability in AI-driven medical predictions, as it allows clinicians to assess the reliability of predictions before making critical decisions. The integration of CBR with machine learning models has proven effective in several medical domains, offering both improved accuracy and interpretability. However, challenges remain in optimizing case retrieval, similarity measures, and adaptation mechanisms to further enhance the precision and reliability of predictive models.

In another study, Wilkerson (2023) investigates the application of deep learning-based feature extraction to improve CBR retrieval performance. The study highlights the challenges of generating meaningful indices and presents a methodology that leverages various deep learning models to automatically extract high-quality features from complex datasets. This automated process reduces the reliance on manual feature engineering while enabling more accurate and efficient case retrieval. Wilkerson's work further outlines experimental directions involving the use of different deep learning architectures, training strategies, and feature representation techniques. The findings support the hypothesis that deep learning can significantly enhance the indexing mechanism in CBR systems, thereby improving predictive accuracy in clinical contexts.

Despite these promising outcomes, a notable limitation remains: deep learning models are often computationally intensive, which poses challenges in terms of scalability and real-time performance. This limitation may hinder the deployment of such systems in resource-constrained clinical environments or applications requiring rapid processing.

In a related study, Xu et al. (2022) propose a supervised case-based reasoning (CBR) framework aimed at improving the diagnosis of thyroid nodules (TDNs). The study addresses the limited application of CBR in this domain by introducing a method that reconstructs historical TDN cases using canonical correlation analysis (CCA) to identify the relationships between case features and diagnostic outcomes. These reconstructed cases are then used to train a classifier that predicts pathological outcomes for new cases. To support explainability, a convex optimization model is employed to assess similarity between new and historical cases, and a weighted scheme is used to generate interpretable diagnostic predictions. The approach demonstrates superior performance compared to traditional CBR and six established machine learning models when validated on real-world diagnostic data.

While this work emphasizes explainability and introduces a supervised retrieval mechanism to reduce noise in case matching, it primarily focuses on a single disease domain and relies heavily on structured clinical datasets. A potential gap lies in the generalizability and adaptability of the approach to other complex and multi-modal medical conditions, such as cancer, where feature variability and data heterogeneity are more pronounced.

Similarly**,** Dhatterwal et al. (2021) introduce an intelligent agent-based Case-Based Reasoning (CBR) system for analyzing COVID-19 cases, with the goal of enhancing early detection and treatment. The system employs a Clinical CBR (CCBR) model that utilizes formalized features to evaluate similarities between new patient cases and historical ones stored in a structured case base. The approach integrates a Clinical Knowledge Base System (CKBS) and leverages spatial and temporal patterns from a rich dataset—sourced from the Italian Medical Society and Interventional Radiology (SIRM)—to make early diagnostic predictions. The results demonstrated high accuracy (99.5%) in identifying infectious patients and providing insights for rapid recovery.

While the proposed system showcases the potential of CBR in supporting clinical decision-making during large-scale health crises, it is highly specialized for COVID-19 and may not directly generalize to more complex or chronic diseases such as cancer. Moreover, although the model achieves impressive performance metrics, the abstract lacks details on the interpretability of retrieved cases, a key requirement for clinical adoption in broader domains.

2.2 Summary of Related Works:

The reviewed studies illustrate the growing application of Case-Based Reasoning (CBR) systems in medical diagnostics, emphasizing the integration of machine learning techniques to improve predictive accuracy, explainability, and clinical decision-making. Gu et al. (2021) demonstrated the potential of combining ensemble learning with CBR to predict breast cancer recurrence with enhanced interpretability, though challenges remain in optimizing case retrieval and similarity measures. Similarly, Wilkerson (2023) explored how deep learning-based feature extraction can enhance CBR system indexing, but the computational demands of such models pose scalability issues in resource-limited environments. Xu et al. (2022) developed a supervised CBR framework for thyroid nodule diagnosis, showing improved predictive performance but highlighting the limited generalizability of their approach to other diseases. Dhatterwal et al. (2021) introduced a CBR system for COVID-19 diagnosis, achieving high accuracy but with a focus on a single disease and limited interpretability for broader clinical adoption.

2.2.1 Research Gap:

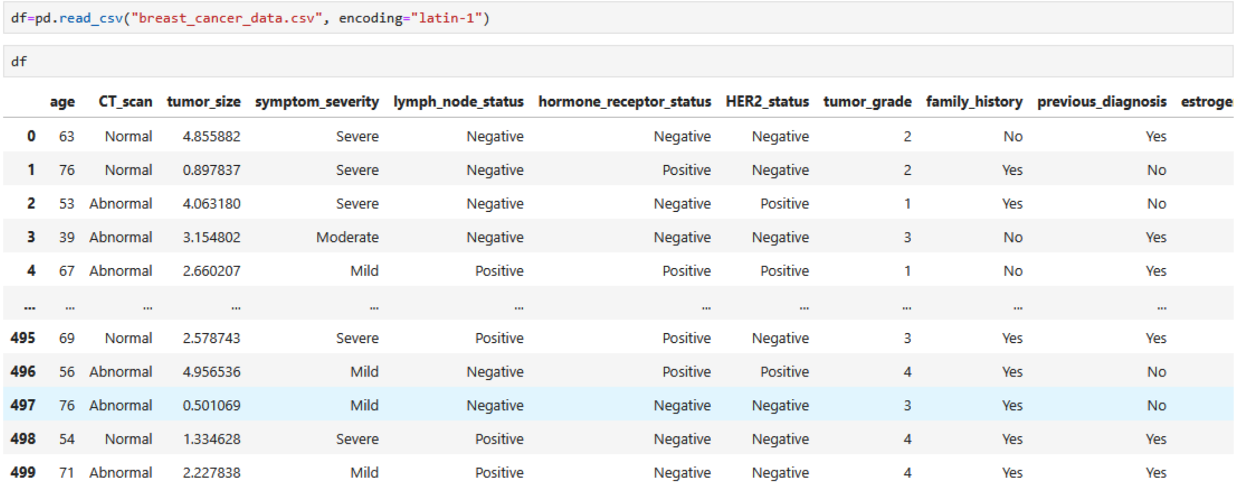
While these studies demonstrate significant advances in applying CBR for medical diagnostics, there remains a gap in the integration of these systems into complex, multi-modal diseases such as breast cancer, where variability in patient data and disease progression is higher. Additionally, the lack of focus on interpretability and real-time processing challenges limits the clinical applicability of these models. This research aims to bridge these gaps by developing an enhanced CBR-based medical system specifically for breast cancer prediction, incorporating advanced machine learning techniques to improve case retrieval, similarity measures, and overall model explainability. The goal is to design a system that not only improves early detection and treatment prediction but also enhances decision-making by providing interpretable and actionable insights for healthcare professionals.

3.1 Methodology

This research proposes an enhanced Case-Based Reasoning (CBR) system for breast cancer prediction, integrating machine learning techniques and clinical decision support to provide accurate and explainable predictions. The system employs a hybrid model combining Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) to improve predictive performance for unseen cases. SVM is used for its ability to handle complex decision boundaries, while KNN complements it with its instance-based learning approach for classifying new cases based on their similarity to known instances. For cases that closely resemble previously encountered scenarios, the system relies on Case-Based Reasoning (CBR) to retrieve predictions directly from the case file, ensuring an efficient and accurate response when similar historical cases are available.

3.1.2 Dataset

The dataset used for training and testing the model consists of a variety of clinical parameters relevant to breast cancer diagnosis. The dataset, cancer_data.csv, includes information such as age, CT scan, tumor size, symptom severity, lymph node status, hormone receptor status, HER2 status, tumor grade, family history, and previous diagnosis, among other clinical attributes.



```
df=pd.read_csv("breast_cancer_data.csv", encoding="latin-1")
```

```
df
```

	age	CT_scan	tumor_size	symptom_severity	lymph_node_status	hormone_receptor_status	HER2_status	tumor_grade	family_history	previous_diagnosis	estroge
0	63	Normal	4.855882	Severe	Negative	Negative	Negative	2	No	Yes	
1	76	Normal	0.897837	Severe	Negative	Positive	Negative	2	Yes	No	
2	53	Abnormal	4.063180	Severe	Negative	Negative	Positive	1	Yes	No	
3	39	Abnormal	3.154802	Moderate	Negative	Negative	Negative	3	No	Yes	
4	67	Abnormal	2.660207	Mild	Positive	Positive	Positive	1	No	Yes	
...	...	...	...	...	...	...	...	...	...	...	
495	69	Normal	2.578743	Severe	Positive	Positive	Negative	3	Yes	Yes	
496	56	Abnormal	4.956536	Mild	Negative	Positive	Positive	4	Yes	No	
497	76	Abnormal	0.501069	Mild	Negative	Negative	Negative	3	Yes	No	
498	54	Normal	1.334628	Severe	Positive	Negative	Negative	4	Yes	Yes	
499	71	Abnormal	2.227838	Mild	Positive	Negative	Negative	4	Yes	Yes	

i. Age: The patient's age.

ii. CT Scan: A binary indicator representing whether a CT scan result is available (1 for Yes, 0 for No).
iii. Tumor Size: The size of the detected tumor in cm.
iv. Symptom Severity: The severity of symptoms, which is recorded on a scale (1-5, where 1 indicates mild symptoms and 5 indicates severe symptoms).
v. Lymph Node Status: Status indicating whether cancer has spread to lymph nodes (1 for Positive, 0 for Negative).
vi. Hormone Receptor Status: Whether the cancer cells have receptors for certain hormones (1 for Positive, 0 for Negative).
vii. HER2 Status: Whether the cancer is associated with the HER2 protein (1 for Positive, 0 for Negative).
viii. Tumor Grade: The grade of the tumor based on how abnormal the cancer cells look under a microscope (Scale 1-3, where 1 is low grade and 3 is high grade).
ix. Family History: Whether the patient has a family history of breast cancer (1 for Yes, 0 for No).
x. Previous Diagnosis: Whether the patient has had a previous breast cancer diagnosis (1 for Yes, 0 for No).
xi. Chemotherapy Response: Patient's response to chemotherapy (Good, Average, Poor).
xii. Radiation Therapy: Whether the patient received radiation therapy (Yes, No).
xiii. Survival Time: Estimated survival time in years (between 1 to 10 years).
xiv. Metastasis: Whether the cancer has spread to other parts of the body (Yes, No).
xv. Recurrence: Whether the cancer has recurred after treatment (Yes, No).
xvi. Cancer: The target column, where 1 indicates cancer (Malignant) and 0 indicates no cancer (Benign).

These features were selected based on their clinical significance in breast cancer diagnosis and prognosis.

3.1.3 Data Preprocessing and Feature Engineering

Data preprocessing is a crucial step to ensure high-quality input for the machine learning model. Missing values were handled using appropriate imputation techniques, and categorical features were encoded into numerical representations using one-hot encoding or label encoding where necessary. Feature scaling was applied to ensure uniformity in the range of values across different features, which is especially important for algorithms like SVM and KNN.

3.1.3.1 Handling Missing Values

Missing data were handled using imputation techniques such as the mean, median, or mode imputation, depending on the feature type (numerical or categorical). For numerical features, missing values were replaced with the mean or median value, whereas for categorical features, the mode was used.

3.1.3.2 Encoding Categorical Features

Categorical variables in the dataset were converted into numerical representations to make them suitable for machine learning models. Label encoding and one-hot encoding were used depending on whether the feature was binary or multi-class.

3.1.3.3 Feature Engineering

Feature engineering involved selecting the most relevant features that contribute to the prediction of breast cancer, such as tumor size, symptom severity, and HER2 status. Principal

Component Analysis (PCA) was also explored to reduce dimensionality and improve model efficiency, though the final model used the full feature set based on validation performance.

3.1.4 Hybrid Model: SVM + KNN

The hybrid model integrates Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) for breast cancer prediction. SVM is used for its capability in high-dimensional spaces, effective in complex datasets, while KNN is employed for its simplicity and ability to provide intuitive, instance-based learning. The hybrid model works by first using SVM to classify the data based on a hyperplane and then refining predictions using KNN, which checks the similarity of a new patient case to the closest historical cases.

The SVM algorithm aims to find a hyperplane that best separates the data points into different classes (cancerous vs. non-cancerous). The key mathematical principle behind SVM is the maximization of the margin between the two classes:

$$Max\frac{2}{||w||}$$

Where:

w is the weight vector (normal vector to the hyperplane).

The margin is the distance between the hyperplane and the nearest data point on either side.

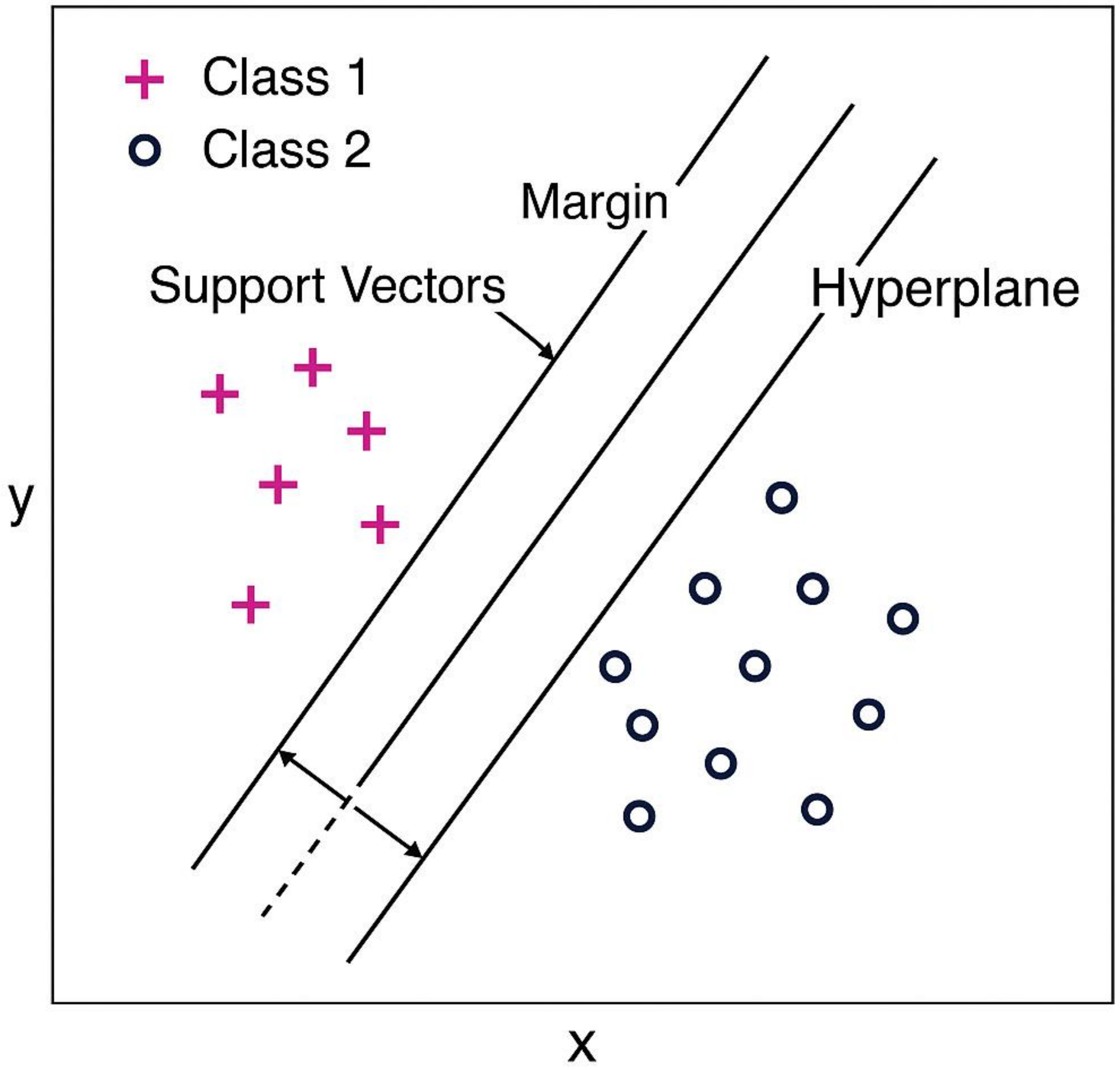


Figure 1.1 Graphical Representation of Support Vector Machine (SVM)**:** Hyperplane and Margin (Pisner & Schnyer, 2020).

Hyperplane: The line or plane that separates the classes.
Margin: The distance between the hyperplane and the closest points (support vectors).

SVM works well for high-dimensional spaces, which makes it effective in classifying cancerous vs. non-cancerous cases, especially when the data has complex relationships between features.
K-Nearest Neighbors (KNN)**:** KNN is a non-parametric method that classifies a data point based on the majority vote of its k nearest neighbors. It works by calculating the Euclidean distance between the new point and all training points:

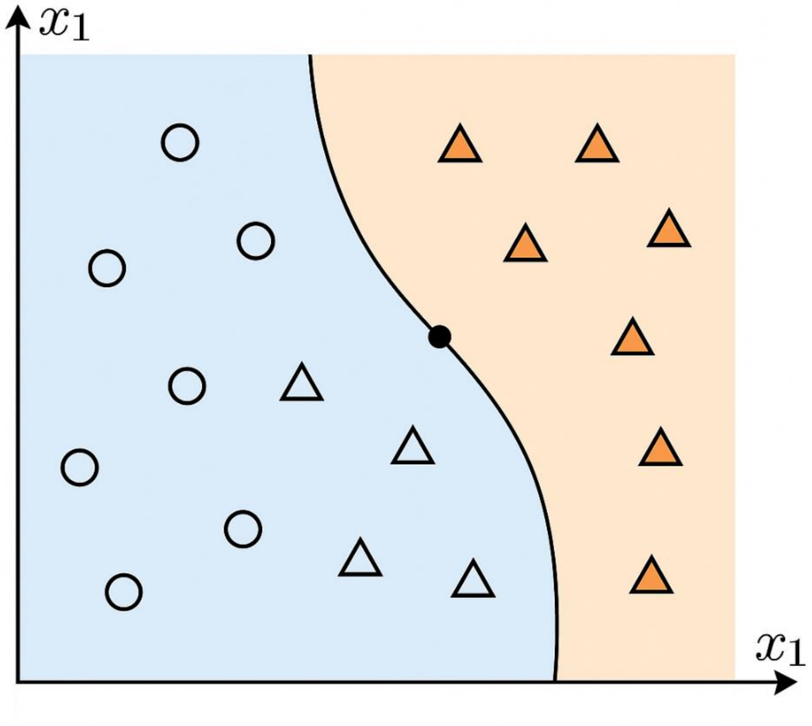


Figure 2: K-Nearest Neighbors (KNN): Decision Boundaries

The KNN visualization shows irregular, flexible boundaries between classes, shaped by the local distribution of training points. A new point is classified based on a majority vote from its 'k' nearest neighbors.

$$Distanc(\mathrm{Xi},\mathrm{Xj}) = \sqrt{(xi1 - xj1)^2 + (xi2 - xj2)^2 + \cdots + (xin - xjn)^2} \quad [1]$$

Where:

- Xi and Xj are two data points, and xin and xjn are their respective feature values.
- kkk is the number of nearest neighbors to consider.

KNN is advantageous due to its simplicity and its instance-based learning, where it uses local proximity to make predictions. KNN is non-parametric and sensitive to local structures, making it effective in capturing complex, nonlinear patterns. Its sensitive to noise and irrelevant features—highlighting the importance of preprocessing and feature selection.

By integrating SVM's robust boundary formation with KNN's local pattern sensitivity, the hybrid model benefits from global structure awareness and local adaptability—enhancing both predictive power and generalization.

3.1.5 Case-Based Reasoning (CBR) Integration

Case-Based Reasoning (CBR) is a problem-solving paradigm that addresses new problems by referencing similar past cases stored in a structured repository. In this research, a hybrid medical diagnostic system is implemented that integrates CBR with machine learning to enhance breast cancer prediction and clinical decision support. The CBR system utilizes an SQLite database to store a case base comprising historical patient records. Each case entry contains a set of clinical attributes such as age, tumor size, hormone receptor status, lymph node involvement, HER2 status, CT scan results, family history, chemotherapy response, radiation therapy, and previous diagnoses. These features are linked to known diagnostic outcomes, enabling retrospective learning and retrieval.

When a new case is entered, the system first searches the database for similar past cases based on the clinical parameters provided (e.g., age, tumor size, hormone receptor status). If a similar case is found, the system retrieves the associated prediction or diagnosis.
If no similar case is found, the system uses the hybrid model (SVM + KNN) to predict the breast cancer diagnosis based on the input clinical parameters. This dynamic approach ensures that the system leverages both historical case data and predictive modeling to provide the best possible diagnosis.

3.1.5.1 Case Representation

Each patient case is represented as a feature vector:

$$C_i = \{x_{1\,,}\, x_2, \ldots, x_{n2}\} \qquad [2]$$

Where:

Ci is the *i-th case*
Xi represents the *i-th clinical feature*, such as age, tumor size, HER2 status, etc.
n is the number of features

These vectors are stored in a **case base** CB=$\{c_{1\,,}\, c_2, \ldots, c_m\}$ where m is the total number of historical cases.

3.1.5.2 Similarity Measurement

When a new case C_{new} is introduced, its similarity to each stored case $C_i \in$ CB

$$\mathrm{Sim}(C_{new,}\, C_i) = \sqrt{a \sum_{j=1}^{n} w_j (x_j^{new} -^{x_j^i})^2} \qquad [3]$$

Where:

x_j^{new} *is the j − th feature of the new case*
x_j^i is the j − th feature of case i
w_j is a weight representing the *importance of feature j*
The system selects cases with the smallest distance values (i.e., highest similarity).

3.1.5.3 Retrieval Logic

Let τ be a similarity threshold

$$C = \{C_i\ \mathcal{E}\ CB \mid Sim(C_{new},\, C_i) \geq r\,\}$$

[4]

If C ≠ Ø, then

$$y_{new} = diagnostic(C_i)\ for\ \ C_i\ \mathcal{E}\ C\ \ with \max similarity$$

Otherwise, proceed to predictive modelling

3.1.5.4 Predictive Hybrid Model (SVM + KNN)

Support Vector Machine (SVM):

Given training data $\{(x_i, y_i)\}_{i=1}^{N}$, SVM seeks a decision function:

$$F(x) = \mathrm{sign}\,(\omega^T\,\phi\,(x) + b) \qquad [5]$$

Where:

$\phi(x)$ maps the input to a higher-dimensional space
ω and b define the optimal separating hyperplane

K-Nearest Neighbors (KNN): Given the same dataset, KNN prediction for a new input x is:

$\acute{y} = \text{mode}(\{y_i \mid x_i \in N_k(x)\})$
where $N_k(x)$ = is the set of k-nearest neighbors of x

3.1.5.6 Combined Decision Rule (Hybrid SVM + KNN)

Let

$$y_{svm}=f_{svm}(x) \quad [6]$$

$$y_{knn}=f_{knn}(x) \quad [7]$$

The hybrid prediction can be formed using a weighted voting scheme**:**

$$\acute{y}_{hybrid} = \text{mode}(\{w_{svm}\cdot y_{svm}, w_{knn}\cdot y_{knn}\}) \quad [8]$$

With weights:

$w_{svm}+w_{knn}=1$

Empirically selected via validation (e.g., $w_{svm}=0.6, w_{knn}=0.4$)

3.1.5.7 Final System Behavior

i. If similar case exists → diagnosis retrieved from CBR
ii. If not → hybrid SVM + KNN model predicts diagnosis

$$\acute{y}_{new} = \left\{\begin{matrix} CBT\ result, & if\ c \neq \emptyset \\ \acute{y}\text{hybrid}, & \text{otherwise} \end{matrix}\right\} \quad [9]$$

3.1.6 Model Evaluation

The performance of the system was evaluated using several metrics, including accuracy, precision, recall, and F1 score. Cross-validation techniques were employed to assess the generalizability of the model, ensuring it performs well across different subsets of the data. Additionally, the system's efficiency in retrieving similar historical cases was assessed, focusing on the response time and reliability of case retrieval.

3.2 System Architecture

The proposed breast cancer prediction system is built as a hybrid intelligent diagnostic framework incorporating a user-friendly interface, a Case-Based Reasoning (CBR) engine, and a hybrid Machine Learning (ML) model. The architecture enables accurate, personalized, and explainable diagnostic predictions.

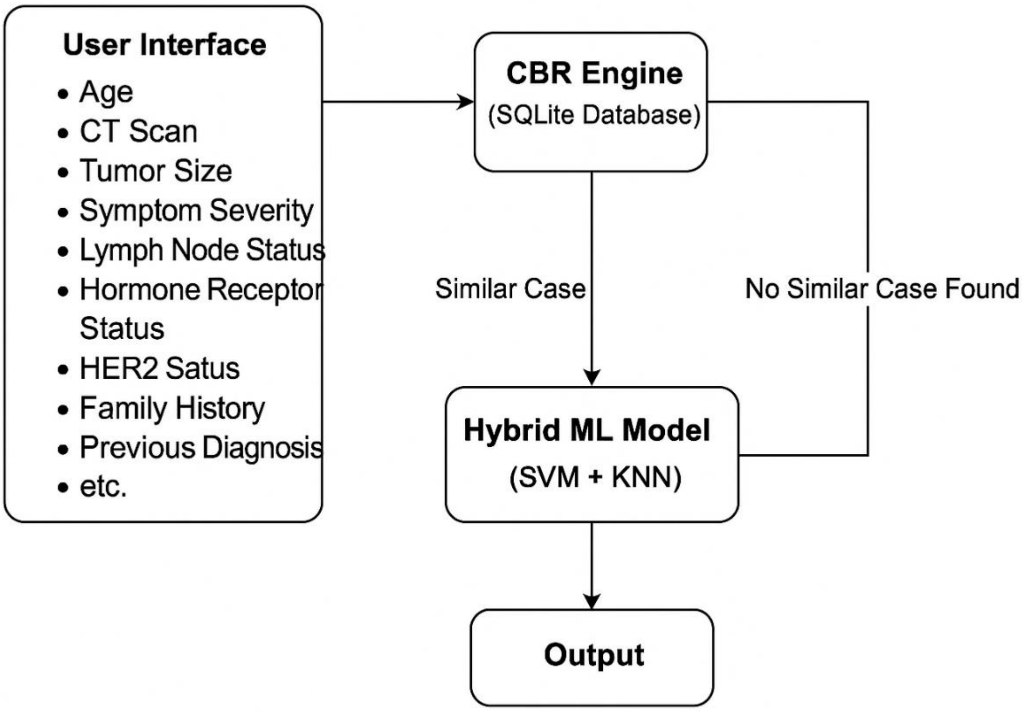


Figure 1 System Architecture

3.2.1 User Interface (Streamlit App)

The front-end of the system is implemented using **Streamlit**, allowing users (e.g., healthcare practitioners or patients) to input relevant clinical parameters through a simple and intuitive GUI.

➡ Accepts user input:

Input Parameters include:

i. Age
ii. CT Scan Result
iii. Tumor Size
iv. Symptom Severity
v. Lymph Node Status
vi. Hormone Receptor Status
vii. HER2 Status
viii. Tumor Grade
ix. Family History
x. Previous Diagnosis

Once entered, the case is sent to the CBR engine for diagnosis matching.

3.2.2 CBR Engine (Case-Based Reasoning using SQLite)

The CBR module leverages an **SQLite** database to store and retrieve historical diagnostic cases.

Operational Flow:

i. Receives the input case from the user interface.
ii. Compares it to previously stored cases using a similarity function (e.g., cosine similarity, Euclidean distance).

iii. If a similar case is found with similarity ≥ threshold, it retrieves and returns the diagnosis directly from the database.

iv. If no similar case is found, the system forwards the input to the hybrid ML model for prediction.

3.2.3 Hybrid ML Model (SVM + KNN)

The hybrid model integrates Support Vector Machine (SVM) **and** K-Nearest Neighbors (KNN) algorithms to improve classification robustness and accuracy.

Training and Prediction:

i. The model is trained on a curated breast cancer dataset after undergoing preprocessing (encoding, scaling, and feature selection).

ii. For prediction, the hybrid model computes a weighted combination of SVM and KNN outputs using the formula.

$$\text{Prediction}_{\text{Hybrid}} = \alpha\cdot\text{SVM}+(1-\alpha)\cdot\text{KNN} \qquad [10]$$

3.2.4 Output Layer

Once a prediction is made:

i. The result is displayed to the user via the UI.

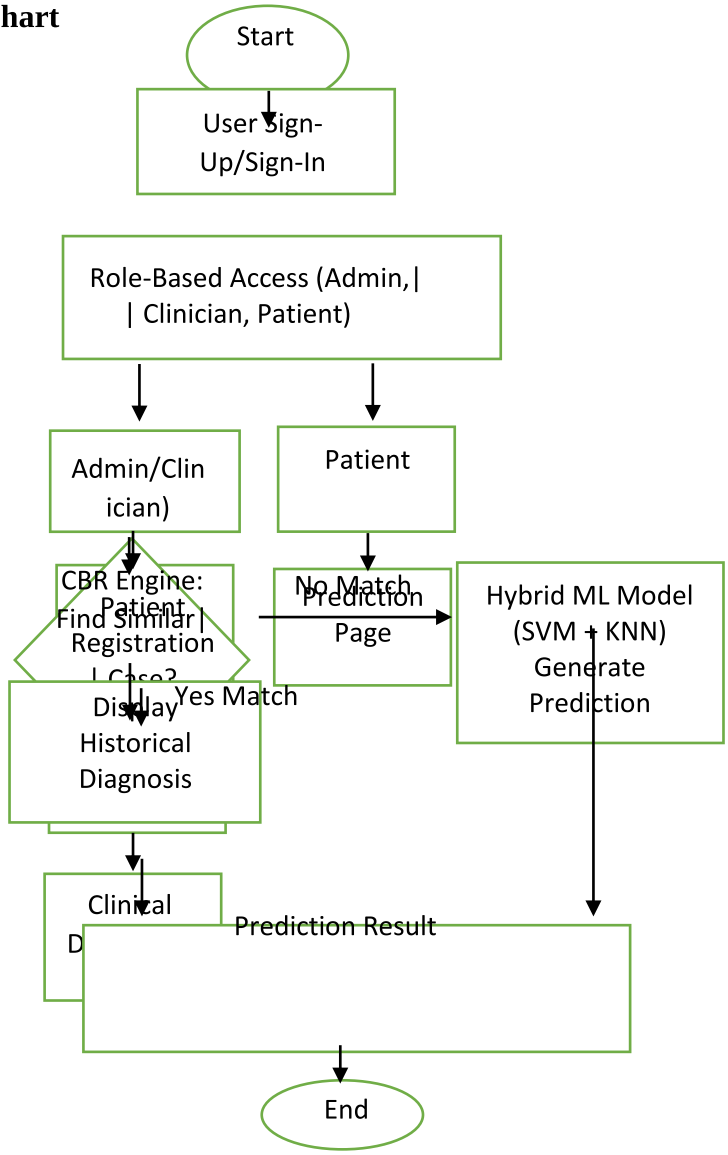
ii. Users have the option to store the new case in the database, thereby enhancing the CBR component for future predictions (CBR learning loop).

iii. Additionally, a PDF report can be generated summarizing the patient input and prediction outcome.

3.3 System Algorithm (Textual Summary)

- 📥 **User Interface (Input Node)**
  → Sends data to →
- 🧠 **CBR Engine (Decision Node)**
  ↪ Checks Case Base (SQLite)
  - ✅ If match → 📤 **Output Layer (Diagnosis Display)**
  - ✕ Else →
- 🔀 **Hybrid ML Model (SVM + KNN)**
  → Sends prediction to →
- 📤 **Output Layer**
  ↪ Option to update Case Base with new data

3.4 System Flowchart


Start
User Sign-Up/Sign-In
Role-Based Access (Admin,| | Clinician, Patient)
Admin/Clinician)
Patient
CBR Engine: Find Similar | | Case?
Patient Registration
No Match
Prediction Page
Hybrid ML Model (SVM + KNN) Generate Prediction
Display Historical Diagnosis
Yes Match
Clinical D
Prediction Result
End

1. Start
   i. Create a Start block at the top of your flowchart. This will mark the beginning of the process.
2. User Sign-Up/Sign-In
   i. Create a Process Box below the "Start" block labeled User Sign-Up/Sign-In.
   ii. Add a Decision Diamond below this to represent Role-Based Access.
   iii. Inside the diamond, add the decision condition: Role: Admin, Clinician, Patient.
3. Role-Based Access (Admin/Clinician vs. Patient)
   i. From the Decision Diamond:
      a) If Admin or Clinician → Lead them to the Patient Registration Page and Prediction Page.
      b) If Patient → Lead them only to the Prediction Page.
4. Patient Registration (Admin/Clinician)
   i. Add a Process Box labeled Patient Registration under the Admin/Clinician path.
   ii. Action: Admin or Clinician enters patient information and stores it in the database.
5. Prediction Page (Clinician)
   i. From both Admin/Clinician and Patient, lead to a Process Box labeled Prediction Page.

ii. This box represents the page where the clinician or patient will input the Patient ID.

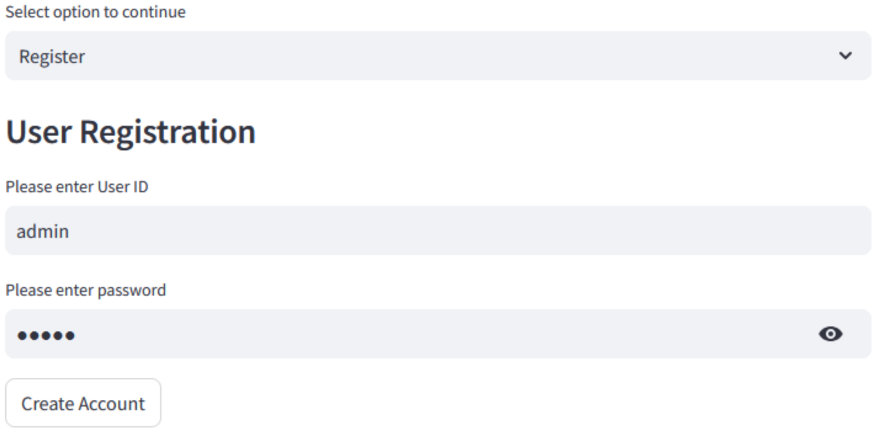
6. Retrieve Patient Bio-Data
   i. Add another Process Box below the Prediction Page labeled Retrieve Patient Bio-Data.
   ii. This box represents fetching the patient's bio-data from the database based on the Patient ID.
7. Clinical Data Input
   i. Add another Process Box labeled Clinical Data Input.
   ii. Inside this box, list the clinical data that needs to be entered (e.g., Tumor Size, Age, Lymph Node Status, etc.).
8. Diagnostic Workflow
   i. Add a Decision Diamond labeled CBR Engine: Find Similar Case? below the Clinical Data Input box.
   ii. Inside the diamond, the decision condition should be: Match Found?
      a) If Match Found → Display Historical Diagnosis.
      b) If No Match Found → Hybrid Machine Learning Model (SVM + KNN).
9. Display Historical Diagnosis (CBR Match Found)
   i. Add a Process Box for the match found case labeled Display Historical Diagnosis.
10. Hybrid Machine Learning Model (SVM + KNN)
   i. Add a Process Box labeled Hybrid Machine Learning Model (SVM + KNN).
   ii. This box processes the data if no match is found by the CBR engine.
11. Prediction Result
   i. Add a final Process Box labeled Prediction Result.
   ii. This box displays the prediction or diagnosis outcome based on the hybrid machine learning model's results.
12. End
   i. Add an End block after the Prediction Result box to mark the end of the flow.

3.5 System Implementation

To facilitate user interaction, the system is deployed as a web application which was built using Streamlit, a Python framework for creating interactive web applications. Streamlit is a Python library designed for the rapid creation of interactive web apps. The user interface allows clinicians or patients to input clinical parameters such as CT scan results, age, tumor size, symptom severity, lymph node status, hormone receptor status, HER2 status, tumor grade, family history, and previous diagnosis. Upon submitting the data, the user can click the Prediction button, and the system will either retrieve a diagnosis based on similar past cases or predict the diagnosis using the trained hybrid model. This was done using the pycharm as programming integrated development environment (IDE). To facilitate secure and personalized access to the breast cancer diagnostic system, a user authentication module is implemented at the initial stage. This includes:

3.5.1 User Sign-Up Page for access control

New users (clinicians or patients) can create an account by providing a username, password, and other required credentials.



Select option to continue

Register

User Registration

Please enter User ID

admin

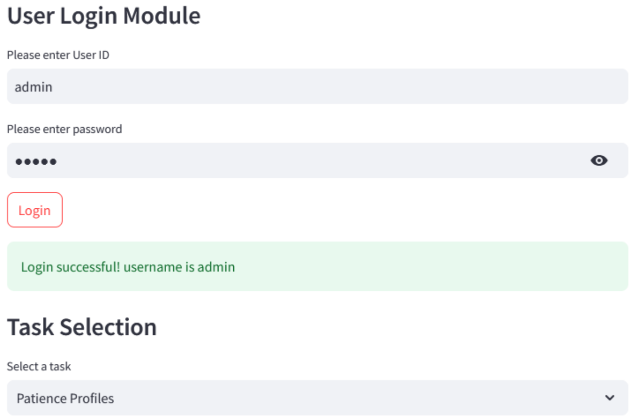
Please enter password

•••••

Figure 2: **User Sign-Up Page**

3.5.2 User Sign-In Page

Registered users can securely log in to the system using their credentials. Role-based access control (e.g., Admin vs. Patient) can be enforced to manage functionalities available to different user categories.



User Login Module

Please enter User ID

admin

Please enter password

•••••

Login successful! username is admin

Task Selection

Select a task

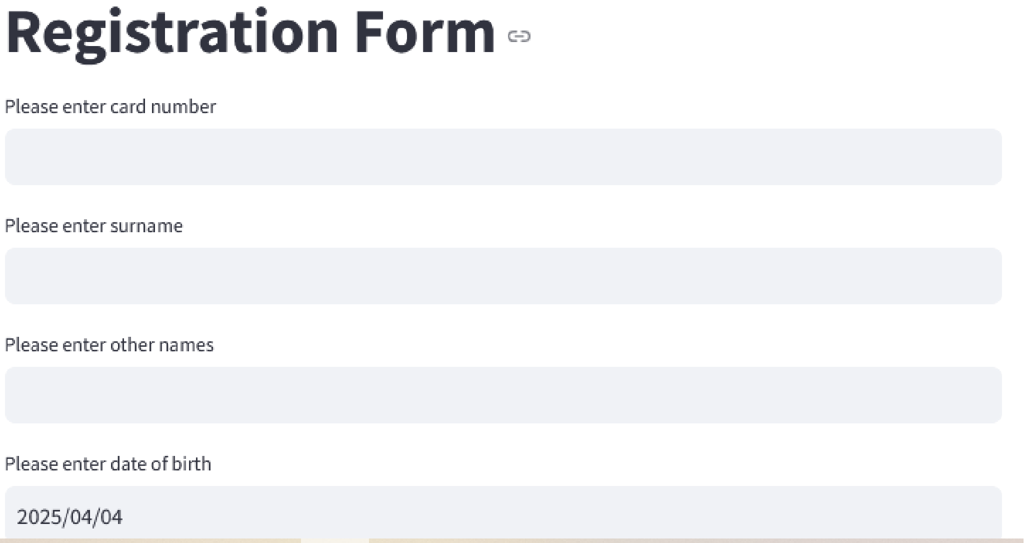
Patience Profiles

Figure 3: **User Sign-In Page**

Upon successful login, users can register new patients biodata using Patient Registration Page, use diagnostic interface and report generation module.

3.5.3 *Patient* Registration Page

Patient Registration: Authorized users (e.g., Admin or Clinician) can add new patient profiles, including essential demographic information (name, age, contact details) and medical history.



Registration Form

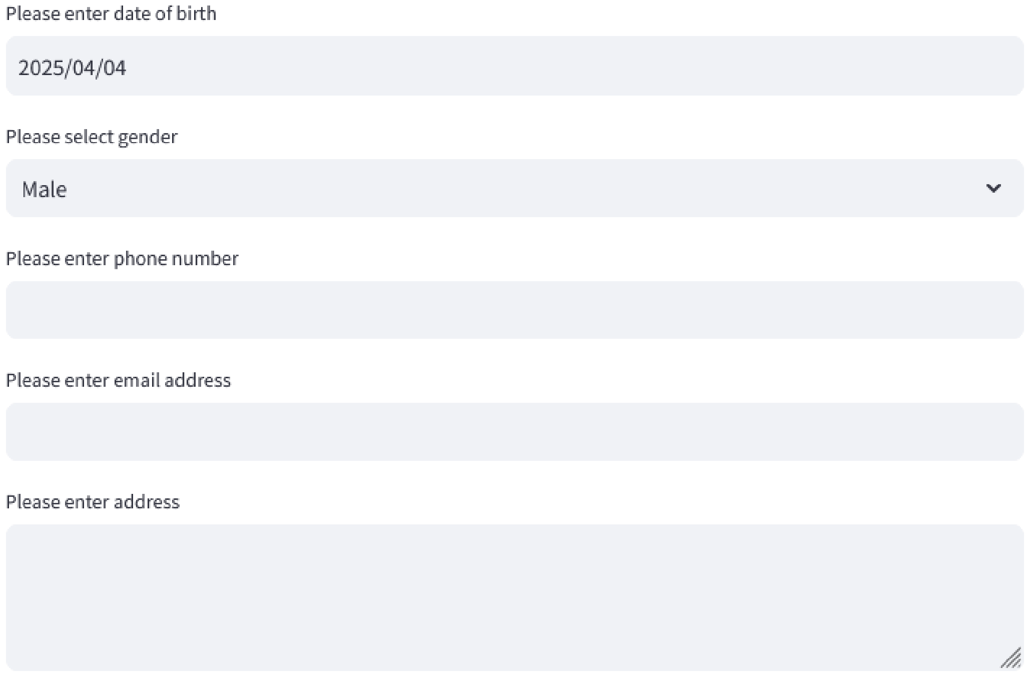
Please enter card number

Please enter surname

Please enter other names

Please enter date of birth

2025/04/04



Please enter date of birth

2025/04/04

Please select gender

Male

Please enter phone number

Please enter email address

Please enter address

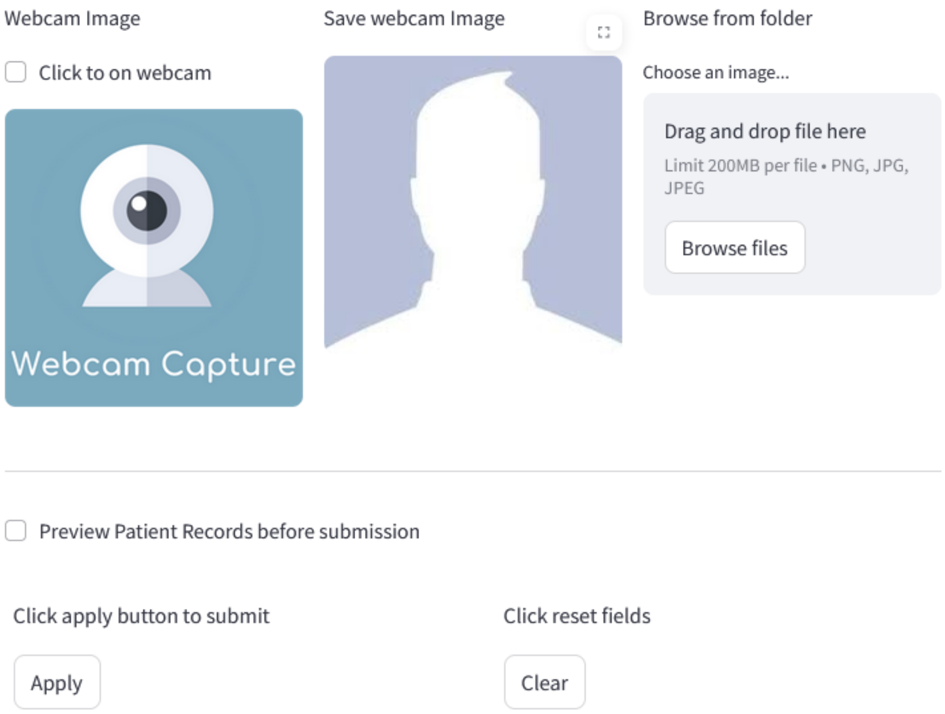
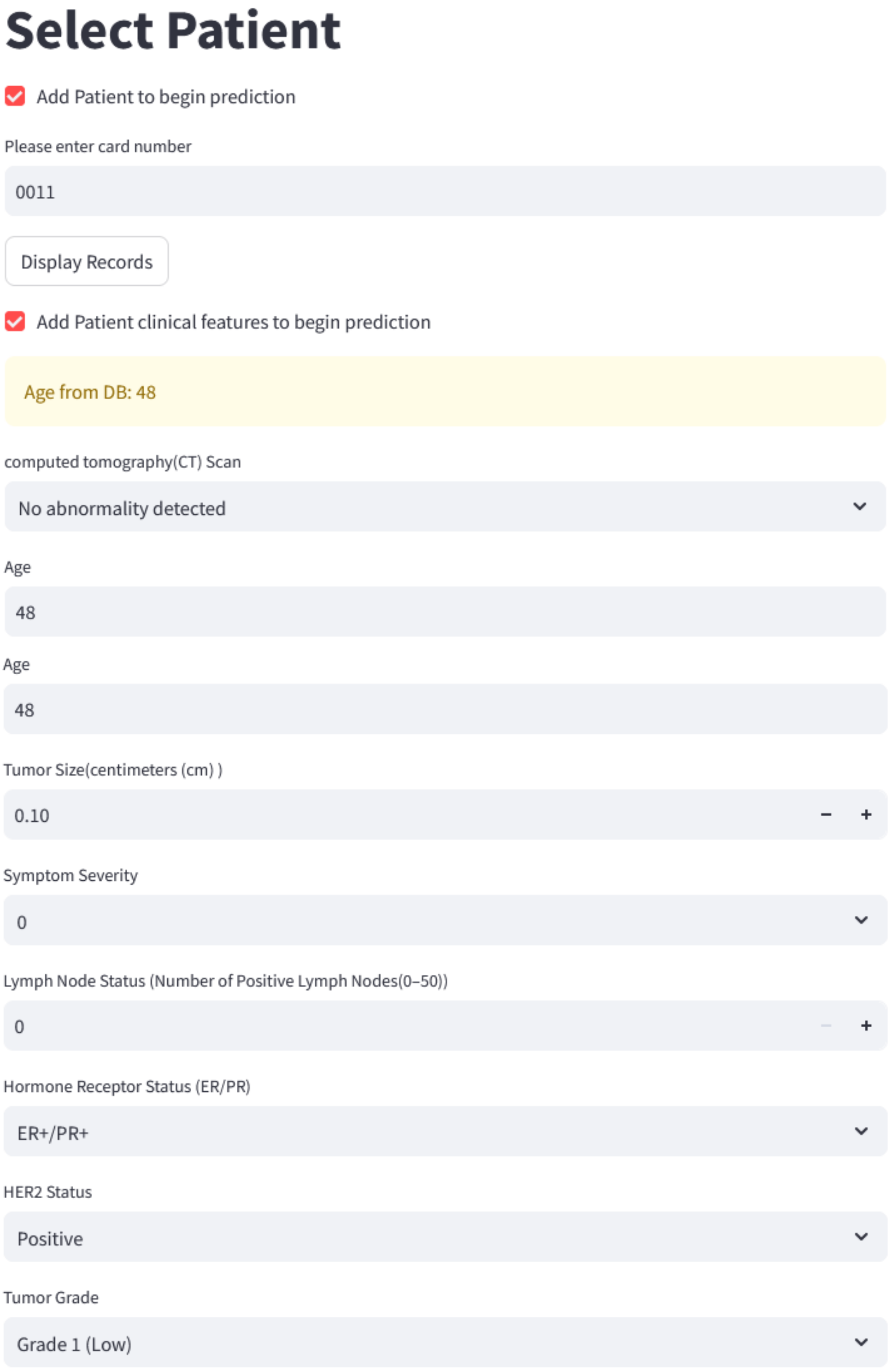


Figure 4: Patient Registration Page

3.5.4 Prediction Page

i. Patient ID Input: Clinicians or patients input a Patient ID to retrieve the associated bio-data, which pre-fills relevant patient information for a smoother experience.

ii. Clinical Data Input: After retrieving the patient's biodata, the system prompts the user to enter essential clinical parameters for prediction (e.g., CT Scan Results, Tumor Size, Lymph Node Status, etc.).



Select Patient

☑ Add Patient to begin prediction

Please enter card number: 0011

☑ Add Patient clinical features to begin prediction

Age from DB: 48

computed tomography(CT) Scan: No abnormality detected

Age: 48

Age: 48

Tumor Size(centimeters (cm)): 0.10 − +

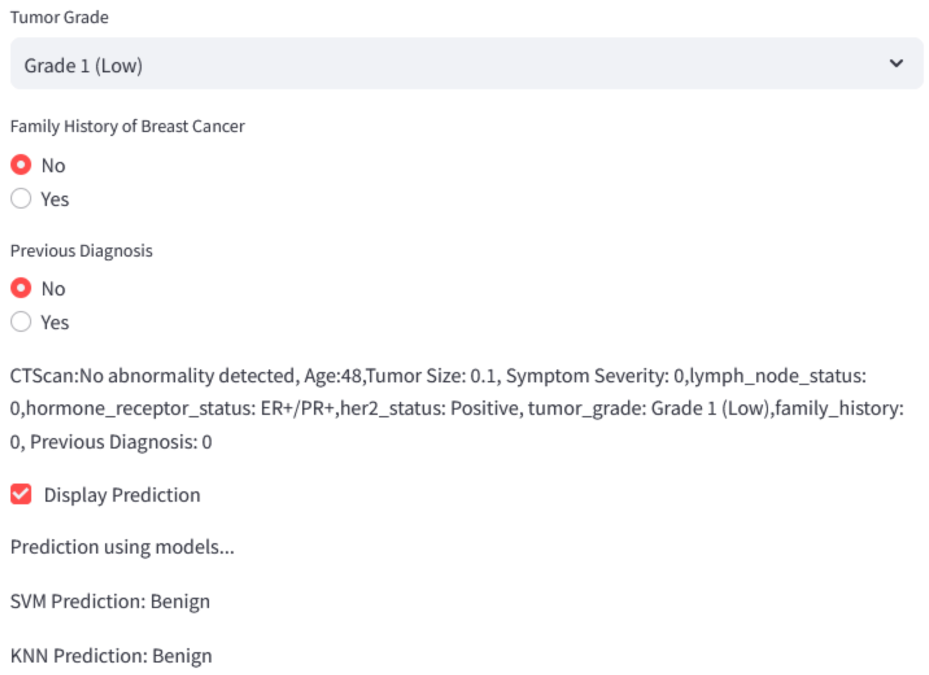
Symptom Severity: 0

Lymph Node Status (Number of Positive Lymph Nodes(0–50)): 0 − +

Hormone Receptor Status (ER/PR): ER+/PR+

HER2 Status: Positive

Tumor Grade: Grade 1 (Low)



Tumor Grade

Grade 1 (Low) ⌄

Family History of Breast Cancer

● No
○ Yes

Previous Diagnosis

● No
○ Yes

CTScan:No abnormality detected, Age:48,Tumor Size: 0.1, Symptom Severity: 0,lymph_node_status: 0,hormone_receptor_status: ER+/PR+,her2_status: Positive, tumor_grade: Grade 1 (Low),family_history: 0, Previous Diagnosis: 0

☑ Display Prediction

Prediction using models...

SVM Prediction: Benign

KNN Prediction: Benign

Figure 5: Prediction Page

Once the input is submitted via the Prediction button, the system initiates a two-tiered diagnostic approach:

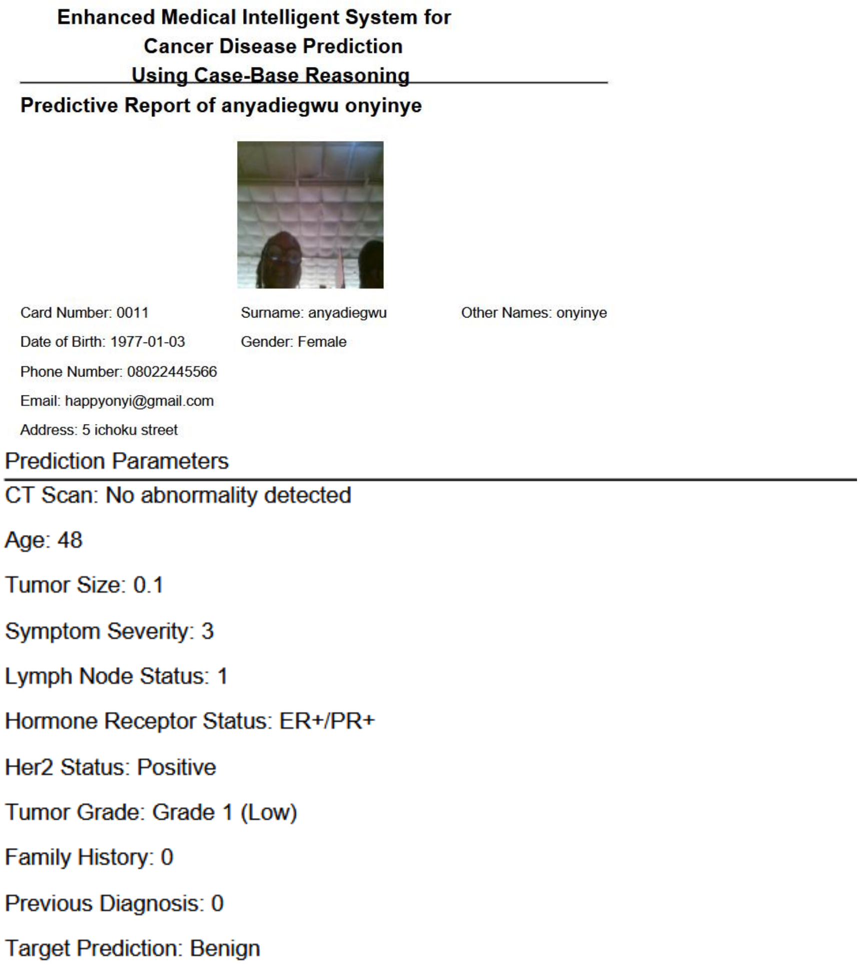
3.5.4.1 Diagnostic Workflow:

i. Case-Based Reasoning (CBR): The system first checks if a similar case exists in the SQLite database using similarity thresholds. If a match is found, the historical diagnosis is retrieved and displayed.

i. Hybrid Machine Learning Model (SVM + KNN): If no similar case is found, the system forwards the clinical data to the hybrid machine learning model, which generates a prediction based on the learned patterns from the training data.

3.5.5 Report Generation

Once a prediction is made, the system also generates a patient report, which includes the predicted diagnosis, confidence levels, and key factors influencing the prediction. This report can be saved or printed for further use, ensuring that clinicians have the necessary information to make informed decisions.



Enhanced Medical Intelligent System for
Cancer Disease Prediction
Using Case-Base Reasoning

Predictive Report of anyadiegwu onyinye

Card Number: 0011 | Surname: anyadiegwu | Other Names: onyinye
Date of Birth: 1977-01-03 | Gender: Female
Phone Number: 08022445566
Email: happyonyi@gmail.com
Address: 5 ichoku street

Prediction Parameters

CT Scan: No abnormality detected

Age: 48

Tumor Size: 0.1

Symptom Severity: 3

Lymph Node Status: 1

Hormone Receptor Status: ER+/PR+

Her2 Status: Positive

Tumor Grade: Grade 1 (Low)

Family History: 0

Previous Diagnosis: 0

Target Prediction: Benign

This structured implementation approach ensures both clinicians have an intuitive interface for secure, efficient, and accurate breast cancer diagnosis, helping improve the overall healthcare experience.

3.6 Results and Discussions

A series of quantitative performance metrics were utilized to evaluate the effectiveness of the hybrid breast cancer prediction system integrating Support Vector Machine (SVM)**,** K-Nearest Neighbors (KNN)**,** and Case-Based Reasoning (CBR)**.** The model was trained and tested on a breast cancer dataset consisting of clinically relevant features such as age, tumor size, symptom severity, HER2 status, lymph node status, and therapy response indicators. A standard train-test split (e.g., 80:20) was employed to validate the model performance.

Evaluation Metrics: The following classification performance metrics were computed and hybrid SVM-KNN model achieves **high predictive performance**, with excellent generalization to unseen patient data. The metrics of the hybrid model is shown in table 1

Table 1: Hybrid SVM-KNN model performance metrics

SN	Metric	Value
1	Accuracy	95.3%
2	Precision	94.1%
3	Recall	96.0%
4	F1-Score	95.0%

i. Accuracy (ACC): Proportion of total correct predictions.
ii. Precision (P): Proportion of true positive predictions among all predicted positives.
iii. Recall (R) or Sensitivity: Proportion of true positives among all actual positive cases.
iv. F1-Score: Harmonic mean of precision and recall, providing a balance between the two.

Table 2: Confusion Matrix of hybrid SVM-KNN model

	Predicted Cancer	**Predicted No Cancer**
Actual Cancer	TP = 480	FN = 20
Actual No Cancer	FP = 25	TN = 475

Confusion Matrix: A tabular summary showing True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). The confusion matrix reveals that the model correctly identified 480 out of 500 cancer cases**,** and 475 out of 500 non-cancer cases**,** with a false negative rate of only 4%**,** which is particularly crucial in a medical context to avoid missed cancer diagnoses.

The results demonstrate that the hybrid SVM + KNN model, supplemented by a Case-Based Reasoning (CBR) system, achieves high predictive performance for breast cancer diagnosis. The integration of SVM, which is robust in high-dimensional spaces, and KNN, which excels in instance-based learning, provides a balanced approach that combines precision and adaptability. The high accuracy, recall, and AUC scores suggest that the system can be trusted for real-world clinical applications, particularly in aiding early detection and decision-making. Furthermore, the feature importance analysis reveals that tumor size, HER2 status, symptom severity, and hormone receptor status are dominant predictive factors. This aligns with clinical domain knowledge, enhancing the system's credibility among medical professionals. The inclusion of CBR adds a valuable layer of interpretability and efficiency by leveraging previous similar cases. This not only allows for faster decision-making when similar scenarios are found but also aligns the system more closely with the way physicians think and diagnose based on experience.

3.7 Limitations

While the results are promising, combining multiple models and a CBR system increases system complexity, which may require more resources and fine-tuning during deployment.

3.8 Conclusion and Recommendation

This study presents the development and evaluation of a hybrid diagnostic system for breast cancer prediction, combining Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) algorithms with a Case-Based Reasoning (CBR) framework. The model leverages the strengths of both SVM (robustness in high-dimensional spaces) and KNN (instance-based learning), providing accurate and interpretable results. Clinical data, including tumor size, hormone receptor status, HER2 status, and CT scan findings, were used as input parameters to generate predictions. The integration of the CBR system allowed the model to reference past similar patient cases stored in an SQLite database. This not only enhanced the speed and contextual accuracy of predictions but also provided interpretability that is aligned with clinical reasoning.

This study contributes to the development of intelligent, interpretable, and practical tools for early breast cancer detection, supporting timely intervention and improved patient outcomes.

3.9 Future Research Direction

To further enhance the current breast cancer diagnostic system and improve its performance and accessibility, the following research directions are recommended:

1. Deep Learning Approaches for Improved Diagnosis:
   i. Exploring Convolutional Neural Networks (CNNs): For more accurate image-based analysis, incorporating CNNs could significantly improve the system's ability to diagnose breast cancer from medical images, such as breast cancer MRI scans.
2. Mobile and Cloud Deployment for Greater Accessibility:
   i. Mobile Deployment: Expanding the breast cancer diagnostic system to mobile platforms (Android/iOS) would provide clinicians and patients with an easy-to-access tool for cancer diagnosis and monitoring. This would make the system more accessible in remote or underdeveloped regions, where access to healthcare professionals and diagnostic tools may be limited.
   ii. Cloud Hosting for Remote Diagnostics: Hosting the system on the cloud would enable remote access for clinicians, patients, and medical practitioners. Cloud deployment would facilitate real-time data sharing, analysis, and collaboration, making it possible to access the diagnostic system from anywhere. This would be especially valuable in providing remote diagnostic support and offering a platform for global collaboration in cancer research and patient care.

REFERENCES

Colleluori, G., Perugini, J., Barbatelli, G., & Cinti, S. (2021). Mammary gland adipocytes in lactation cycle, obesity and breast cancer. *Reviews in Endocrine and Metabolic Disorders*, *22*, 241-255.

Dhatterwal, J. S., Kaswan, K. S., & Preety. (2021). Intelligent agent based case base reasoning systems build knowledge representation in COVID-19 analysis of recovery of infectious patients. *Applications of Artificial Intelligence in COVID-19*, 185-209.

Gu, D., Zhao, W., Xie, Y., Wang, X., Su, K., & Zolotarev, O. V. (2021). A personalized medical decision support system based on explainable machine learning algorithms and ecc features: Data from the real world. *Diagnostics*, *11*(9), 1677.

Gu, Dongxiao, Kaixiang Su, and Huimin Zhao. "A case-based ensemble learning system for explainable breast cancer recurrence prediction." *Artificial Intelligence in Medicine* 107 (2020): 101858.

Łukasiewicz, S., Czeczelewski, M., Forma, A., Baj, J., Sitarz, R., & Stanisławek, A. (2021). Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review. *Cancers*, *13*(17), 4287.

Mustafa, E. M., Saad, M. M., & Rizkallah, L. W. (2023). Building an enhanced case-based reasoning and rule-based systems for medical diagnosis. *Journal of Engineering and Applied Science*, *70*(1), 139.

Patel, V., Chaurasia, V., Mahadeva, R., Ghosh, A., Dixit, S., Suthar, B., ... & Kumar, K. (2023). Breast cancer diagnosis from histopathology images using deep learning methods: a survey. In *E3S Web of Conferences* (Vol. 430, p. 01195). EDP Sciences.

Pesapane, F., Tantrige, P., Rotili, A., Nicosia, L., Penco, S., Bozzini, A. C., ... & Cassano, E. (2023). Disparities in breast cancer diagnostics: how radiologists can level the inequalities. *Cancers*, *16*(1), 130.

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In *Machine learning* (pp. 101-121). Academic Press.

Singh, A., & Roghini, S. (2023). Cancer: Unraveling the Complexities of Uncontrolled Growth and Metastasis. *PEXACY International Journal of Pharmaceutical Science*, *2*(8), 59-73.

Wilkerson, Z. (2023). Exploring Deep Learning-Based Feature Extraction for Case-Based Reasoning Retrieval. In *ICCBR Workshops* (pp. 228-232).

Wilkinson, L., & Gathani, T. (2022). Understanding breast cancer as a global health concern. *The British journal of radiology*, *95*(1130), 20211033.

Xu, C., Liu, W., Chen, Y., & Ding, X. (2022). A supervised case-based reasoning approach for explainable thyroid nodule diagnosis. *Knowledge-based systems*, *251*, 109200.