A Project Report on

Secure Machine Learning for Pathological Assessment

submitted in partial fulfillment for the award of

Bachelor of Technology

in

Computer Science and Engineering

by

K Mowlya Sai Sundari (Y21ACS466) **M Asritha (Y21ACS506)**

K Venkata Manoj (Y21ACS464) **K Kannaiah (Y21ACS484)**

Under the guidance of

Mrs. M. Karuna, M. Tech (PhD)

Asst. Prof

Department of Computer Science and Engineering

Bapatla Engineering College

(Autonomous)

(Affiliated to Acharya Nagarjuna University)

BAPATLA – 522 102, Andhra Pradesh, INDIA

2024-2025

Department of
Computer Science and Engineering

CERTIFICATE

This is to certify that the project report entitled **Secure Machine Learning for Pathological Assessment** that is being submitted by K Mowlya Sai Sundari(Y21ACS466), M Asritha(Y21ACS506),K Venkata Manoj(Y21ACS464)and K Kannaiah(Y21ACS486) in partial fulfillment for the award of the Degree of Bachelor of Technology in Computer Science and Engineering to the Acharya Nagarjuna University is a record of bonafide work carried out by them under our guidance and supervision.

Date:

Signature of the Guide
Mrs. M. Karuna
Asst. Professor

Signature of the HOD
Dr. M. Rajesh Babu
Associate Professor

DECLARATION

We declare that this project work is composed by ourselves, that the work contained herein is our own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

K. Mowlya Sai Sundari(Y21ACS466)

M. Asritha(Y21ACS506)

K. Venkata Manoj(Y21ACS464)

K. Kannaiah(Y21ACS484)

Acknowledgement

We sincerely thank the following distinguished personalities who have given their advice and support for successful completion of the work.

We are deeply indebted to our most respected guide **Mrs. M. Karuna**, Asst. Professor, Department of CSE, for her valuable and inspiring guidance, comments, suggestions and encouragement.

We extend our sincere thanks to **M. Rajesh Babu**, Assoc. Prof. & Head of the Dept. for extending his cooperation and providing the required resources.

We would like to thank our beloved Principal **Dr. N. Rama Devi** for providing the online resources and other facilities to carry out this work.

We would like to express sincere thanks to our project coordinator **Dr. P. Pardhasaradhi,** Prof. Dept. of CSE for his helpful suggestions in presenting this document.

We extend our sincere thanks to all other teaching faculty and non-teaching staff of the department, who helped directly or indirectly for their cooperation and encouragement.

K. Mowlya Sai Sundari(Y21ACS466)

M. Asritha(Y21ACS506)

K. Venkata Manoj (Y21ACS464)

K. Kannaiah(Y21ACS484)

Table of Contents

List of Figures

List of Tables

List of Equations

Abstract

Breast cancer is one of the most common and life-threatening diseases among women globally, thus early and precise detection is critical to improve survival rates and treatment results. To do this, a secure and efficient breast cancer categorization system is created by combining machine learning and cryptographic methods. Unlike standard methods that use Support Vector Machines (SVM) for model training, Random Forest and XGBoost are used to improve predictive accuracy. However, securing sensitive patient data remains a top priority in medical applications. To overcome this, CKKS homomorphic encryption is used, which enables encrypted data processing without the need for decryption, ensuring data privacy and security. Before training the XGBoost model, the dataset is thoroughly preprocessed, including missing value management, feature standardization, and SMOTE-based class imbalance correction. During prediction, fresh patient data is encrypted and securely sent, and homomorphic computations are carried out without disclosing sensitive information. The encrypted results are then returned, decrypted, and evaluated to determine the final diagnosis. To validate the effectiveness of this strategy, performance is measured using accuracy, precision, recall, and F1-score, resulting in a highly accurate and privacy-preserving breast cancer detection system. By merging modern machine learning models with safe cryptographic approaches, this framework provides both medical dependability and patient data confidentiality, providing a reliable solution for breast cancer diagnosis.

Keywords: Breast Cancer, Machine Learning, Random Forest, XGBoost, Homomorphic Encryption, CKKS, Data Privacy, Predictive Accuracy, SVM, Feature Standardization, SMOTE, Class Imbalance, Accuracy, Precision, Recall, F1-Score, Medical Diagnosis.

1 Introduction

Medical diagnosis, especially of breast cancer, has emerged as a key domain of interest for contemporary healthcare. As breast cancer is among the most common malignancies in females worldwide, correct and early detection is crucial in enhancing survival as well as optimal treatment strategies. Medical data and their complexity as well as sensitiveness, on the other hand, present big challenges in conceiving automated schemes that are simultaneously accurate and privacy-protective.

This research intends to investigate and propose a secure and smart system for breast cancer classification by combining enhanced machine learning methodologies with contemporary cryptography techniques. Using a numerical data set with diagnostic features and using preprocessing methods like mean imputation, standardization, and class balancing using SMOTE, our intention is to develop models capable of learning from past experiences and correctly classifying whether a tumor is benign or malignant.

The use of highly effective machine learning models, such as but not limited to Random Forest and XGBoost Classifier, facilitates the development of stable predictive models able to detect subtle patterns in health data. In order to maintain patient confidentiality, the architecture leverages CKKS homomorphic encryption based on the TenSEAL library, providing support for computation in encrypted form—keeping the data private even at the stage of model inference. This solution answers one of the most important challenges in healthcare AI: ensuring the secure processing of sensitive data.

Lastly, this work hopes to advance the emerging discipline of secure medical AI systems and provide insightful contributions towards privacy-aware diagnostic tools for health practitioners,

scientists, and programmers. By interlacing machine learning strengths and cryptographic technology prowess, we hope to make breast cancer diagnosis systems more trustworthy, accurate, and secure—providing stronger and more confidential decision-support for clinicians and patients alike.

1.1 Machine Learning

Machine learning (ML) is a form of artificial intelligence (AI) that allows computers to learn automatically from data and enhance their performance on a given task over time, without explicit programming. Machine learning algorithms, in the context of medical diagnosis, can sift through large amounts of clinical data, identify patterns, and provide predictions that can help healthcare professionals make informed decisions. Such algorithms are used to discern latent connections in data sets and produce correct results based on historical observations.

In conventional programming, a human coder creates step-by-step instructions to instruct the computer on how to perform a task. Such instructions tend to follow a sequential logic or rule-based system, e.g., an if-then logic. However, machine learning is based on data-driven methods where models are trained on past data to discover patterns, associations, and relationships that can then be generalized to new cases.

Whereas artificial intelligence is a generic field that encompasses all technologies simulating human intelligence, machine learning is a more specialized subset involving algorithms that learn from and predict based on data. These models change and grow as additional data is added to them, rendering them extremely efficient in dynamic and data-intensive situations such as in healthcare.

1.1.1 Supervised Learning

Supervised learning is a form of machine learning where the model learns from a labeled dataset. The training set's each data point is associated with a corresponding label, allowing the model to acquire the mapping of inputs to desired outputs. With training complete, the model will predict the output for new, unseen data by applying the learned patterns.

This method is employed extensively in classification problems, such as predicting that a tumor is benign or cancerous from diagnostic characteristics. After the model's training and validation, the trained model can subsequently be tested with test data in order to quantify how well the model makes predictions.

Supervised learning is particularly valuable in medical diagnosis, where labeled data sets—usually curated and validated by subject matter experts—can be used to train models that replicate clinical decision-making with high accuracy and consistency.

1.1.2 Unsupervised Learning

Unsupervised learning focuses on exploration and pattern discovery in unlabeled data or data without known outcomes. Unlike supervised learning, where models are trained using labeled outputs, unsupervised learning methods need to find hidden patterns, clustering, or relationships within the data without any external information.

A ubiquitous technique in unsupervised learning is clustering, in which like points are categorized together due to commonalities. In medicine, this may be applied for the identification of disease subtypes, the determination of associations between symptoms, or the identification of trends in patient records that aren't necessarily immediately obvious.

While unsupervised learning isn't normally applied to classification tasks such as breast cancer diagnosis per se, it can be very useful in the exploratory data analysis step in order to know the distribution and relationships of the data prior to using supervised models.

1.2 Introduction to Cryptography

Cryptography is both the study and practice of methods for protecting communication and data against adversaries. It is the process of transforming information into a mode that is not readable to unauthorized users, such that only designated recipients can read and interpret the original message. Cryptography is extremely important in information security as it achieves confidentiality, integrity, authentication, and non-repudiation.Contemporary cryptographic techniques are extensively used across different fields, such as finance, communications, and health, to encrypt sensitive data like passwords, money information, and medical records against cyber attacks.

1.2.1 Role of Cryptography in Data Security

a. Cryptography ensures that data is encrypted before transmission or storage, making it inaccessible to unauthorized users.

b. Even if intruders intercept the data, they are not able to interpret its meaning without the related decryption key.

c. Such a process protects sensitive data such as patient histories in medical use or financial information in banking platforms.

1.2.2 Types of Cryptography

a. Symmetric Cryptography – Utilizes the identical key for encryption and decryption. It is fast

and suitable for encrypting a lot of data but calls for secure key sharing among parties.

b. Asymmetric Cryptography – Requires a set of two keys: one for public encryption and one for private decryption. It improves security since no secret key need be shared.

c. Homomorphic Encryption – A relatively new cryptographic method that enables computation to be executed directly on ciphertext without first decrypting it. This is particularly valuable in privacy-sensitive use cases such as medical diagnosis.

1.2.3 Importance of Cryptography in Healthcare

a. Secures patient information in electronic health records and diagnosis programs from unauthorized use.

b. Provides secure communication among physicians, laboratories, and patients through digital platforms.

c. Supports privacy-preserving machine learning by encrypting sensitive health information while still enabling model training and prediction, as applied in this project.

1.2.4 Real-World Applications of Cryptography

a. Healthcare – Protected storage and transportation of healthcare records, privacy-aware diagnosis employing mechanisms such as Homomorphic Encryption.

b. Banking and Finance – Securing online transactions, digital signatures, and access to accounts.

c. Communication – End-to-end encryption of messages in messaging platforms and secure mail services to ensure no eavesdropping.

d. E-Governance – Reliable digital identities and document signing for applications such as e-voting and digital certificates.

1.3 Homomorphic Encryption

Homomorphic encryption is a sophisticated cryptographic method through which computations can be carried out directly on encrypted information without decrypting it first. This implies that sensitive information is kept confidential throughout the process—right from analysis and processing.

In classical encryption, the data would have to be decrypted before they can be used, posing a potential threat. Homomorphic encryption avoids this risk by allowing for operations such as addition and multiplication to be performed on the encrypted data, and when decrypted, the output would be equal to the result if the operations had been conducted directly on the unencrypted original data.

1.3.1 Types of Homomorphic Encryption

a. **Partially Homomorphic Encryption (PHE)** – Only supports one operation type, e.g., either addition or multiplication. Illustration: RSA supports multiplication.

b. **Somewhat Homomorphic Encryption (SHE)** – Supports a restricted number of operations (several additions and multiplications) before the ciphertext is too noisy to be viable.

c. **Fully Homomorphic Encryption (FHE)** – Allows addition and multiplication an unlimited number of times on encrypted data. It is strongly secure but very computationally demanding.

d. **CKKS Scheme (Cheon–Kim–Kim–Song)** – An approximate homomorphic encryption of the type particularly designed for real-number arithmetic, e.g., used in machine learning models. It is the scheme employed in our project for secure medical diagnosis.

1.3.2 Advantages of Homomorphic Encryption

a. Data Privacy – Sensitive information (e.g., patient health records) can be analyzed without revealing the raw data.

b. Security in Outsourced Computation – Facilitates secure processing of data in untrusted environments, e.g., cloud platforms.

c. Compliance – Assists in compliance with data protection laws such as HIPAA and GDPR by keeping data encrypted across the lifecycle.

1.3.3 Applications of Homomorphic Encryption

a. Medical Diagnosis – Patient information can be passed into disease prediction machine learning models securely without invading patient confidentiality.

b. Financial Services – Secure computation on encrypted financial information for fraud detection or risk analysis.

1.4 Objective

This project will create a safe and precise medical diagnosis system for breast cancer classification using machine learning models combined with homomorphic encryption. Through the examination of enriched numerical data obtained from the Wisconsin Breast Cancer Diagnostic dataset, the system employs advanced classifiers like Random Forest and XGBoost to accurately predict disease outcomes.

In order to provide data privacy and alignment with healthcare data protection requirements, the project uses the CKKS homomorphic encryption scheme so that encrypted data can be utilized for prediction without revealing sensitive patient information.

2 Literature Review

Over the last few years, the combination of cryptographic methods and machine learning has been a key factor in making privacy-preserving medical applications possible. The increased threat to the security of confidential health information has resulted in the creation of privacy-enhanced models for disease diagnosis and prediction.

Al Badawi and Faizal Bin Yusof [1] proposed a privacy-preserving framework for pathological evaluation with machine learning and homomorphic encryption. Their system supports encrypted inference on patient data without exposing any sensitive information, thereby maintaining data confidentiality standards in healthcare. This research formed the basis for integrating encrypted computation with clinical prediction systems.

The homomorphic encryption algorithm employed in most such models is the CKKS (Cheon-Kim-Kim-Song) scheme, proposed by Cheon et al. [2]. The scheme enables approximate arithmetic computations directly over ciphertext, and so it is most suitable for machine learning inference computations, particularly real-valued numerical data computations.

The Wisconsin Breast Cancer Diagnostic dataset, initially gathered by Wolberg and Mangasarian [3], [4], is still among the most commonly used datasets in breast cancer classification research. It contains real-world diagnostic features and labeled outcomes needed to train and test ML models.

Apart from that, various research has exhibited the possibility of integrating support vector machines and encryption. Zhu et al. [5] described a nonlinear SVM-based secure medical prediagnosis system with encrypted inputs, enjoying high classification accuracy while securing

patient information. Zhang et al. [6] took it to the next level by developing a privacy-preserving multiclass SVM framework that was appropriate for clinical diagnosis on cloud platforms.

Chen and Zheng [7] applied a linear regression model to encrypted data and proved that simple ML models could be efficiently migrated to encrypted environments. Their contribution illustrates the possibility of privacy-preserving machine learning (PPML) in secure areas like medicine.

In addition, several other works investigated the application of homomorphic encryption in large-scale biomedical contexts. Gursoy et al. [8] and Blatt et al. [9] discussed genomic privacy by applying full homomorphic encryption (FHE) to genome-wide association study and genotype imputation, respectively. These pieces of work stress the general feasibility of encrypted ML in healthcare and bioinformatics.

More recent breakthroughs in encrypted deep learning by Lee et al. [10] also demonstrate increased interest in ensuring that AI and developed a support vector machine classification method that can function solely on encrypted datasets, resulting in good classification accuracy without compromising data confidentiality.

While all these contributions form a strong foundation for secure and private AI in medicine, few of them are specifically on breast cancer diagnosis or do not leverage high-performance ensemble models like Random Forest or XGBoost. Our contribution attempts to bridge this gap through creating a strong breast cancer prediction system utilizing high-performance classifiers and leveraging CKKS homomorphic encryption for secure inference over healthcare data.

3 Problem Statement

The increasing global incidence of breast cancer has highlighted the need for more accurate and accessible diagnostic options. Despite advances in medical technology, current screening procedures frequently fall short of providing early and precise detection, particularly in low-resource settings. Traditional techniques, such as manual mammography assessments and rudimentary machine learning models, are prone to discrepancies, especially where patient data privacy is critical.

One of the most difficult challenges is finding a balance between diagnostic accuracy and data security. Many AI-driven solutions require access to raw medical data, putting healthcare networks at risk of data breaches and privacy abuses. Malicious exploitation of these vulnerabilities can compromise both individual patient privacy and the overall healthcare system. Furthermore, traditional models like SVM frequently hit their performance limits, potentially ignoring early malignancy signs.

Advanced models, such as XGBoost, improve diagnostic performance, but they rely on centralized data processing, which violates tight privacy rules such as HIPAA and GDPR. Medical datasets complicate matters further by introducing missing values and class imbalances. This underscores the critical need for a safe, privacy-preserving diagnostic system that achieves high accuracy while maintaining patient anonymity. Such a solution should smoothly combine clinician interpretability with strong protection against illegal data access

4 System Analysis

Machine learning has also emerged as a critical tool in the medical field, particularly for life critical tasks like the classification of breast cancer. Yet although these models deliver accurate predictions, they usually don't cover the important side of data privacy. With the highly sensitive nature of medical records, the necessity for secure and privacy-protection diagnostic systems is now greater than ever. This work presents a secure machine learning system that supports breast cancer prediction while maintaining patient data confidentiality through homomorphic encryption, in this case, the CKKS scheme.

4.1 Existing System

Breast cancer classifiers have conventionally employed machine learning algorithms like Random Forest, Support Vector Machine, or XGBoost on clinical datasets like the Wisconsin Breast Cancer Diagnostic dataset. These algorithms review patient feature value data to detect tumors as being benign or malignant. In a majority of present systems, data from patients are either processed at the local sites or sent across in plain form to cloud sites where inference happens. Although these systems can be made highly accurate, they do not have the means to provide data privacy during communication and processing phases. Without encryption, data is vulnerable to threats while being transferred, resulting in severe privacy issues and rendering the system non-compliant with regulations like HIPAA or GDPR.

4.2 Limitations

The primary shortcomings of existing systems are their failure to safeguard sensitive patient data. Because the data is transmitted without encryption, it is susceptible to interception, tampering, or

misuse. These systems also suffer from model inversion attacks in which attackers try to use model outputs to recover original data. Moreover, the lack of secure protocols bars such models from being used in actual clinical settings that require strict confidentiality. Lastly, although conventional models provide high predictive accuracy, they compromise interpretability and data protection, which are necessary for medical use.

4.3 Proposed System

The suggested system embeds privacy-preserving computation into the breast cancer diagnosis pipeline. It allows secure prediction via CKKS homomorphic encryption along with efficient machine learning algorithms such as XGBoost and Random Forest. In the proposed architecture, patient data is encrypted on the client side prior to being transferred to the server. The server never gets the raw data. Rather, it carries out encrypted inference with a homomorphically compatible variant of the machine learning model. The output, also in encrypted format, is then sent back to the client where it is decrypted and presented. Through this process, the system ensures that patient information is kept confidential throughout.

The preprocessing step involves critical steps like missing value handling with mean imputation, label encoding the target variable, feature value standardization, and class balancing using SMOTE. These preprocessing methods enhance model accuracy and minimize bias. The preprocessed data is then utilized to train individual Random Forest and XGBoost models, which are subsequently used for homomorphic inference. This method not only provides correct classification outcomes but also ensures complete privacy, which makes it ideal for real-world deployment in clinical environments.

4.4 Architecture

The architecture of the system consists of a server and client setting, as illustrated by the data flow diagram in Figure 4.1. At the client level, CKKS encryption parameters are initialized first. The user enters new patient data, which are promptly encrypted through the CKKS scheme. This encrypted patient record is securely sent to the server. At the server side, the data is processed through inference via a homomorphically compatible machine learning model, either Random Forest or XGBoost. The prediction is calculated based on all encryptions, without any leakage of sensitive information. The encrypted output is passed back to the client, where it is decrypted and shown to the user. The overall architecture guarantees that data confidentiality is preserved throughout the whole diagnostic process, and raw data never leaves the user's device insecure.

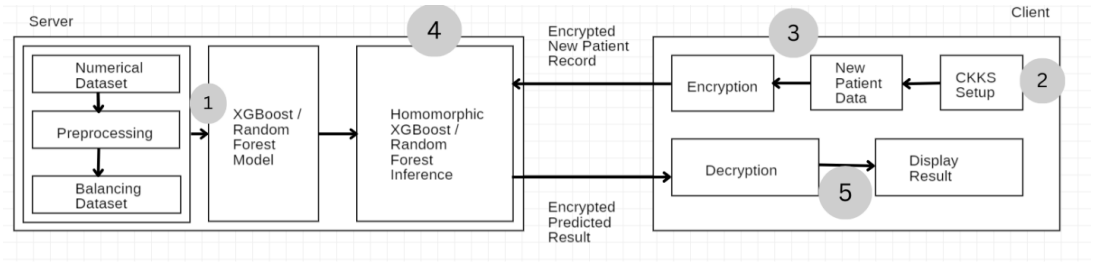

Figure 4.1 Architecture

Figure 4.1 depicts this architecture and emphasizes the seamless integration between secure encryption and effective machine learning inference, allowing a strong and privacy-conscious diagnostic system for breast cancer classification.

5 Methodology

This chapter discusses the procedures and methods used to create a privacy-protecting breast cancer classifier based on machine learning models supplemented with homomorphic encryption. The aim of this project is to create a secure medical diagnostic device that can efficiently predict the risk of breast cancer while maintaining confidentiality of the patient's data throughout processing.

A. First, we downloaded the dataset from the Kaggle website, namely the Wisconsin Breast Cancer Diagnostic dataset, which has diagnostic measurements of breast tumor cell data. For additional richness in the dataset, 12 more numerical features were incorporated, and thus the data was made stronger for classification purposes.

B. Once we obtained the dataset, we performed preprocessing. It involved missing value handling using mean imputation, label encoding target labels (malignant or benign), and feature value standardization to have uniformity. We also used SMOTE (Synthetic Minority Over-sampling Technique) for balancing the dataset to avoid model training bias owing to class imbalance.

C. After the data was cleaned and prepared, we trained two machine learning models: Random Forest and XGBoost. Both models were trained on the enriched dataset to determine the best classifier. These models were then transformed into a homomorphic encryption-compatible form.

D. During the last phase, we deployed secure inference with the CKKS homomorphic encryption scheme via the TenSEAL library. In this system, users encrypt their diagnostic input values

locally on the client side before sending them to the server. The server directly performs inference on the encrypted values and sends back the encrypted prediction, which is decrypted locally by the client.

This method guarantees that the data are never revealed in their raw state, ensuring absolute privacy during the diagnostic process while still providing precise classification.

5.1 Dataset

A dataset is an organized set of data to be used in analysis and training models. We utilized the Wisconsin Breast Cancer Diagnostic dataset for this project, which includes precise diagnostic measurements of breast tumor samples. The dataset was downloaded from Kaggle and then augmented with 12 extra numerical features concerning tumor attributes to enhance model performance.

Every row of the dataset denotes a patient sample, and every column denotes a particular diagnostic attribute extracted from fine needle aspirate (FNA) images of breast mass. The dataset has features like mean radius, texture, perimeter, area, smoothness, compactness, concavity, and symmetry, and so on. The target column denotes whether or not the tumor is benign or malignant.

The large dataset was saved in an Excel file and comprised solely numeric data, making it compatible for machine learning model training and secure inference through the CKKS encryption scheme.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	id	diagnosis	radius_me	texture_m	perimeter	area_mea	smoothne	compactn	concavity_	concave p	symmetry	fractal_dir	radius_se	texture_se	perimeter	area_se	smoothne	compactn	concavity_
2	842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095	0.9053	8.589	153.4	0.006399	0.04904	0.05373
3	842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.0186
4	84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.00615	0.04006	0.03832
5	84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956	1.156	3.445	27.23	0.00911	0.07458	0.05661
6	84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.01149	0.02461	0.05688
7	843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345	0.8902	2.217	27.19	0.00751	0.03345	0.03672
8	844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467	0.7732	3.18	53.91	0.004314	0.01382	0.02254
9	84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835	1.377	3.856	50.96	0.008805	0.03029	0.02488
10	844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063	1.002	2.406	24.32	0.005731	0.03502	0.03553

Fig 5.1 Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
560	925277	B	14.59	22.68	96.39	657.1	0.08473	0.133	0.1029	0.03736	0.1454	0.06147	0.2254	1.108	2.224	19.54	0.004242	0.04639	0.06578
561	925291	B	11.51	23.93	74.52	403.5	0.09261	0.1021	0.1112	0.04105	0.1388	0.0657	0.2388	2.904	1.936	16.97	0.0082	0.02982	0.05738
562	925292	B	14.05	27.15	91.38	600.4	0.09929	0.1126	0.04462	0.04304	0.1537	0.06171	0.3645	1.492	2.888	29.84	0.007256	0.02678	0.02071
563	925311	B	11.2	29.37	70.67	386	0.07449	0.03558	0	0	0.106	0.05502	0.3141	3.896	2.041	22.81	0.007594	0.008878	0
564	925622	M	15.22	30.62	103.4	716.9	0.1048	0.2087	0.255	0.09429	0.2128	0.07152	0.2602	1.205	2.362	22.65	0.004625	0.04844	0.07359
565	926125	M	20.92	25.09	143	1347	0.1099	0.2236	0.3174	0.1474	0.2149	0.06879	0.9622	1.026	8.758	118.8	0.006399	0.0431	0.07845
566	926424	M	21.56	22.39	142	1479	0.111	0.1159	0.2439	0.1389	0.1726	0.05623	1.176	1.256	7.673	158.7	0.0103	0.02891	0.05198
567	926682	M	20.13	28.25	131.2	1261	0.0978	0.1034	0.144	0.09791	0.1752	0.05533	0.7655	2.463	5.203	99.04	0.005769	0.02423	0.0395
568	926954	M	16.6	28.08	108.3	858.1	0.08455	0.1023	0.09251	0.05302	0.159	0.05648	0.4564	1.075	3.425	48.55	0.005903	0.03731	0.0473
569	927241	M	20.6	29.33	140.1	1265	0.1178	0.277	0.3514	0.152	0.2397	0.07016	0.726	1.595	5.772	86.22	0.006522	0.06158	0.07117
570	92751	B	7.76	24.54	47.92	181	0.05263	0.04362	0	0	0.1587	0.05884	0.3857	1.428	2.548	19.15	0.007189	0.00466	0

Fig 5.2 Dataset

5.2 Preprocessing

Data preprocessing is the most important phase of developing a robust and efficient prediction model. In our project, the raw diagnostic data of breast cancer frequently included missing values, class imbalance, and feature variation in scale. These issues had to be resolved to ready the dataset for correct classification with the help of machine learning methods. The preprocessing step had various sub-steps including data cleaning, feature selection, feature engineering, normalization, and data splitting, all designed to improve the model's predictive power.

5.2.1 Data Cleaning

The process of cleaning data was vital to maintain the accuracy and integrity of the dataset used for diagnosis of breast cancer. Missing values in the dataset were managed with mean imputation, where the missing values were replaced with the mean of corresponding columns.

The approach ensured all useful records were preserved without any significant bias introduced. We also checked and removed any duplicate records that might introduce bias in the training process. As the dataset was numerical and clean, there were no inconsistencies or irrelevant entries that needed further deletion.

5.2.2 Feature Selection

Feature selection is essential to minimize dimensionality and enhance the performance of the model. The initial dataset contained a number of diagnostic features obtained from breast tissue images. Every feature explained an attribute of the cell nuclei, including radius, texture, perimeter, area, smoothness, concavity, and symmetry. We chose a mix of both original and 12 more engineered numerical features, with a focus on those that had high correlation with the target output. The target feature was whether the tumor was benign or malignant, which was represented numerically during model training.

5.2.3 Feature Engineering

Feature engineering in this work entailed augmenting the dataset with more numeric features using domain knowledge. We calculated statistical aggregates like mean, standard error, and worst values over original features to identify more subtle patterns. These new features assisted the model in identifying more subtle relationships between diagnostic measurements and malignancy probability. This step greatly enhanced data richness and facilitated the development of models with increased predictive capability.

5.2.4 Normalization

To ensure that all features are scaled consistently, we used standardization, a normalization process that rescales the data to have a mean of 0 and standard deviation of 1. Standardization was important because features such as area and radius were larger in numerical value than features like smoothness and symmetry, potentially overwhelming the model training otherwise. Through bringing all features to the same scale, we ensured that the model treated all inputs equally, resulting in stable and accurate predictions.

5.2.5 Data Splitting

The last operation in the preprocessing stage was dividing the dataset into training and test sets. We employed a split ratio of 80:20, using 80% of the data for training the machine learning algorithms (Random Forest and XGBoost) and setting aside 20% of the data for testing model performance. This method guaranteed that the models trained were tested on unseen data, enabling us to determine their generalization ability before they could be used in a privacy-preserving environment based on homomorphic encryption.

5.3 Algorithms

The last step in the suggested system is classification, where the processed and cleaned data is fed into various machine learning models to train and test their performance. The algorithms employed are Random Forest, XGBoost, and CKKS Homomorphic Encryption. Random Forest and XGBoost are supervised learning algorithms employed for binary classification problems like predicting whether a tumor is benign or malignant. These models were trained and tested on the same feature-rich dataset

5.3.1 Random Forest

Random Forest is a widely used ensemble machine learning algorithm that can be used for both classification and regression problems. It works by creating multiple decision trees while training and aggregating their predictions for enhanced accuracy and prevention of overfitting. The decision trees are each built from a different random subset of the dataset, so the model becomes stronger and less prone to noisy data and outliers.

While training, every tree in the forest is trained on a bootstrapped sample of the data and, at every split in the tree, a random subset of features is examined. This introduces randomness to form a diverse ensemble of learners that improves the predictive performance overall. The final prediction for classification is done by majority voting, and for regression, it is done by the average of predictions by all individual trees.

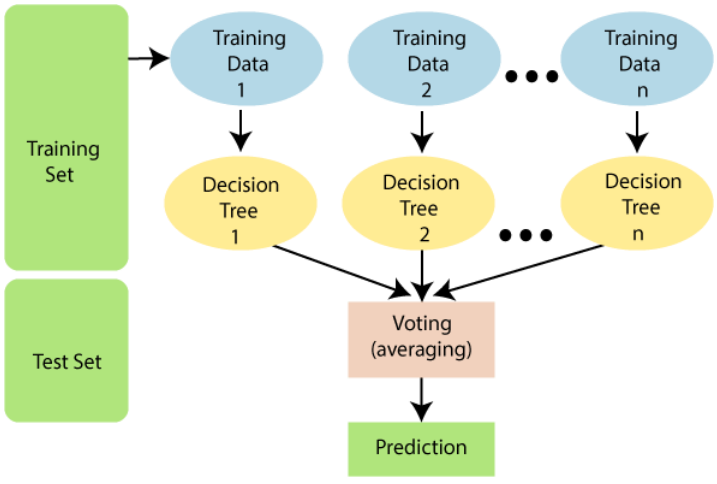

Fig 5.3 Random Forest

The power of Random Forest is that it can generalize well without needing extensive

hyperparameter tuning. It also gives information about the importance of every feature, which assists in interpreting the decisions of the model. In our project, the Random Forest classifier was trained on the processed and encrypted dataset, and it performed well in classifying breast cancer diagnoses securely and accurately.

5.3.2 XGBoost

XGBoost is a sophisticated version of the gradient boosting algorithm famous for its efficiency, performance, and scalability. It's an ensemble learning technique that constructs models one by one, where each subsequent model improves on the mistakes of earlier models. This boosting technique allows the model to pay more attention to the hard-to-predict cases, leading to better accuracy.

The strength of XGBoost lies in its ability to manage sparse data efficiently, provide regularization to prevent overfitting, and parallelize tree construction. XGBoost, when training, optimizes a given loss function using gradient descent to refine performance incrementally with each pass. It also internally manages missing data, which improves its resilience in practice.

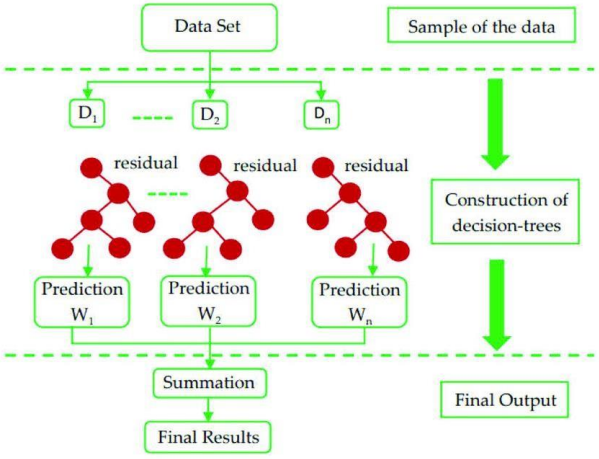

Fig 5.4 XGBoost

XGBoost was used here because it excels in classification tasks, particularly with structured medical data. As implemented with our breast cancer data, it resulted in highly accurate predictions and hence ideal for a clinical setup where dependability matters most. Both performance and explainability qualities that make XGBoost a great selection in predictive modeling across health-related uses.

5.3.3 CKKS Homomorphic Encryption

CKKS (Cheon–Kim–Kim–Song) is a homomorphic encryption scheme tailored to approximate arithmetic of real or complex numbers. Contrary to classical encryption schemes for exact integers or bits, CKKS supports encrypted computation over floating-point data. This renders it extremely well-adapted for privacy-preserving machine learning computations, e.g., secure medical diagnosis, in which numerical operations are executed over encrypted patient information without decryption. CKKS facilitates direct addition and multiplication operations on encrypted data, for which it's specifically suited for secure inference in ML models.

The mathematical basis of CKKS is to encode vectors of real or complex numbers into polynomials and subsequently perform operations in a ring structure. The central formula expressing the heart of CKKS is:

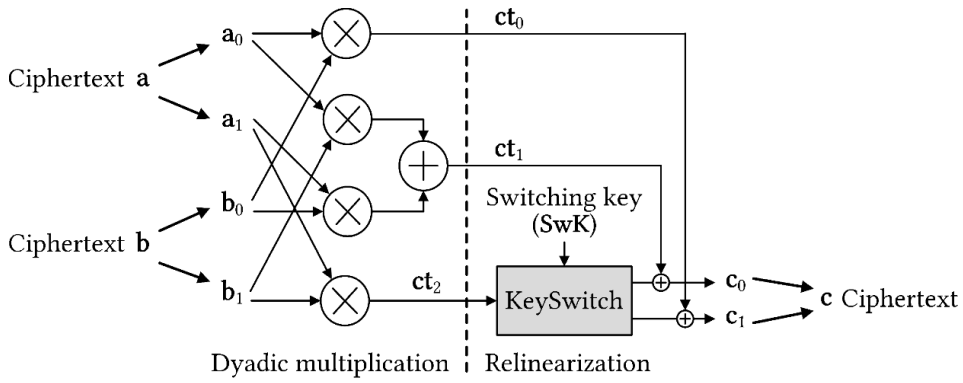

Fig 5.5 CKKS Homomorphic Encryption

The approximation is good enough for most ML and statistical applications that can handle small numerical noise, particularly when dealing with high-dimensional real-valued data.

CKKS also introduces a notion of "scaling factor" to handle precision in the encrypted space. Every ciphertext has an associated scaling factor that grows upon multiplication and is periodically rescaled to limit error growth. In the context of our model, CKKS allows to make predictions on encrypted patient features by using trained ML models without revealing sensitive medical information throughout the computation process. It is especially useful in cooperative healthcare environments, where data privacy is most important.

5.4 Evaluation of Algorithms

This section addresses the performance evaluation measures employed to measure the effectiveness of the proposed machine learning models. Both classification measures are used to judge the quality of the predictions generated by the Random Forest and XGBoost algorithms employed for breast cancer prediction.

5.4.1 Classification Model Evaluation

a. **Accuracy:**

Accuracy quantifies the general validity of the model by computing the ratio of accurate results (true positives and true negatives) over the total cases analyzed. Accuracy provides a basic idea of the frequency of times the classifier correctly classifies cases.

$$\boldsymbol{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}}$$

Eq.5.1

b. **Precision**

precision finds out how many of the anticipated positive cases really were positive. It is particularly useful when it is costly to be a false positive.

$$\boldsymbol{Precision = \frac{FP}{FP + TN}}$$

Eq. 5.2

c. **Recall**

Recall, or sensitivity or true positive rate, assesses how the model performs to find true positive cases among all the actual positive cases. Recall is important in medical diagnosis where false negative finding may have very adverse effects.

$$\boldsymbol{Recall = \frac{TP}{TP + FN}}$$

Eq. 5.3

d. **F1 Score**

The F1 Score is the harmonic mean of precision and recall. It is a balanced measure that considers both false positives and false negatives, providing a more comprehensive view of the model's accuracy.

$$\boldsymbol{F1\_Score = \frac{2 * Precision * Recall}{Precision + Recall}}$$

Eq. 5.4

6 System requirements and specifications

This chapter provides a summary of the hardware and software specifications required during the development and deployment of the project. It also describes the software tools and libraries utilized during the development, providing an extensive overview of technical setup required for the project.

System Requirement Specification (SRS) is an important component of the software development life cycle. It describes the entire scope, functionality, and limits of the system. An SRS is a map that describes the project goals so that it fulfills the user requirements with the least amount of risk and deviation. The SRS is distinctive in that it contains functional and non-functional requirements and acts as a guide throughout the development process.

6.1 Hardware Requirements

i. Processor: Intel Core i3 or above

ii. Hard Disk: Minimum 500 GB

iii. RAM: 4 GB or higher

iv. Peripherals: Standard accessories such as monitor, keyboard, and mouse

6.2 Software Requirements

i. Operating System: Windows 10 or above

ii. Programming Language: Python

iii. IDE/Environment: Anaconda Navigator with Jupyter Notebook and Google colab

iv. Web Framework: Streamlit

v. Encryption Library: TenSEAL

vi. AI Tools Used: ChatGPT, Google Bard

6.2.1 Software Libraries

i. **NumPy**: NumPy is a Python library for doing efficient numerical computations. It offers quick multidimensional array objects as well as capabilities for mathematical operations such as linear algebra and Fourier transform. NumPy arrays outperform Python lists for data processing workloads.

ii. **Pandas**: Pandas is a Python library that allows you to manipulate and analyze data. It includes simple data structures such as DataFrames for working with structured data, making it perfect for cleaning and preparing patient information for machine learning.

iii. **Matplotlib:** Matplotlib is a Python charting toolkit that can produce static, animated, and interactive displays. It integrates seamlessly with NumPy arrays and is often used for displaying model performance metrics and training data.

iv. **Scikit-learn:** Scikit-learn is a comprehensive machine learning framework that offers efficient tools for classification, regression, and clustering. It makes it easier to implement algorithms like Random Forest and XGBoost while also allowing for preprocessing, model evaluation, and hyperparameter customization.

v. **XGBoost:** The classification model was built using XGBoost, a high-performance gradient boosting toolkit. It is designed for high speed and performance, making it excellent for processing big feature sets.

vi. **TenSEAL:** TenSEAL is a package that performs homomorphic encryption on tensors to enable encrypted machine learning algorithms. It supports the CKKS technique, which allows for privacy-preserving predictions on encrypted patient data without revealing sensitive information to the server.

vii. **Streamlit**: Streamlit is an open-source Python framework for developing and deploying interactive web apps for machine learning and data research. In this project, it acts as the front-end interface, allowing users to enter data, read predictions, and interact securely with the encrypted backend model.

7 System design

UML diagrams are employed to illustrate the organization of the systems in terms of classes, attributes, relations, and operations among objects. Execution of the entire program takes place in 4 principal steps. The four principal modules are realized as described below:

7.1 Use case diagram

Use-case diagrams display the functionality and scope of the system as a whole. Use-case diagrams also define the interaction between the actors and the system. The Fig. 7.1 displays the use case representation of the system

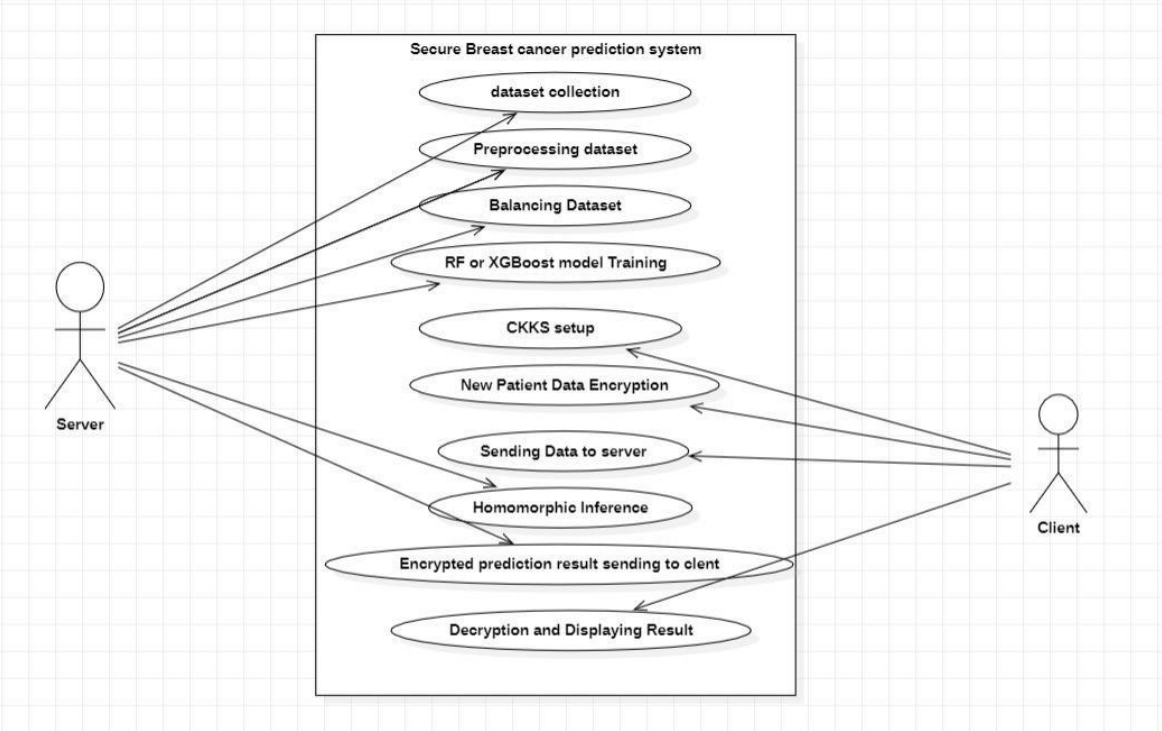

Fig 7.1 Use Case Diagram

7.2 Class diagram

Class diagram is actually a graphical representation of the static structure of the system and is a representation of many aspects of the application. The Fig. 7.2 shows the class diagram representation of the system.

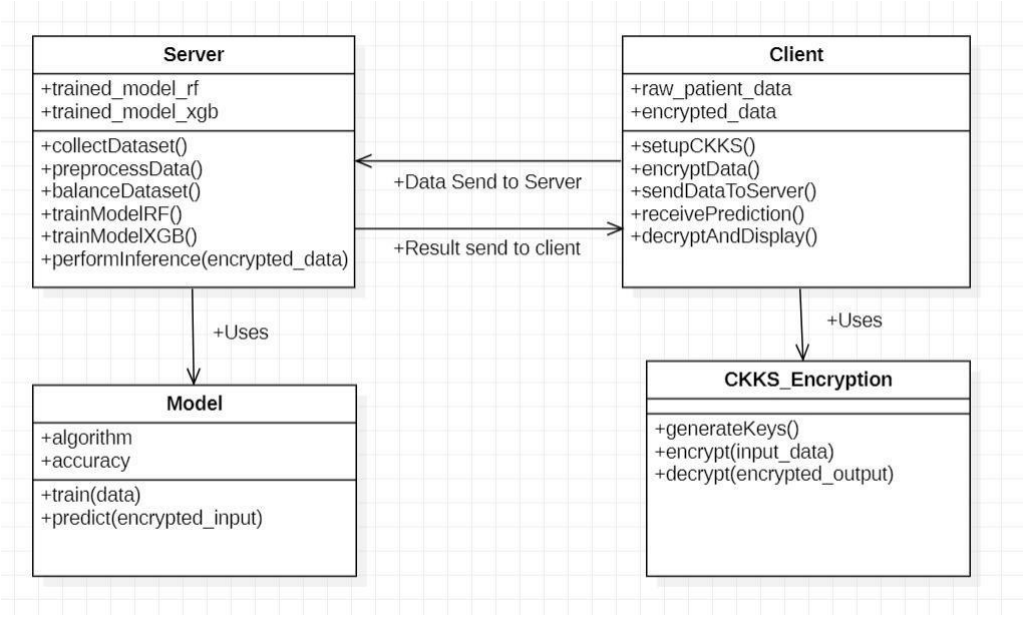

Fig 7.2 Class Diagram

7.3 Activity diagram

Activity diagram is actually a flowchart to indicate the flow between two activities can be defined as a system operation. Control flow is indicated from one operation to another. Such flow may be sequential, branched, or concurrent. In UML, an activity diagram illustrates a snapshot of the behavior of a system by defining the sequence of actions in a process. The Fig. 7.3 illustrates the activity diagram representation of the system

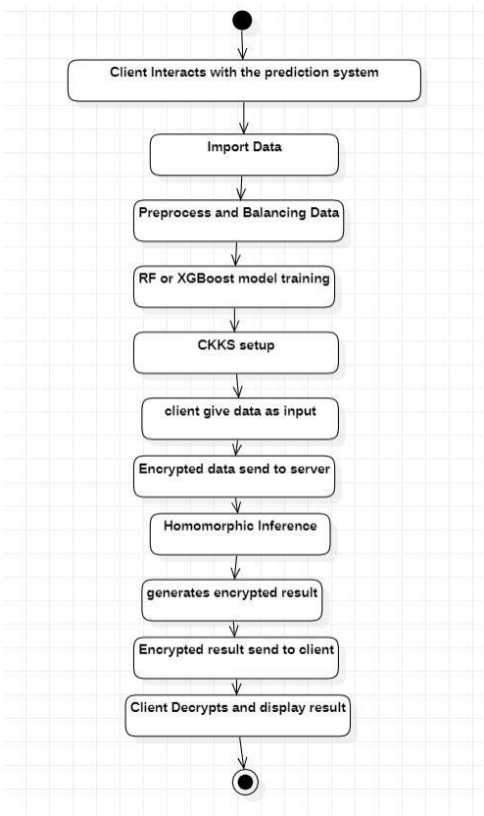

Fig 7.3 Activity Diagram

7.4 Sequence diagram

A sequence diagram is an interaction diagram because it shows how-and in what order-a group of objects work together. The Fig. 7.4 shows the system's sequence diagram notation. Software developers and business people use these types of diagrams to find requirements for a new system or to model an existing process. Like the class diagram, developers like to think sequence diagrams were created just for them.

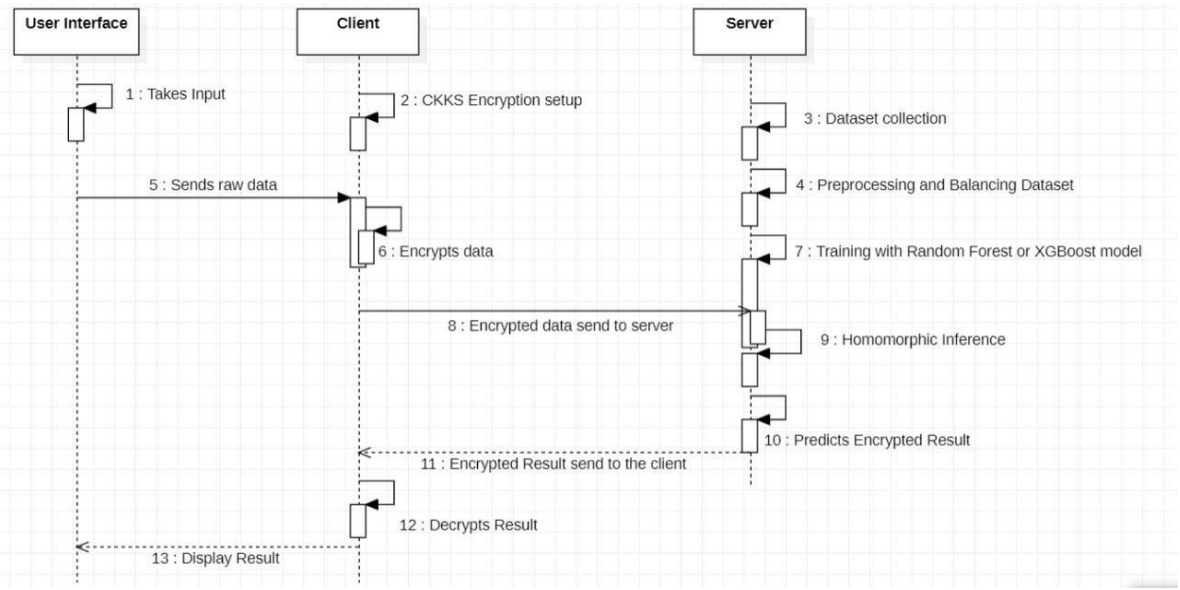

Fig 7.4 Sequence Diagram

8 Testing

Software testing is a significant task for guaranteeing the reliability, functionality, and security of the Secure Breast Cancer Prediction System. It checks whether the system is behaving as expected with various components such as machine learning prediction, homomorphic encryption, and client-server communication. Testing process included different phases such as unit, integration, system, acceptance, and performance testing.

8.1 Unit Testing

Individual module unit testing of data preprocessing, model training, encryption settings, and prediction logic was done. Functionality was tested using valid and invalid inputs to avoid incorrect outputs, especially for the encryption and decryption modules, and the model inference when the inputs are numbers.

8.2 Integration Testing

Integration testing was conducted to validate the interaction among various modules such as client encryption, server inference, and processing responses. Integration testing validated that the encrypted patient data was sent, processed securely on the server, and encrypted results received back and decrypted correctly on the client side.

8.3 System Testing

System testing was used to verify the end-to-end system process from client input all the way through to the final prediction output. It made certain that user interface, model predictions,

secure communication, and result display were in harmony with each other in real-time.

8.4 Acceptance Testing

Acceptance testing entailed end-user verification to ensure that the system fulfilled all functional needs like usability, correct prediction, and secure communication. It was certified and tested on the basis of dependability in delivering encrypted diagnosis results via a simple user interface.

8.5 Performance Testing

Performance testing confirmed the system's response time and computational performance in training models, encryption, and inference. It allowed for the identification of delays introduced by encryption and ensured the system is efficient and secure even for bigger data sets.

9 Implementation

This section outlines step-by-step deployment of our homomorphic encryption-based secure breast cancer classification system using machine learning. It includes dataset preprocessing, model training, CKKS-based encrypted inference, and deployment through an easy-to-use Streamlit web application.

9.1 Loading the Dataset

The data that is used in this project is the Wisconsin Breast Cancer Diagnostic Dataset, which is downloaded via Kaggle and supplemented by another 12 numeric attributes. It is then imported using pandas.read_csv(). The DataFrame is later used for preprocessing.

9.2 Dropping Unnecessary Columns

The "id" column is deleted by drop() as it has no contribution towards the prediction task. The "diagnosis" column is utilized as y, the target variable, and the others are placed into X for training.

9.3 Handling Missing Values

For ensuring model stability, missing values in the data are handled with SimpleImputer with strategy being 'mean'. This replaces the null values with the mean of every column.

9.4 Label Encoding the Target Variable

The 'diagnosis' column contains categorical data ("M" for malignant and "B" for benign). They

are converted to numeric using LabelEncoder, where "M" is assigned 1 and "B" is assigned 0.

9.5 Feature Scaling

Feature scaling is done with the help of StandardScaler from Scikit-learn. This is to ensure that every feature contributes to the learning process equally by bringing all the values to a similar scale.

9.6 Imbalanced Data Handling using SMOTE

The dataset had class imbalance (more benign cases than malignant). This was tackled using SMOTE (Synthetic Minority Oversampling Technique) with a sampling strategy of 0.75. This oversampled the minority class so that a balanced dataset could be achieved.

9.7 Splitting the Dataset

After preprocessing, the data was divided into a training set and a testing set using train_test_split with 20% as the test size. Stratified sampling was applied to ensure class distribution in both sets.

9.8 Model Training

The Random Forest and XGBoost algorithms were individually trained on the preprocessed breast cancer dataset to make predictions as malignant or benign for tumors. The two algorithms were optimized to reach high accuracy and reliability in their predictions as the basis for safe inference within the encrypted setting.

9.8.1 XGBoost Classifier

An XGBoost classifier was trained with 500 estimators, a maximum depth of 8, and a learning rate of 0.03. The model had high predictive accuracy on unseen data.

9.8.2 Random Forest Classifier

100 estimators were used to train a Random Forest classifier. This ensemble model yielded stable and interpretable results, and its overall robustness helped.

9.9 Feature Importance Extraction

Both models were trained after which feature importances were derived from them. The average of the feature importance values was computed and used as a threshold value in CKKS-encrypted predictions. These importances and thresholds were subsequently employed in the encrypted inference pipeline.

9.10 Homomorphic Encryption Using CKKS

For ensuring data privacy, the encryption scheme CKKS of TenSEAL was utilized. The encryption context was parameterized with the polynomial modulus degree of 8192 and coefficient modulus bit sizes as [60, 40, 40, 60] for maintaining a trade-off between efficiency and security. Encryption keys like Galois keys and relinearization keys were created for enabling functionalities like rotation and rescaling on ciphertexts. The trained models, along with preprocessing modules such as the scaler, imputer, and SMOTE, and the CKKS context, were serialized using Python's pickle module to support secure, encrypted inference later on.

9.11 Encrypted Prediction Logic

The prediction was computed based on the encrypted dot product between the input vector encrypted with the input feature importance vector encrypted. The resultant value was then compared with the encrypted threshold to establish whether or not the sample was predicted benign or malignant. All computation occurred within the encrypted space, leaving only the ultimate output to be decrypted.

9.12 Streamlit-Based Web Application

The secure diagnostic model encrypted was implemented as a Streamlit web app in order to provide safe and convenient breast cancer prediction. Users may input 42 numerical features, 30 of which are taken from the original dataset and 12 additional engineered ones. Users have the choice of selecting between the XGBoost or Random Forest model to perform the prediction. All the input features are encrypted via the CKKS scheme prior to inference, thereby preserving data privacy throughout the process. The encrypted result is decrypted to show the diagnosis, which shows whether the case is malignant or benign. To have optimal performance, the pre-trained models and encryption environment are loaded effectively by applying the @st.cache_resource decorator in Streamlit.

9.13 Model Serialization

The trained models, preprocessors, feature importance vectors, threshold values, and CKKS encryption contexts were all stored in.pkl format using the pickle module. That made it easier to load them and directly consume them in the web application.

9.14 Evaluation Metrics

While emphasis was on secure prediction, internal validation revealed that both models performed very accurately. The XGBoost classifier performed marginally better than the Random Forest model in precision and recall. Other metrics like confusion matrix and F1-score were also tested during training.

9.15 GitHub Repository Link

https://github.com/MowlyaSai/Secure_machine_learning_for_pathological_assessment.git

10 Results

In this chapter, we present the outcome of our encrypted diagnostic program for breast cancer. The Streamlit interface is shown in Figure 10.1. Depending on the input, the encrypted prediction result is either malignant (Figure 10.2) or benign (Figure 10.3). Random Forest and XGBoost model performance metrics are indicated in Figures 10.4 and 10.5, and their corresponding evaluation plots are in Figures 10.6 and 10.7. From the comparison in Table 10.1, XGBoost is slightly better than Random Forest in overall classification performance under encryption.

10.1 Interface

The interface is created with the help of the Streamlit Python framework. When run, the app is opened in the browser and facilitates the user to enter 42 numeric features (30 of the original Wisconsin dataset and 12 custom features).

It is created in a manner where the user can choose between the Random Forest or XGBoost model to carry out an encrypted diagnosis. When input is made, data is encrypted through the use of the CKKS scheme, the prediction is carried out on the encrypted values, and the decrypted result is presented to the user.

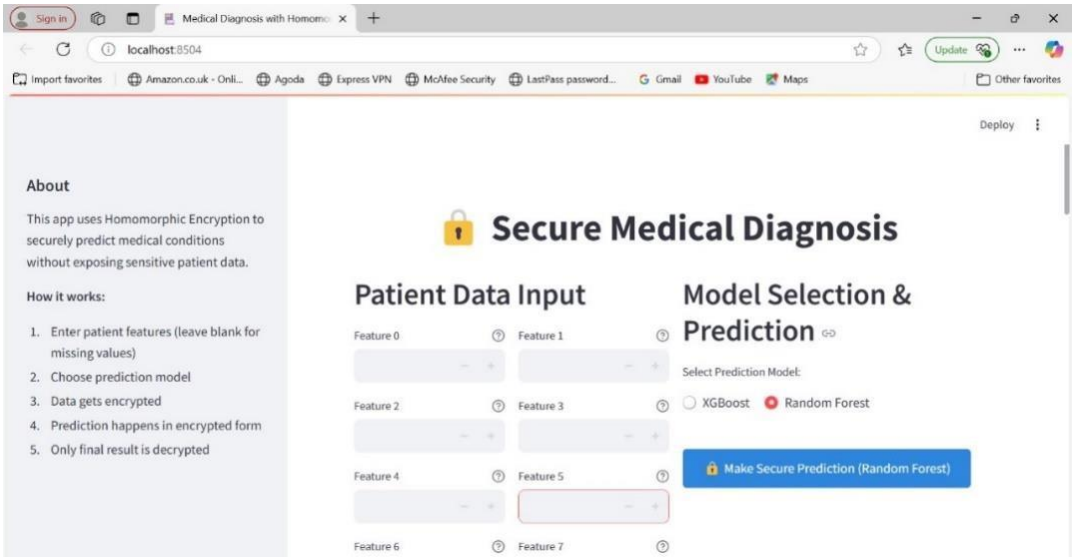

Figure 10.1 User Interface

10.2 Output when Malignant Data is Input

When the user provides input data that is for a malignant case and chooses a prediction model, the encrypted input is processed and the decrypted result is shown as malignant on the screen.

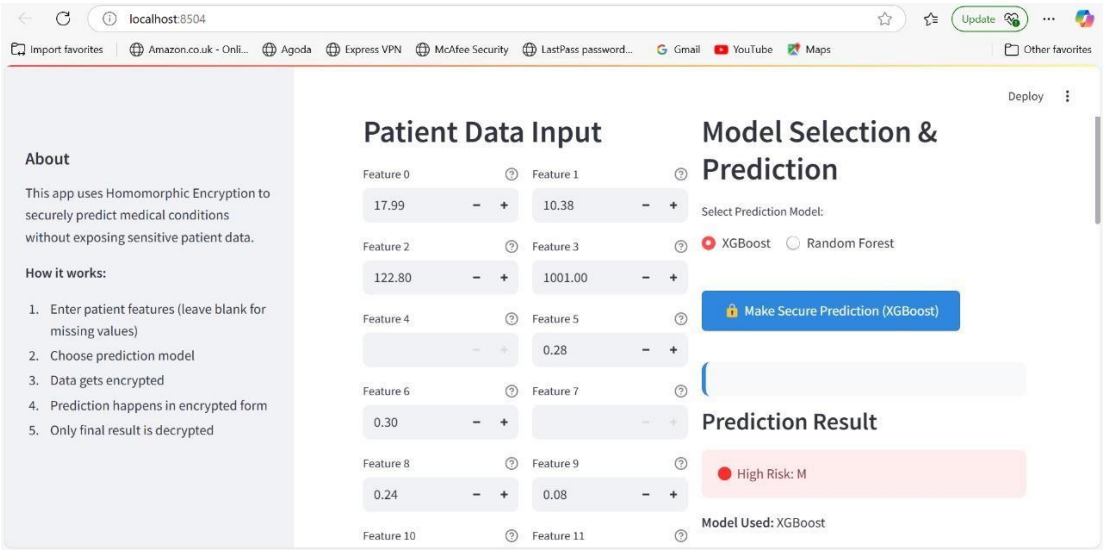

Figure 10.2 Result when Malignant Data is Input

10.3 Output when Benign Data is Input

When the user provides as input data that corresponds to a benign case with either of the models, the encrypted input is computed and the decrypted output is reflected as benign in the result window.

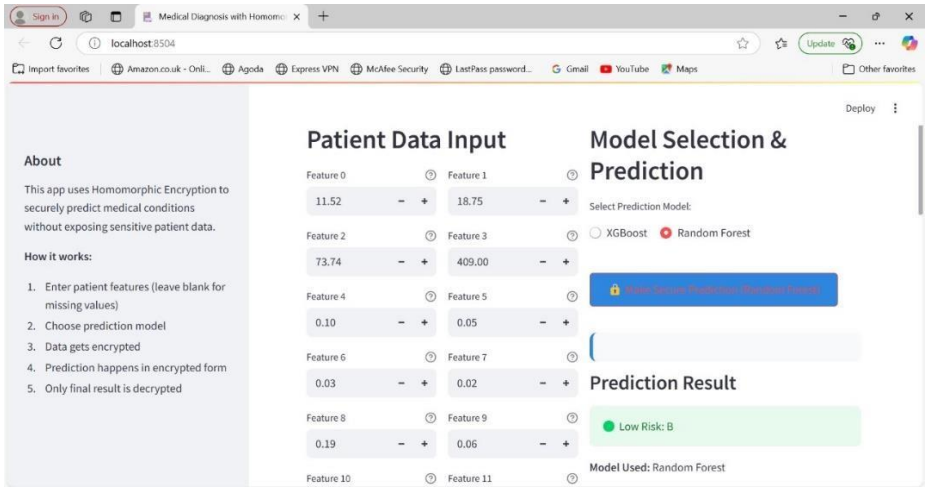

Figure 10.3 Result when Benign Data is Input

10.4 Performance Metrics of Random Forest Model

Figure 10.4 presents the performance of the Random Forest model on the test set. The values are accuracy, precision, recall, and F1-score.

```
Random Forest Model Evaluation:
Accuracy: 0.9760
Precision: 1.0000
Recall: 0.9434
F1-Score: 0.9709
```

Figure 10.4 Performance of Random Forest Model

10.5 XGBoost Model Performance Metrics

Figure 10.5 shows the classification report of the XGBoost model, including common performance metrics employed for the evaluation.

```
XGBoost Model Evaluation:
Accuracy: 0.9920
Precision: 1.0000
Recall: 0.9811
F1-Score: 0.9905
```

Figure 10.5 XGBoost Model Performance

10.6 Performance Plots

Bar plots in Figures 10.6 and 10.7 graphically show the evaluation metrics (accuracy, precision, recall, F1-score) for Random Forest and XGBoost models, respectively.

Figure 10.6 Random Forest & XGBoost Performance Graph

10.7 Comparison

Table 10.1 compares the performance of Random Forest and XGBoost models based on several performance metrics to select the high-performing algorithm for encrypted breast cancer diagnosis.

Table 10.1 Comparison of Random Forest and XGBoost Models

Serial No	Algorithm	Accuracy	Precision	Recall	F1-Score
1	Random Forest	97.60%	100%	94.34%	97.07%
2	XGBoost	99.20%	100%	98.11%	99.05%

11 Conclusions

In this project, we implemented a secure and efficient machine learning-based diagnostic system for breast cancer classification with homomorphic encryption. We aimed at protecting the privacy of sensitive medical information while achieving high prediction accuracy. The Wisconsin Breast Cancer Diagnostic dataset, which was augmented with 12 more numerical features, was utilized to train the models.The data went through preprocessing phases like mean imputation for missing values, label encoding of the target variable, feature standardization, and class balancing via SMOTE. These phases helped to ensure that the models got clean and balanced data, which is important in ensuring consistent performance.

We trained two classification models—Random Forest and XGBoost—on the preprocessed data. Both performed extremely well, with XGBoost performing 99.2% and Random Forest performing at 97.6%. All critical evaluation measures such as precision, recall, and F1-score also proved the effectiveness of the models to separate benign and malignant cases.

In order to maintain data privacy, we employed the CKKS homomorphic encryption scheme via the TenSEAL library, which supports encrypted inference without revealing raw input data. The models, along with preprocessing modules and encryption context, were serialized for usage. The whole system was integrated into an interactive Streamlit web app, which supports users in inputting 42 numerical features and receiving encrypted diagnostic results in a convenient way.
In general, this project is able to prove that high-performing machine learning models can indeed be integrated with privacy-preserving encryption methods, providing a viable solution for secure medical diagnosis.

12 Future Enhancement

There are a few auspicious avenues in which this project can be further developed to enhance predictive accuracy and security. One significant addition would be incorporating deep learning models like Convolutional Neural Networks(CNNs) or Deep Neural Networks (DNNs), which have proven extremely accurate in classification tasks in the medical domain. These models can learn intricate patterns in high-dimensional data autonomously. Based on data, including more clinical datasets containing imaging, genomic, or pathology records may greatly enhance the input feature space and enable more comprehensive diagnosis. Multimodal input can allow the model to better identify subtle malignancy indicators and thereby achieve higher diagnostic accuracy.

From a security and privacy standpoint, testing out alternative privacy-protecting methods like Federated Learning or Secure Multi-Party Computation (SMPC) can further protect patient data in decentralized settings. They can allow collaborative model training between institutions without having to share raw data, enhancing confidentiality.

In addition, improving the scalability and performance of the encryption module by optimizing CKKS parameters or switching to GPU-accelerated homomorphic encryption libraries can result in accelerated encrypted inference. Lastly, integrating the whole system into a cloud-based or mobile platform can make the diagnostic tool more convenient for healthcare professionals and patients, enabling real-world deployment in remote or resource-constrained regions.Such future upgrades would not only make the system more accurate and user-friendly but also establish a sound ground for building robust, secure, and scalable medical AI diagnostic solutions in the future.

13 References

[1] Al Badawi, A., & Faizal Bin Yusof, M. (2024). “Private pathological assessment via machine learning and homomorphic encryption.” *BioData Mining*, 17(33).

[2] Cheon, J. H., Kim, A., Kim, M., & Song, Y. (2017). “Homomorphic encryption for arithmetic of approximate numbers”. *Advances in Cryptology*–ASIACRYPT 2017, pp. 409–437. Springer.

[3] Wolberg, W., & Mangasarian, O. (1990). “Wisconsin Breast Cancer Dataset. UCI Machine Learning Repository”. https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic

[4] Wolberg, W., Mangasarian, O., Street, N., Street, W. (1995). “Breast Cancer Wisconsin (Diagnostic)”. https://doi.org/10.24432/C5DW2B

[5] Zhu, H., Liu, X., Lu, R., Li, H. (2016). “Efficient and privacy-preserving online medical prediagnosis framework using nonlinear SVM.” *IEEE J Biomed Health Inform*, 21(3), 838–850.

[6] Zhang, M., Song, W., Zhang, J. (2020). “A secure clinical diagnosis with privacy-preserving multiclass support vector machine in clouds”. *IEEE Syst J*, 16(1), 67–78.

[7] Chen, B., & Zheng, X. (2022). “Implementing Linear Regression with Homomorphic Encryption.” *Procedia Comput Sci*, 202, 324–329. https://doi.org/10.1016/j.procs.2022.04.044

[8] Gursoy, G., Chielle, E., Brannon, C. M., Maniatakos, M., Gerstein, M. (2022). “Privacy-preserving genotype imputation with fully homomorphic encryption.” *Cell Syst*, 13(2), 173–182.

[9] Blatt, M., Gusev, A., Polyakov, Y., Goldwasser, S. (2020). “Secure large-scale genome-wide association studies using homomorphic encryption.” *Proc Natl Acad Sci*, 117(21), 11608–11613.

[10] Lee, J. W., Kang, H., Lee, Y., et al. (2022). “Privacy-preserving machine learning with fully homomorphic encryption for deep neural network”. *IEEE Access*, 10, 30039–30054.