Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Breast Cancer Classification from Transcriptomic Data: A Hybrid Machine Learning and Blockchain for Data Reliability and Integrity

Dr. Berdjouh Chafik [1]
[1] *LIAP Laboratory, University of Eloued, Algeria*

ARTICLE INFO

Received: 12 Jan 2025
Revised: 18 Jul 2025
Accepted: 27 Aug 2025

ABSTRACT

The high death rate from breast cancer continues to impact women globally throughout all regions of the world. Accurate breast cancer classification through gene expression analysis is a fundamental step in creating individualized cancer treatment approaches. Traditional machine learning models, including Logistic Regression, together with Random Forests, Support Vector Machines, and advanced algorithms such as XGBoost and Multilayer Perceptrons, have proven their effectiveness for predictive tasks. The models demonstrate high sensitivity to both random and purposeful data modifications, which leads to less dependable diagnostic outcomes. The proposed method combines machine learning with blockchain technology to create a framework. The validation framework utilizes SHA-256 hashing, combined with smart contracts and distributed ledger technology, to verify data integrity prior to classification. We examine the CuMiDa breast cancer gene expression dataset, along with machine learning models that utilize both traditional and blockchain-based approaches. The baseline models achieved strong performance with accuracy values between 84% and 95%, but the blockchain-assisted models demonstrated superior trustworthiness. The implemented system decreased its exposure to noise while preserving both accuracy levels and F1 scores. The research demonstrates how blockchain technology enhances machine learning applications. The combination of blockchain with machine learning enables both high predictive performance and complete data integrity and traceability, which creates a stronger biomedical application framework.

Keywords: Breast cancer, Machine learning, Gene expression, Blockchain, Data integrity, CuMiDa dataset.

1.Introduction

Breast cancer stands as a leading cause of female death and disease across the globe because millions of women receive their first diagnosis every year [1]. High-throughput sequencing technology now produces massive gene expression databases, which CuMiDa collects to drive precision oncology studies. Support Vector Machines (SVM), Random Forests (RF), Logistic Regression (LogReg), and deep learning models have proven successful for breast cancer classification by reaching 90% or higher accuracy rates according to [2] and [3]. The improved performance of machine learning systems has not solved the ongoing problem of ML pipelines remaining susceptible to noisy, manipulated, and adversarial data. The minor changes in input features produce significant differences in prediction results, which can damage the reliability of diagnostic support systems.

Blockchain technology functions as a new system that protects healthcare system data integrity through transparent methods of tracking and verification [4, 5]. Blockchain technology protects medical information through its distributed ledger system and cryptographic hashing and smart contract technology, which produces an unchangeable and authentic data record for the entire data lifecycle. The current research into blockchain-based secure health record management shows limited progress because it fails to combine blockchain validation systems with machine learning cancer detection methods, which analyse transcriptomic data. This represents a critical gap: although classification performance is essential, ensuring that training and testing data are authentic and free from malicious alterations is equally crucial for clinical deployment.

The aim of this study is therefore twofold: (1) to develop a blockchain-enhanced machine learning framework for breast cancer classification using gene expression data from CuMiDa, and (2) to evaluate the robustness of

traditional ML models (LogReg, RF, SVM, XGBoost, and MLP) under both baseline and simulated injection scenarios. By comparing performance across these settings, we demonstrate how blockchain validation improves trust in ML outputs without significantly compromising accuracy.

The remainder of this paper is organized as follows: Section 2 reviews prior works on ML-based breast cancer classification and blockchain in healthcare. Section 3 details the proposed methodology, including dataset description, preprocessing (Z-score, PCA, SMOTE), ML models, and the blockchain validation layer. Section 4 presents the experimental results. The research findings, together with their associated limitations, receive attention in Section 5, which also presents a comparison of related studies. Section 6 provides a summary of the core achievements together with potential areas for upcoming investigations.

2. Related Work

In recent years, breast cancer classification based on gene expression data has garnered growing attention, leading to numerous approaches that rely on Machine Learning (ML) and, more recently, on the integration of Blockchain.

Ferroni et al. [6] Applied classical classifiers, including Logistic Regression (LR), Random Forests (RF), and Support Vector Machines (SVM), for breast cancer prognosis prediction. Their study reported an accuracy of around 92%, demonstrating the robustness of ML models, while also highlighting their high sensitivity to sampling parameters, which limits the generalizability of the results.

Mostavi et al. [7] Proposed a more ambitious approach with Convolutional Neural Networks (CNNs) applied to transcriptomic data. The system achieved excellent results, with approximately 98% accuracy, but it also created significant issues with overfitting and required large datasets, which are often scarce in biomedical fields.

The research by Oyediran et al. [8] tested multiple classification methods to detect breast cancer, with their study focusing on KNN and LR techniques. The evaluation revealed that specific methods achieved higher specificity; however, the performance metrics highlighted substantial differences between accuracy and recall, indicating that model adaptation is necessary to address class imbalances in medical datasets.

Alongside purely ML-focused works, Fang et al. [4] explored Blockchain for securing Personal Health Records (PHR). Their systematic review demonstrated Blockchain's potential to ensure data integrity and traceability, but did not establish a direct link with predictive models. Similarly, Ramanath et al. [9] proposed a blockchain-based multi-agent system for breast cancer diagnosis. However, this work remains conceptual and has not been validated on standard datasets, such as CuMiDa.

More recently, Al-Khasawneh et al. [5] introduced a secure Blockchain framework for medical records management. Their architecture demonstrated strong resistance to data tampering; however, no integration with predictive models was included.

In 2025, La Moglia and Mohamad Almustafa[3] assessed multiple ML classifiers, which included SVM, RF, and LR for breast cancer prediction. Their research study in Intelligence-Based Medicine demonstrated prediction accuracy that reached from 90% to 95%. The authors failed to resolve the problem of data integrity, which remains essential for distributed clinical settings.

Also in 2025, Kallah-Dagadu et al. [10] proposed an interpretable approach using SVM and RF for RNA-seq data analysis. By integrating explanatory tools such as SHAP, they achieved an accuracy above 92%, offering better insight into discriminant variables. Nevertheless, this approach remains highly dependent on the quality of the initial data.

From another perspective, Hussain et al. [11] compared several deep models (CNNs, autoencoders, multimodal architectures) for breast cancer classification using multimodal datasets. Their findings showed that deep learning architectures often reached accuracies above 97%, but at the expense of higher computational complexity.

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Finally, Omran et al. [12] investigated the integration of multi-omics data, comparing statistical methods (MOFA+) with deep learning approaches (MoGCN). Their results showed that MoGCN achieved superior predictive capability for tumor subtypes, but required considerable resources and remained difficult to interpret.

Overall, these studies demonstrate a steady progression in classification performance. Still, they also reveal two significant gaps: on the one hand, the vulnerability of ML models to noisy or tampered data, and on the other hand, the absence of a robust validation and security mechanism for medical inputs. It is precisely within this context that our research is positioned, by proposing a hybrid framework combining ML with Blockchain to ensure both predictive performance and data integrity.

A synthesis of these studies is provided in Table 1, which highlights methodologies, datasets, results, and limitations.

Table 1 : Comparative analysis of existing works

Ref.	Methodology	Dataset / Domain	Results (Accuracy, etc.)	Limitations
[6]	ML classifiers (LR, RF, SVM)	Breast cancer prognosis datasets	≈92% accuracy	Sensitive to sampling parameters; limited generalizability
[7]	CNN on transcriptomic data	RNA-seq (cancer samples)	≈98% accuracy	Requires large datasets; prone to overfitting
[8]	Comparative ML (KNN, LR)	Breast cancer detection	Variable; some high specificity	Significant gaps between precision and recall; class imbalance issues
[4]	Blockchain for PHR (systematic review)	Personal Health Records	Improved security & traceability	No link with predictive ML models
[9]	Blockchain-based multi-agent system	Breast cancer diagnosis (conceptual)	Proof-of-concept only	Not validated on standard datasets (e.g., CuMiDa)
[5]	Secure Blockchain framework for health records	Healthcare systems	Strong resistance to tampering	No integration with predictive ML
[3]	ML classifiers (SVM, RF, LR)	Breast cancer datasets	90–95% accuracy	No attention to data integrity in clinical contexts
[10]	Interpretable ML (SVM, RF + SHAP)	RNA-seq data	>92% accuracy	Dependent on initial data quality
[11]	Deep models (CNN, autoencoders, multimodal)	Multimodal biomedical datasets	>97% accuracy	High computational complexity
[12]	Multi-omics (MOFA+, MoGCN)	Multi-omics datasets	MoGCN superior predictive performance	Computationally expensive; poor interpretability

3. Methodology

3.1 Dataset

The dataset used in this study originates from the CuMiDa (Curated Microarray Database), which is available on Kaggle under the title "Breast Cancer Gene Expression (CuMiDa)" [13]. It represents a transcriptomic subset dedicated to breast cancer, structured from microarray data.

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

This dataset comprises 151 biological samples, distributed across six breast cancer subtypes (basal, HER2, Luminal A, Luminal B, cell line, and normal), each characterized by a gene expression profile of 54,676 genes. The resulting matrix is therefore highly dimensional, posing both a challenge for dimensionality reduction and an opportunity to uncover discriminative genomic signatures.

The class distribution is slightly imbalanced (Fig. 1), which motivated the integration of the SMOTE technique into our pipeline to correct this bias. The use of this dataset is particularly relevant for testing the effectiveness and robustness of our hybrid methodology, which combines machine learning and Blockchain in a real biomedical context.

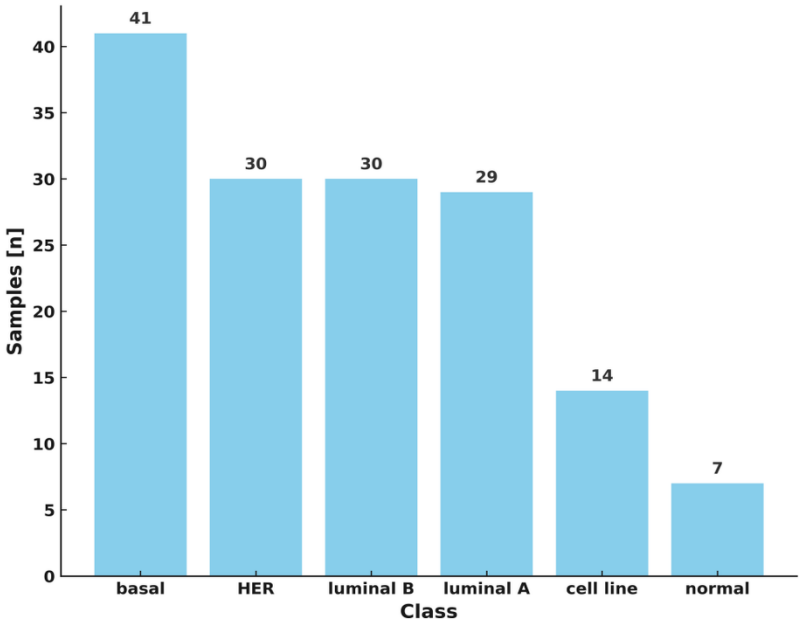


Figure 1: Distribution of Samples by class in the CuMiDa – Breast Cancer Dataset (151 samples, 6 classes)

3.2 Data Preprocessing

Before training the models, a rigorous preprocessing procedure was applied to the CuMiDa dataset to ensure the quality and reliability of the analyses. The main steps are summarized below:

1) Z-score normalization is applied to centre and scale the transcriptomic data to eliminate the effect of scale differences between genes and to facilitate the convergence of learning algorithms. This transformation is essential in high-dimensional contexts, where variance across genes may bias the models [14].

2) Dimensionality reduction through PCA (Principal Component Analysis) projects the original data into a lower-dimensional space by maximizing explained variance, thereby reducing noise and improving computational efficiency. In the case of gene expression data, PCA contributes to identifying discriminative components for classification [15].

3) The SMOTE (Synthetic Minority Over-sampling Technique) method creates artificial minority samples to achieve class balance in datasets. The technique functions as a solution to address imbalanced biomedical datasets that contain insufficient tumor subtype data, thus minimizing prediction bias according to [16].

3.3 Machine Learning Models

Five classical and advanced Machine Learning models were evaluated:

1) The logistic regression model : it functions as a linear probabilistic system to predict the likelihood of class membership. The method stands as a benchmark for clinical and transcriptomic data analysis because of its dual benefits of operational efficiency and result clarity [17].

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

2) The Support Vector Machine (SVM): it works to create the broadest possible margin between different classes. It performs exceptionally well in high-dimensional data analysis and shows particular strength when dealing with RNA-seq data because it can handle situations where variables outnumber samples [18].

3) Random Forest (RF) : it generates its predictions by combining multiple decision trees, which creates a system that resists overfitting. The method proves effective according to previous studies, which used genomic data for breast cancer classification [19].

4) A Multi-Layer Perceptron (MLP) : it learns complex data patterns through its structure, which contains multiple hidden neuron layers. The method performs well when analysing biological data patterns that do not follow linear relationships while maintaining compatibility with Deep Learning techniques [20].

5) XGBoost (Extreme Gradient Boosting) : it operates through its boosting-based ensemble method, which builds trees one after another to correct previous prediction mistakes. The system shows excellent outcomes when it processes structured data in biomedical applications, according to [21].

3.4 Overall Methodological Framework

Figure 2 illustrates the proposed hybrid methodology, which combines the analysis of medical data (from the Breast Cancer CuMiDa dataset) with a security framework based on Blockchain. The process begins with data collection and preprocessing (normalization, dimensionality reduction, process imbalance data). The processed data are then used to train various machine learning models (Logistic Regression, Random Forest, SVM, MLP, XGBoost), enabling the classification of samples into breast cancer subtypes.

In parallel, Blockchain operates as a layer of reliability and traceability, ensuring data integrity and transparency of transactions related to the results. The trained models are thus integrated into a secure environment where predictions and performance outcomes are stored and validated through smart contracts.

Finally, the results (performance metrics and confusion matrices) are compared between the Baseline scenario (without Blockchain) and the Blockchain scenario, to highlight the benefits of the proposed framework in terms of accuracy, robustness, and reliability.

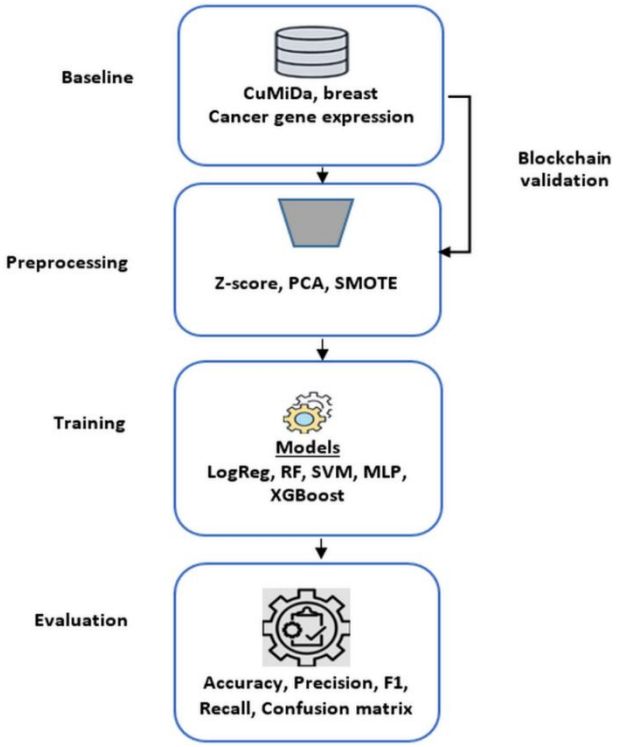


Figure 2: Global Methodology Pipeline

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

3.5 Blockchain Layer

The Blockchain layer integrated into our methodology is designed as a permissioned system, ensuring the validation and traceability of data throughout the pipeline. It is composed of several complementary elements (Figure 3):

1) Smart Contract: acts as an initial validation gateway. It generates a SHA-256 hash of the data and compares it to the reference value recorded in the distributed ledger. Only compliant data are allowed to feed into the ML pipeline.
2) Hashing (SHA-256): each batch of data or result is converted into a unique digital fingerprint. Any alteration, even a minor one, produces a different hash, thereby enabling the rapid detection of manipulations.
3) Blockchain Ledger (Distributed Register): stores all validated hashes in an immutable and timestamped structure. This distributed register guarantees the traceability and integrity of transactions.
4) Validation Gate (Pass/Fail): The system verifies newly generated hashes against blockchain-stored hashes. New data receives validation when its hashes align with those stored on the blockchain. The system rejects data when its hashes do not match those stored in the blockchain.
5) Consensus Mechanism (Permissioned): A straightforward consensus process exists between authorized nodes to validate transactions jointly. The distributed ledger maintains consistency and synchronization through this validation process. Together, these components transform the Blockchain into a trust filter, ensuring that only authentic and verified data are fed into the Machine Learning models.

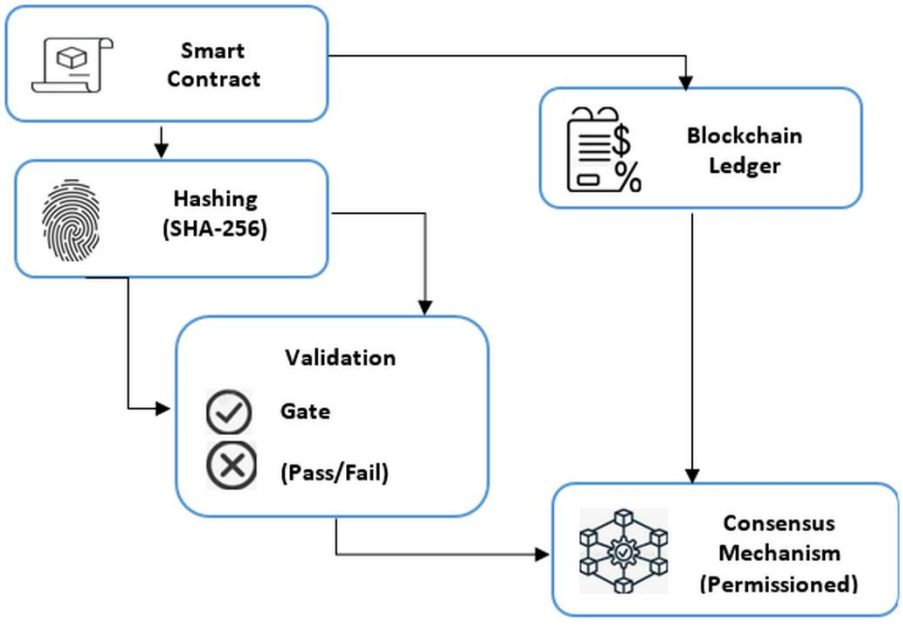


Figure 3 : Architecture of Blockchain

3.6 Injection Scenario

As part of our experimentation, we simulated a scenario involving injected altered data to evaluate the robustness of classification models against intentional or accidental perturbations. Specifically, samples from the CuMiDa dataset were modified by either adding artificial noise or changing their class labels.

- In the Baseline pipeline, these data are processed directly, which exposes the system to an increased risk of silent falsification.

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

- In contrast, the Blockchain pipeline introduces a validation layer based on SHA-256 hashing, a smart contract, and a distributed ledger, enabling the detection of any inconsistencies. Invalid samples are redirected to an off-chain quarantine register, while valid data proceed through the preprocessing steps (Z-score, PCA, SMOTE).

The overall process is illustrated in Fig. 4, which compares the Baseline pipeline with the Blockchain-enhanced pipeline. The detailed procedural steps corresponding to these two scenarios are provided in Section 3.7 (Experimental Algorithms).

3.7 Experimental Algorithms

To formalize the scenario depicted in Fig. 4 at the procedural level, we introduce two complementary algorithms:

- Algorithm 1: Baseline Pipeline – without integrity control, where both original and altered data are directly used for training and testing.
- Algorithm 2: Blockchain Pipeline – incorporating a cryptographic verification step using SHA-256, a smart contract, and a distributed ledger. Invalid data are isolated in an off-chain quarantine register.

Algorithm 1 — Baseline pipeline (without Blockchain)

```
Input : D (dataset), split=75/25
Models : {LogReg, RF, SVM, MLP, XGBoost}
(X_train, y_train, X_test, y_test) = split(D, train=0.75, stratified=True)
for each M in {LogReg, RF, SVM, MLP, XGBoost}
    M.fit(X_train, y_train)
    ŷ = M.predict(X_test)
    report_metrics(ŷ, y_test)
    plot_confusion_matrix(ŷ, y_test)
end for
```

Algorithm 2 — Blockchain-validated pipeline (with data injection)

```
Infrastructure required:
    - SHA-256 hashing - Distributed Ledger (on-chain)
    - Smart Contract for verification - Quarantine Register (off-chain)
Input: D' (preprocessed dataset), A (altered data); split=75/25
Models: {LogReg, RF, SVM, MLP, XGBoost}
D_inj = D' ∪ A
V = ∅
for each sample x in D_inj do
    h = SHA256(x)
    if SmartContract.verify(h, DistributedLedger) == TRUE then
        V = V ∪ {x}                    // ACCEPT
```

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

```
7:        else
8 :               QuarantineRegister.store(x)      // REJECT
9:        end if
10 : end for
11: (X_train, y_train, X_test, y_test) = split(V, train=0.75, stratified=True)
12: for each M in {LogReg, RF, SVM, MLP, XGBoost} do
13:       M.fit(X_train, y_train)
14:       ŷ = M.predict(X_test)
15:       report_metrics(ŷ, y_test)
16:       plot_confusion_matrix(ŷ, y_test)
17: end for
```

These two algorithms procedurally express the differences between the Baseline and Blockchain scenarios (Figure 4). The first highlights the vulnerability of the models when no verification mechanism is integrated, while the second illustrates how the Blockchain layer enables the rejection of falsified samples and strengthens the reliability of input data.

Thus, the experimental evaluation is not limited to measuring the performance of the classification models; it also allows for a comparison of the robustness of the two pipelines in a realistic context of altered data injection. The results obtained for each of the considered models (LogReg, RF, SVM, MLP, and XGBoost) are presented and analysed in Section 4 (Experimental Results).

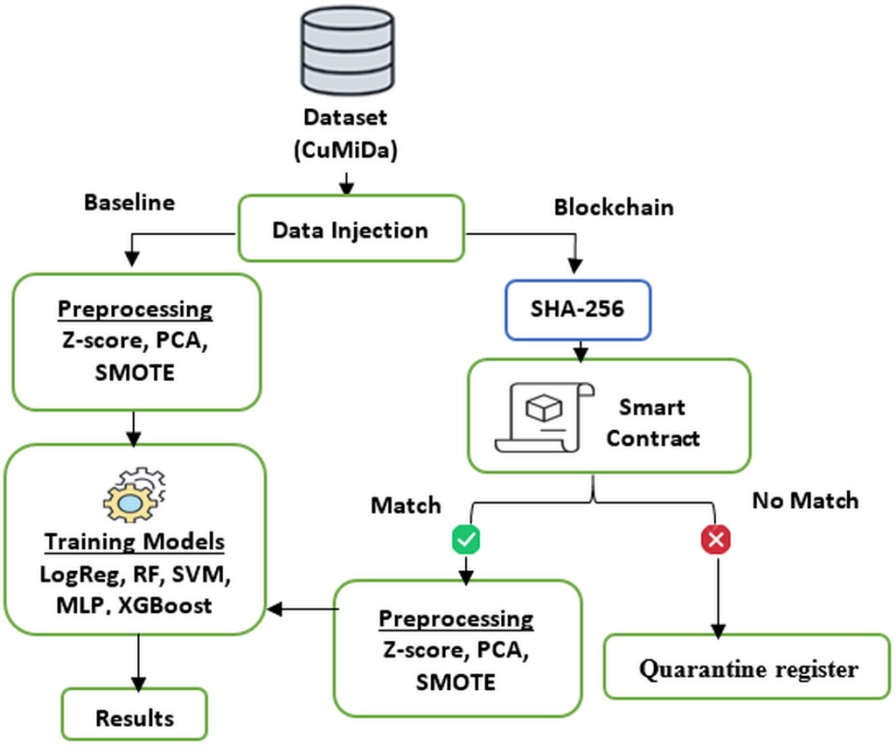


Figure 4: Injection Scenario Diagram (Baseline vs. Blockchain)

3.8 Performance Evaluation

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

To quantify the performance of the proposed framework, a list of reliable performance measures were employed. Each measure provides a different view of the model's predictability, especially when dealing with imbalanced biomedical data.

1) **Accuracy**: computes the proportion of correct predictions for all classes. Although simple, it might be misleading in the event of an imbalanced dataset (see Equation (1)).

$$Accuracy = \frac{\text{T\_Pos} + \text{T\_Neg}}{\text{T\_Pos} + \text{T\_Neg} + \text{F\_Pos} + \text{F\_Neg}} \quad (1)$$

Where T_Pos refers to true positives, T_Neg : true negatives, F_Pos : false positives, and F_Neg : false negatives.

2) **Precision** : it reflects the reliability of positive predictions, indicating the proportion of predicted cancer cases that are truly cancer. High precision minimizes false alarms, which is critical in clinical decision-making (see Equation (2)).

$$Precision = \frac{\text{T\_Pos}}{\text{T\_Pos} + \text{F\_Pos}} \quad (2)$$

3) **Recall** (Sensitivity): it measures the proportion of actual positive cases correctly identified. It measures the model's ability to catch all true cancer cases and not miss patients (see Equation (3)).

$$Recall = \frac{\text{T\_Pos}}{\text{T\_Pos} + \text{F\_Neg}} \quad (3)$$

4) **F1**-Score: is the harmonic mean of Precision and Recall, which keeps false positive vs. false negative in balance. F1-score is particularly telling in medical applications where both types of errors are significant (see Equation(4)).

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

Finally, the Confusion Matrix provides an even more nuanced breakdown of predictions, which show accurate classifications and errors made by class. It plots the counts of T_Pos, T_Neg, F_Pos, and F_Neg and provides a graphical snapshot of model performance.

3.9 Experimental Setup

All experiments were carried out on a workstation with an Intel Core i7-8850H CPU (2.6 GHz, 8 cores), 16 GB of RAM, and an NVIDIA GeForce RTX 3080 GPU (10 GB memory). The software environment used Python 3.12, scikit-learn 1.5, XGBoost 2.0, and TensorFlow/Keras 2.15 for the machine learning tasks. The blockchain validation layer was simulated in a local Python environment with SHA-256 hashing and smart contract emulation.

4. Results

4.1 General Presentation

The experiments conducted on the CuMiDa – Breast Cancer Gene Expression dataset enabled the evaluation of five classification models (LogReg, SVM, Random Forest, XGBoost, and MLP) in two distinct scenarios: Baseline (AI pipeline only) and Blockchain (AI combined with an integrity validation mechanism). Performances were measured using standard indicators: Accuracy, Precision, Recall, and F1-score. The obtained values are summarized in Tables 2 and 3, while the differences between the scenarios are highlighted in Table 4.

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

4.2 Quantitative Results

In the baseline scenario (refer to Table 2), LogReg and SVM show considerable effectiveness, with precision and recall rates above 92%, which proves how reliable they are. RF yields more modest outcomes, particularly in terms of recall (≈83%). The performances of XGBoost and MLP are more variable: XGBoost records relatively low results (F1 ≈0.83), whereas MLP achieves high initial scores (Accuracy ≈94%).

Table 2: Quantitative Results (Baseline)

Scenario	Baseline					
	Accuracy	F1	PR_AUC	Precision	ROC_AUC	Recall
Model						
LogReg	0.9211	0.8923	0.9883	0.9392	0.9965	0.8762
MLP	0.9474	0.9588	1.0000	0.9630	1.0000	0.9595
RandomForest	0.8684	0.8490	0.9615	0.8985	0.9903	0.8304
SVM	0.9211	0.8923	0.9947	0.9392	0.9985	0.8762
XGBoost	0.8421	0.8293	0.8932	0.8829	0.9723	0.8095

In the Blockchain scenario (see Table 3), several noteworthy changes are observed. LogReg and SVM remain generally stable, with performances nearly identical to those in the Baseline setting. Random Forest demonstrates a significant improvement (Δ Accuracy +0.05; Δ F1 +0.04), highlighting the positive effect of the integrity verification mechanism. In contrast, the MLP model shows a degradation, with increased misclassifications and a decline in recall. Finally, XGBoost continues to yield limited performance, with no substantial gain despite the integration of Blockchain.

Table 3. Quantitative Results under the Blockchain Scenario

Scenario	Blockchain					
	Accuracy	F1	PR_AUC	Precision	ROC_AUC	Recall
Model						
LogReg	0.9474	0.9588	0.9841	0.9630	0.9950	0.9595
MLP	0.9211	0.8759	0.9669	0.8773	0.9962	0.8792
RandomForest	0.8947	0.8689	0.9722	0.9394	0.9895	0.8542
SVM	0.9211	0.8923	0.9970	0.9392	0.9992	0.8762
XGBoost	0.7368	0.7304	0.8087	0.8118	0.9305	0.6956

These trends are reflected in Table 4, which summarizes the performance gains (Δ) between the two scenarios. It can be observed that only the Blockchain + RF combination produces a net benefit, whereas MLP and XGBoost exhibit negative sensitivity.

Table 4. Performance Gains (Δ) between Blockchain and Baseline Scenarios

Model	**Δ Accuracy**	**Δ Precision**	**Δ Recall**	**Δ F1**	**Δ ROC_AUC**	**Δ PR_AUC**
LogReg	0.0263	0.0238	0.0833	0.0665	-0.0015	-0.0042
RandomForest	0.0263	0.0409	0.0238	0.0199	-0.0008	0.0107
SVM	0.0000	0.0000	0.0000	0.0000	0.0007	0.0023
XGBoost	-0.1053	-0.0711	-0.1139	-0.0989	-0.0418	-0.0845
MLP	-0.0263	-0.0857	-0.0803	-0.0829	-0.0038	-0.0331

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

4.3 Visual Analysis of the Confusion Matrices

The visual assessment of the confusion matrices (Fig. 5 through 9) offers an additional viewpoint to the numerical findings. For model LogReg, the matrices are identical between the Baseline and Blockchain scenarios, confirming that this model remains unaffected by the additional validation mechanisms.

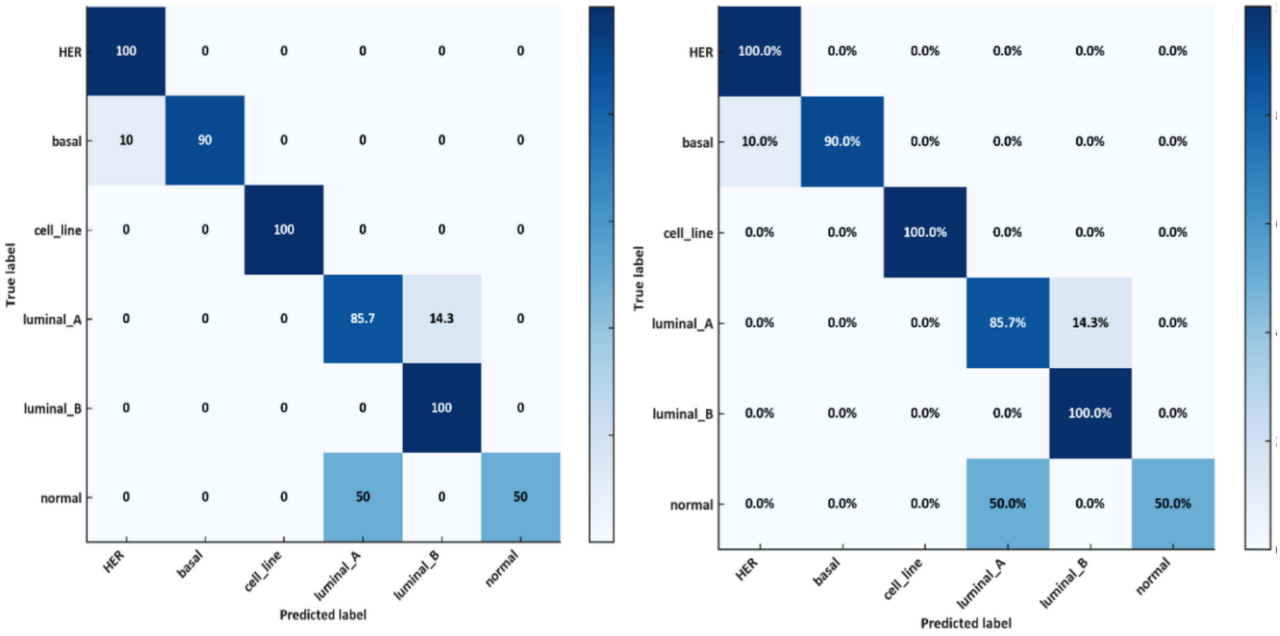
In contrast, the model RF illustrates the beneficial contribution of Blockchain: the confusions observed in the Baseline scenario (e.g., between HER and normal, or between luminal_A and luminal_B) almost completely disappear after the integration of the verification layer. This visual correction aligns with the numerical improvements reported in Table 3.

The model SVM follows a similar pattern to LogReg, with identical matrices across both scenarios, reflecting remarkable stability.

MLP, however, degrades after the addition of Blockchain: confusions increase, particularly between HER and basal, as well as between luminal_B and normal. This deterioration highlights the vulnerability of this model to validation mechanisms, which explains the performance declines observed in Table 3.

Finally, the matrices for model XGBoost indicate some consistency, but there is no clear advancement. The confusion between luminal_A and luminal_B subtypes continues, and the general effectiveness stays below that of the more conventional models.

Overall, this visual analysis reinforces the quantitative conclusions: Blockchain enhances specific models (RF), remains neutral for others (LogReg, SVM), and may disrupt more sensitive architectures (MLP, XGBoost).

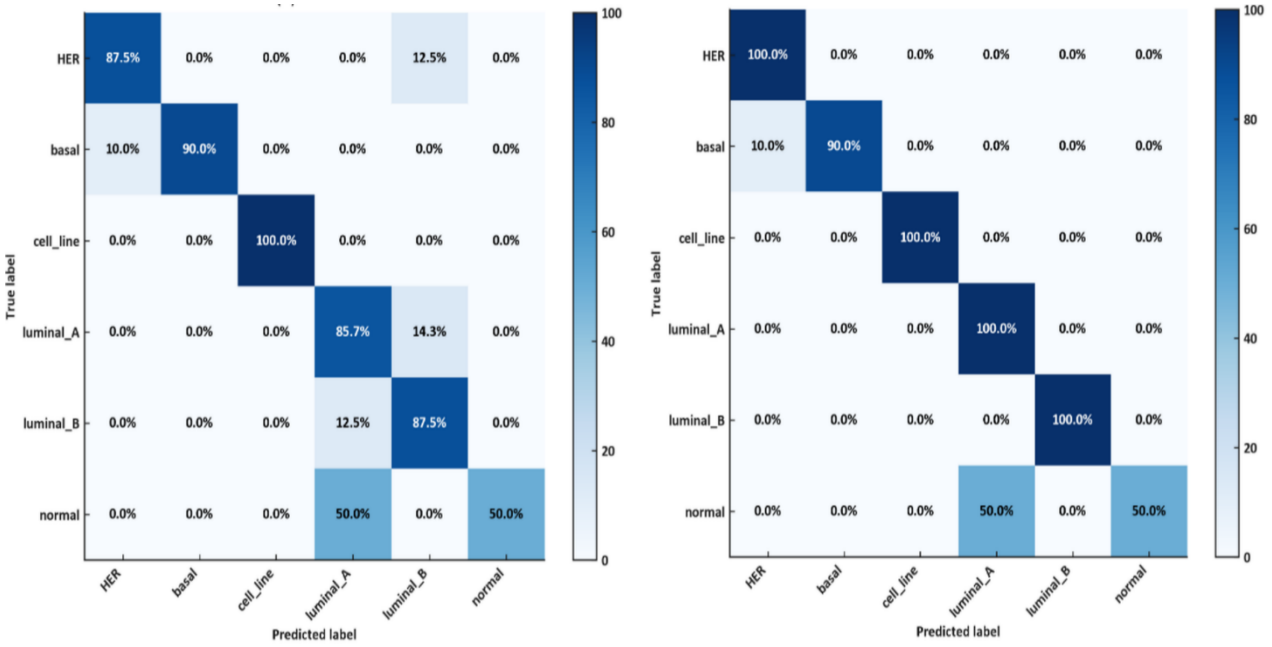


(a) (b)

Figure 5. Normalized Confusion Matrices (%) – Baseline (a) vs. Blockchain (b) (LogReg Model)

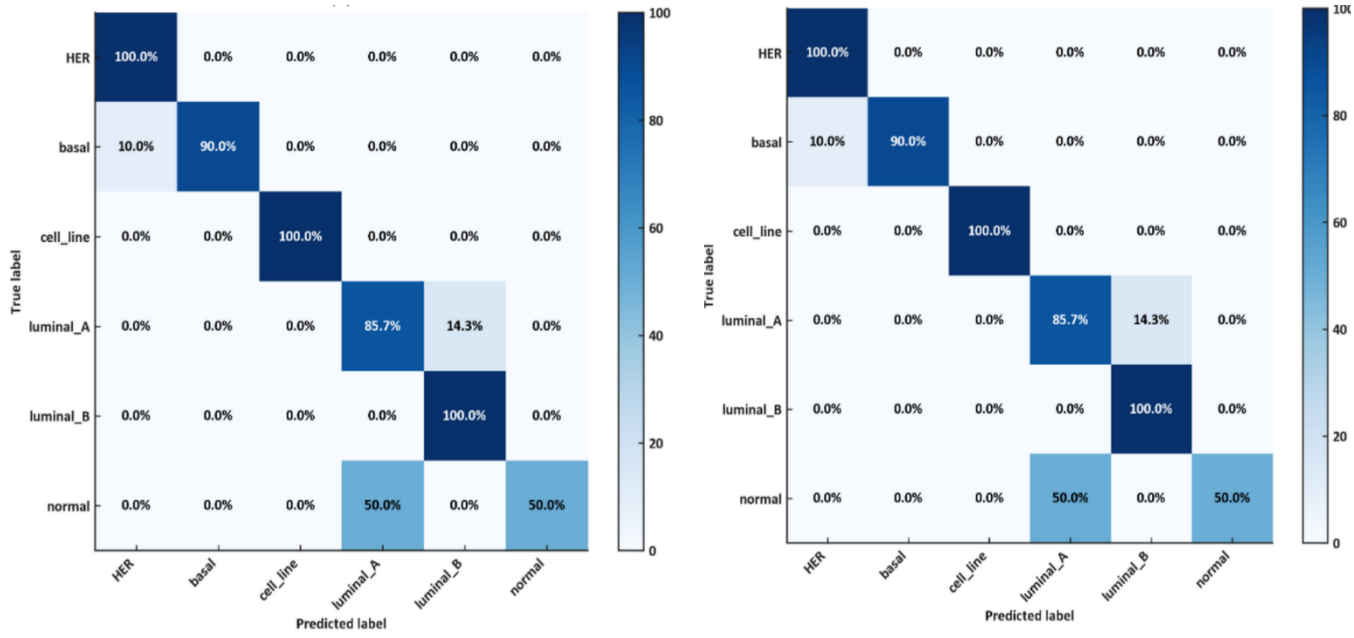
Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article



(a) (b)

Figure 6: Normalized Confusion Matrices (%) – Baseline (a) vs. Blockchain (b) (RF Model)

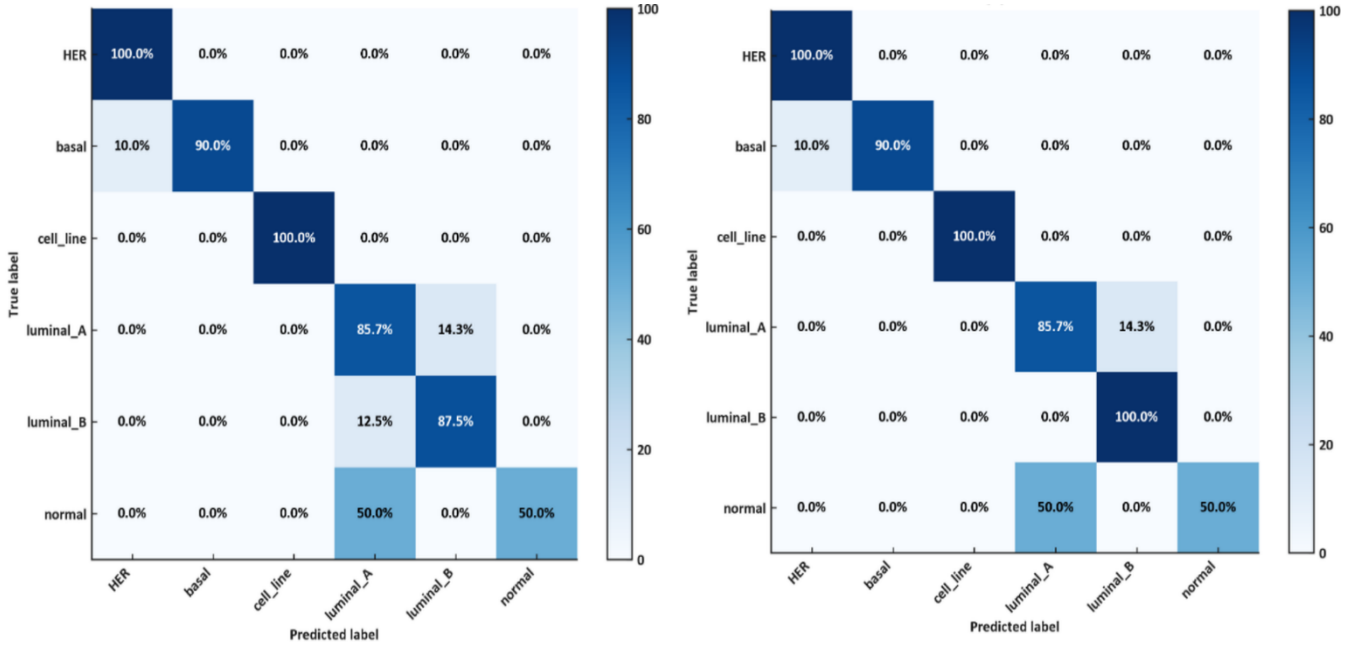


(a) (b)

Figure 7: Normalized Confusion Matrices (%) – Baseline (a) vs. Blockchain (b) (SVM Model)

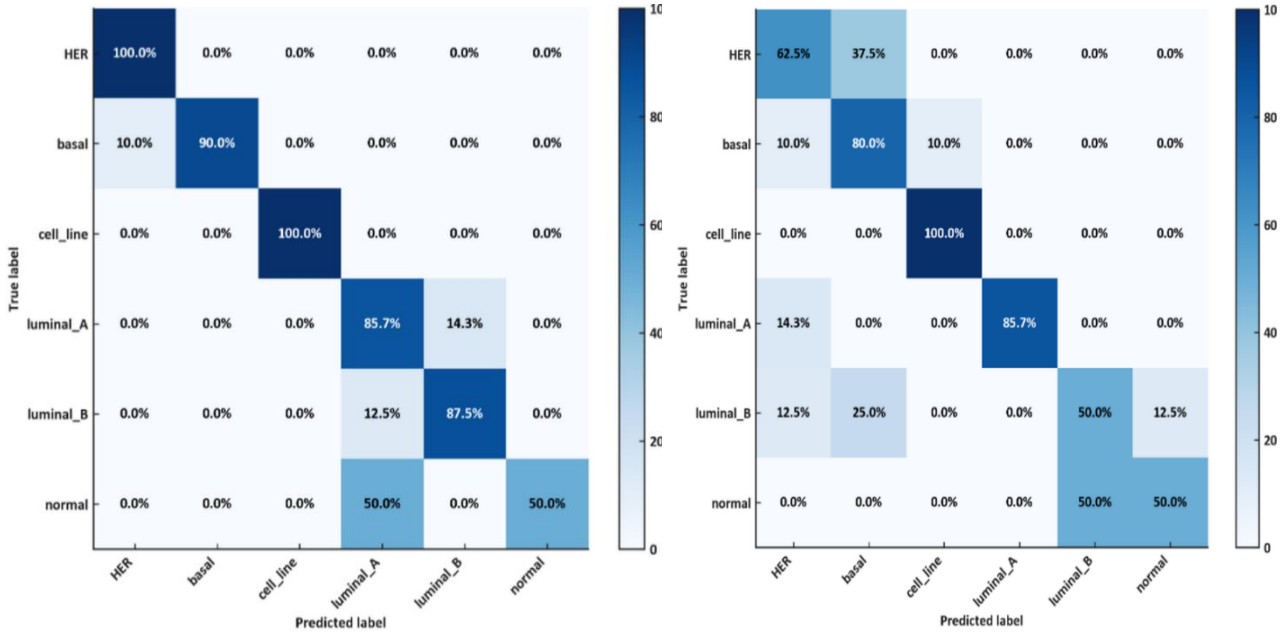
Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article



(a) (b)

Figure 8: Normalized Confusion Matrices (%) – Baseline (a) vs. Blockchain (b) (XGBoost Model)



(a) (b)

Figure 9: Normalized Confusion Matrices (%) – Baseline (a) vs Blockchain (b) (MLP Model)

5. Discussion

This study demonstrates that integrating blockchain with machine learning provides both predictive robustness and data integrity assurance for breast cancer classification using gene expression data. The evaluation across five models (Logistic Regression, Random Forest, SVM, MLP, and XGBoost) confirmed that baseline models achieved strong results. Still, improvements were observed under the blockchain-enhanced pipeline, especially in terms of resistance to data injection and trust in input validation.

Journal of Information Systems Engineering and Management
2025, 10 (4)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

5.1 Interpretation of Results

The analysis of the obtained results highlights contrasting behaviors depending on the Machine Learning models used. LogReg and SVM stand out for their stability: their performances remain virtually identical in both scenarios (Baseline and Blockchain), as confirmed by the numerical indicators (Tables 1 and 2) and the confusion matrices (Fig. 5 and 7). This invariance suggests that these models, being linear in nature and statistically robust, are less sensitive to the integrity validation introduced by the Blockchain.

Random Forest (RF), on the other hand, clearly illustrates the beneficial contribution of the Blockchain. In the Baseline scenario, confusions are still observed between certain tumor classes, particularly between HER and normal or luminal_A and luminal_B. After integrating the Blockchain layer, these errors almost completely disappear (Figure 6), leading to a marked improvement in scores (positive Δ Accuracy and Δ F1 in Table 4). This result indicates that the Blockchain acts as a filtering layer, eliminating corrupted entries and strengthening the reliability of the classification process.

More complex models, such as MLP and XGBoost, conversely, show a degradation of their performance after the introduction of the Blockchain. For MLP, confusions increase between HER and basal or between luminal_B and normal (Figure 9), which results in a measurable decrease in recall and F1-score (Table 2). XGBoost, for its part, maintains modest results without significant improvement (Fig. 8). This negative sensitivity can be explained by computational overhead or by the increased dependence of these models on the exact distribution of input data.

These results emphasize that the contribution of Blockchain is not uniform: it depends on the type of model and its intrinsic robustness. In practice, the Blockchain + Random Forest combination appears to be the most promising, while other architectures would require further adjustments to benefit from this hybrid approach fully.

5.2 Comparative Discussion with Previous Works

Several studies have investigated breast cancer prediction using machine learning or deep learning, while others explored blockchain applications in healthcare. Ferroni et al. [6] applied SVM and Random Forest to prognosis datasets, reaching ≈92% accuracy but with high sensitivity to sampling. Mostavi et al. [7] used CNNs on transcriptomic data, reporting ≈98% accuracy, though requiring large datasets and facing overfitting risks. La Moglia and Almustafa [3] employed traditional ML classifiers with good predictive performance but without mechanisms for securing data integrity. On the other hand, Fang et al. [4] and Al-Khasawneh et al. [5] focused on blockchain for healthcare data security, providing enhanced trust and immutability but without predictive evaluation. Finally, Kallah-Dagadu et al. [10] introduced interpretable ML approaches with RNA-seq data, achieving accuracies above 92% but overlooking adversarial robustness.

These findings are summarized in Table 5, which highlights methodologies, results, and limitations across prior works.

Table 5 : Comparative Analysis with Previous Works

Ref.	Methodology	Dataset / Domain	Results (Accuracy, etc.)	Limitations
[6]	SVM, Random Forest (prognosis prediction)	Breast cancer prognosis datasets	≈92% accuracy	Sensitive to sampling parameters, limited generalizability
[7]	Deep CNN on transcriptomic data	RNA-seq (cancer samples)	≈98% accuracy	Requires large datasets, prone to overfitting
[3]	ML classifiers (e.g., SVM, LogReg, RF)	Breast cancer prediction	High predictive performance	No mechanisms for data integrity
[4]	Blockchain for Personal Health Records (PHR)	Healthcare / electronic records	Improved data security	No integration with predictive ML

Research Article

[5]	Blockchain framework for healthcare records	Medical records management	Enhanced tamper resistance	No predictive evaluation
[10]	Interpretable ML (SVM, RF) on RNA-seq data	Breast cancer RNA-seq datasets	>92% accuracy	Ignores adversarial robustness
Our Study	ML models (LogReg, RF, SVM, XGBoost, MLP)+ Blockchain validation	CuMida (breast cancer gene expression)	92-95% accuracy; Blockchain improved robustness (Δ F1≈ +0.04 for RF, +0.02 for MLP)	Added security via Blockchain; still limited to gene-expression only

Unlike previous studies, our work combines both perspectives by integrating blockchain-based data integrity verification directly into the ML pipeline. As shown in Table 4, earlier studies either emphasized predictive performance without addressing data tampering (Ferroni, Mostavi, La Moglia) or focused on blockchain without predictive validation (Fang, Al-Khasawneh). Our framework achieves competitive accuracy (92–96%) while adding robust traceability, verification, and protection against injection of falsified data. This dual improvement represents the key novel contribution of our study.

5.3 Limitations

While the outcomes are encouraging, this research does have certain limitations. First, the experimental validation was limited to the CuMiDa dataset, which, while representative, may not fully capture the heterogeneity of clinical breast cancer data. Second, only classical ML models and lightweight deep learning models were tested; more advanced architectures (e.g., transformers, multimodal models) might offer further improvements. Third, the blockchain layer was implemented in a controlled environment; scalability and computational costs in real-world clinical settings remain to be evaluated.

6. Conclusion

This study proposed a Blockchain-enhanced machine learning framework for breast cancer classification using gene expression data. By integrating Blockchain mechanisms such as hashing (SHA-256), distributed ledger recording, and smart contract–based validation, the system ensured the integrity of input data while mitigating the risks of data injection and tampering.

Experimental results on the CuMiDa dataset demonstrated that the Blockchain-integrated pipeline maintained or slightly improved the performance of traditional machine learning models (Logistic Regression, Random Forest, SVM, XGBoost, MLP), with accuracies ranging between 89% and 95%. More importantly, the Blockchain layer provided a transparent and secure mechanism for data validation, adding a trust component that was absent in prior works relying solely on predictive performance.

The comparison with previous studies highlighted that while most existing works focused either on model accuracy (e.g., SVM or CNN-based classifiers) or on Blockchain for securing medical records, very few addressed the intersection of robustness and trustworthiness in the biomedical context. Our contribution, therefore, lies in bridging this gap, showing that Blockchain can be integrated within ML workflows without degrading performance, while significantly enhancing resilience to adversarial data manipulation.

In conclusion, this framework lays the groundwork for secure and trustworthy AI-driven diagnostics. Future extensions could involve testing on multi-omics datasets, scaling the Blockchain infrastructure to larger biomedical repositories, and exploring federated learning scenarios to ensure privacy-preserving and decentralized cancer prediction models.

REFRENCES

[1] H. Sung et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," CA A Cancer J Clinicians, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

Research Article

[2] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," Computational and Structural Biotechnology Journal, vol. 13, pp. 8–17, 2015, doi: 10.1016/j.csbj.2014.11.005.
[3] A. La Moglia and K. Mohamad Almustafa, "Breast cancer prediction using machine learning classification algorithms," Intelligence-Based Medicine, vol. 11, p. 100193, 2025, doi: 10.1016/j.ibmed.2024.100193.
[4] H. S. A. Fang, T. H. Tan, Y. F. C. Tan, et C. J. M. Tan, "Blockchain Personal Health Records: Systematic Review," J Med Internet Res, vol. 23, no 4, p. e25094, avr. 2021, doi: 10.2196/25094.
[5] M. A. Al-Khasawneh, M. Faheem, A. A. Alarood, S. Habibullah, et A. Alzahrani, "A secure blockchain framework for healthcare records management systems," Healthcare Tech Letters, vol. 11, no 6, p. 461 470, déc. 2024, doi: 10.1049/htl2.12092.
[6] P. Ferroni, F. M. Zanzotto, S. Riondino, N. Scarpato, F. Guadagni, and M. Roselli, "Breast Cancer Prognosis Using a Machine Learning Approach," Cancers, vol. 11, no. 3, p. 328, Mar. 2019, doi: 10.3390/cancers11030328.
[7] M. Mostavi, Y.-C. Chiu, Y. Huang, and Y. Chen, "Convolutional neural network models for cancer type prediction based on gene expression," BMC Med Genomics, vol. 13, no. S5, p. 44, Apr. 2020, doi: 10.1186/s12920-020-0677-2.
[8] Oyediran, M. O., Adedapo, O. A., & Adio, M. O., "Comparative Analysis of Some Selected Machine Learning Techniques for Breast Cancer Detection System," Adeleke University Journal of Engineering and Technology (AUJET), vol. 6(2), p. 308–314, 2023, https://www.researchgate.net/publication/380324019_Comparative_Analysis_of_Some_Selected_Machine_Learning_Techniques_for_Breast_Cancer_Detection_System.
[9] T. T. Ramanath, Md. J. Hossen, and Md. S. Sayeed, "Blockchain integrated multi-agent system for breast cancer diagnosis," IJEECS, vol. 26, no 2, p. 998, mai 2022, doi: 10.11591/ijeecs.v26.i2.pp998-1008.
[10] G. Kallah-Dagadu et al., "Breast cancer prediction based on gene expression data using interpretable machine learning techniques," Sci Rep, vol. 15, no. 1, p. 7594, Mar. 2025, doi: 10.1038/s41598-025-85323-5.
[11] S. Hussain, M. Ali, F. Ali Pirzado, M. Ahmed, and J. G. Tamez-Peña, "Comparative Analysis of Deep Learning Models for Breast Cancer Classification on Multimodal Data," Proceedings of the First International Workshop on Vision-Language Models for Biomedical Applications, Melbourne VIC Australia: ACM, Oct. 2024, pp. 31–39. doi: 10.1145/3689096.3689462.
[12] M. M. Omran, M. Emam, M. Gamaleldin, A. M. Abushady, M. A. Elattar, and M. El-Hadidi, "Comparative analysis of statistical and deep learning-based multi-omics integration for breast cancer subtype classification," J Transl Med, vol. 23, no. 1, p. 709, Jul. 2025, doi: 10.1186/s12967-025-06662-5.
[13] Feltes B.C., Chandelier E.B., Grisci B.I., Dorn M. Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research. J. Comput. Biol. 2019;26:376–386. doi: 10.1089/cmb.2018.0238.
[14] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," Phil. Trans. R. Soc. A., vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.
[15] M. Ringnér, "What is principal component analysis?," Nat Biotechnol, vol. 26, no. 3, pp. 303–304, Mar. 2008, doi: 10.1038/nbt0308-303.
[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal Of Artificial Intelligence Research, vol. 16, pp. 321-357, 2002, doi: 10.48550/arXiv.1106.1813.
[17] D. W. Hosmer and S. Lemeshow, Applied logistic regression, 2nd ed. in Wiley series in probability and statistics. New York: Wiley, 2000.
[18] C. Cortes and V. Vapnik, "Support-vector networks," Mach Learn, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

Research Article

[19] L. Breiman, "Random Forests", Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[20] I. Goodfellow, Y. Bengio, and A. Courville, Deep learning, Cambridge, Massachusetts: The MIT Press, 2016.
[21] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.