Received: 17 Dec 2025, Accepted: 30 Dec 2025, Published: 04 Jan 2026
Digital Object Identifier: https://doi.org/10.63503/j.ijssic.2025.207

Review Article

Breast Cancer: Aetiology, Diagnosis, Prevention, Treatment, and the Transformative Role of Artificial Intelligence

Blessing Oloko[1], Xiaochun Cheng[2*]

[1] Department of Computer Science, University of Lagos, Faculty of Science, Nigeria.

[2] Swansea University, Welsh computer science department, Swansea, UK.

Karodan01@gmail.com[1], xiaochun.cheng@swansea.ac.uk[2]

*Corresponding author: Xiaochun Cheng, xiaochun.cheng@swansea.ac.uk.

ABSTRACT

Breast cancer is a life-threatening disease affecting women, and it needs effective and efficient diagnostic procedures to detect it early enough to ensure effective management. Artificial Intelligence (AI) offers significant promise for improving breast cancer detection and treatment; however, its application remains constrained by methodological, ethical, and infrastructural limitations. This paper examines a hybrid anomaly-detection model applied to the breast cancer dataset from Kaggle. The dataset is preprocessed and class-balanced using Mean Absolute Error (MAE) thresholding and SMOTE techniques. Twelve ML and DL models were trained and compared using the dataset in terms of typical evaluation metrics. The highest diagnostic performance was obtained with the LightGBM classifier with 0.9912 Accuracy, 0.9880 F1 Score, and a Specificity (1.0000). The Results indicate the usefulness of ensemble techniques in high-accuracy biomedical anomaly detection. Persistent issues with insufficient dataset diversity, model interpretability, and clinical standards should be resolved to facilitate the deployment of reliable AI systems in the fields of breast cancer detection, prevention, and therapy.

Keywords: *Breast cancer, Aetiology, Diagnosis, Treatment, Prevention, Deep learning, Explainable AI.*

1. Introduction

Breast cancer develops as a result of the failure of self-regulation in multicellular organisms. It arises from abnormal communication within an otherwise self-organizing multicellular organism, driven by genetic defects that lead to excessive proliferation of cells and tissues in specific body regions [38]. The malignant characteristics of unchecked proliferation, anti-apoptotic response, angiogenesis, and metastasis are caused by progressive genetic and epigenetic instability [25] [38] [41]. The tumour is a severe and progressive multigenic disease characterized by cell miscommunication, which has detrimental effects on the entire organism [38] [39] [40] [85]. It is not caused by the failure of a single gene or even several genes. Single-nucleotide mutations, copy number variations, and aberrant splicing that affect the isoforms—such as CD44, which is linked to invasion and metastatic dissemination—are examples of the many genomic alterations. Investigations revealed that infection with pathogens, including viruses, can activate transposable elements, which cause genomic rearrangements that can be transmitted into subsequent generations [38]

The majority of Cancers (90-95%) are not inherited [31]. They are brought on by unplanned genetic alterations in old age and environmental influences (e.g., tobacco, radiation). These malignancies are referred to as "spontaneous" or "non-hereditary" cancers. Only 5–10% of women develop breast cancer due to deleterious mutations (hereditary or familial malignancies). The lack of a family history does not always mean that there is no risk. Cancer will be found in about 38 percent of individuals during their lifetime. Preventive measures of non-hereditary cancers are lifestyle habits (not smoking, healthy weight, exercise). The vast majority (approximately 90–95%) are non-hereditary or sporadic cases, arising from somatic mutations that accumulate over an individual's lifetime due to aging and environmental exposures [11] [24] [32] [79].

Multi-omics integration is indispensable for achieving a truly comprehensive understanding of complex biological processes and systems. Integration of AI with multi-omics (genomics, transcriptomics, proteomics, metabolomics) facilitates precision oncology in the areas of Imaging and Digital Pathology (CNNs detect subtle patterns; GANs generate synthetic images for rare cancers), Data Challenges (Heterogeneity, missing values, and dataset bias impede reproducibility and generalizability) and Clinical Workflow Barriers (Integration requires training, infrastructure adaptation, and alignment with regulatory frameworks) [82]. This involves the studies of non-traditional liquid biopsy components, including tumour-educated platelets [4] [7] [51] [67]. Cancer risk factors range from non-modifiable genetic determinants to highly modifiable lifestyle behaviours that increase susceptibility to the disease. Oestrogen is required for the proliferation of epithelial cells. The cumulative exposure to Oestrogen increases risks as observed by early menarche, late menopause, nulliparity, and hormone replacement therapy [3] [38] [41]. Some of the key contributors to carcinogenesis include obesity, sedentary behaviour, and large amounts of alcohol ingestion, smoking, and exposure to ionizing radiation. Certain malignancies, like breast and colon cancer, are known to be predisposed by obesity, which is attributed to diets rich in added sugars. High sugar intake can also promote systemic inflammation, which might also augment the risk of cancer. Since the risk of cancer is not only associated with obesity but also other chronic ailments such as diabetes and cardiovascular disease, it has been suggested to reduce the consumption of added sugars and total intake of a balanced diet rich in fruits, vegetables, whole grains, and lean proteins. The latter are included in the feature sets of AI-assisted risk stratification models and are the primary focus of population-level preventative initiatives. [2] [10] [33] [38] [41] [45]. AI has entered modern biomedicine and is being used in diagnostics, drug discovery, molecular biology, and clinical decision-making.

The diagnostic process comprises physical examination, comprehensive history-taking and symptom assessment, advanced imaging, laboratory tests, differential diagnosis, and tissue biopsy. This structured approach enhances diagnostic accuracy [83]. Mammography is one of the widely used screening procedures for breast cancer. It is still the basis of breast cancer screening at the population level since it is shown to be effective in minimizing the mortality rate related to breast cancer, but a significant number of overdiagnoses is also observed [23][29][38]. Ultrasound is the most widely utilized imaging method for tumour staging and biopsy guidance in the clinical routine of breast cancer screening. Compared to mammography, it is more affordable to buy and maintain, portable, and more adaptable. Increasing the visibility of the thick breast tissue and accurately directing the biopsy are two important functions of ultrasound. However, it is highly operator-dependent [3] [23] [29] [66].

Even though it is the most expensive technique, magnetic resonance imaging (MRI) possesses low specificity and sensitivity in detecting breast cancer compared to mammography and ultrasound. [24] [29] [30] [38] [56]. Conflicting evidence, small sample sizes, and the lack of standardized criteria limit clarity regarding the definitive role and cost-effectiveness of hybrid imaging modalities (PET/CT and PET/MRI) in the overall management of breast cancer [23]. The definitive histopathologic examination of tissue samples remains the gold standard for confirming malignancy, tumour grading, and assessing the status of key receptors (ER, PR, and HER2) [30, 52, 53].. The study focused on Artificial Intelligence models—specifically, machine learning and deep learning approaches—for breast cancer detection and stage classification. These models were compared to evaluate their effectiveness in identifying breast cancer and accurately determining its stage.

2. Literature review (Related works)

Articles and books examining the technical, data, and infrastructural, as well as clinical, ethical, and legal challenges that hinder the widespread adoption and reliability of Artificial Intelligence (AI) in biomedical, biological sciences, and healthcare settings were analysed.

Technical and Methodological Loopholes.

- **Limited Scope of AI Methodologies:** Existing studies primarily emphasize Convolutional Neural Networks (CNNs) and supervised learning models. On the other hand, emerging and potentially transformative approaches—such as reinforcement learning, unsupervised learning, generative artificial intelligence, and Large Language Models (LLMs)—remain insufficiently explored in the academic literature. Expensive and more systematic investigation

of these methodologies could substantially improve the predictive robustness, automation, and interpretability of AI solutions when applied to objective, valid, and reliable biomedical data [2][3][4][6][7][10][12][13][14][15][16][17][20][25][26][29][30][31][32][33][36][38][39][42][46][47][48][49][70].

- **Limited Comparative Model Evaluation:** Few studies conduct standard, head-to-head comparisons of AI architectures (e.g., Convolutional Neural Networks versus Vision Transformers) using large, heterogeneous datasets. This lack of rigorous benchmarking impedes the identification of the most effective models for specific biomedical tasks and undermines replicability, thereby impeding the selection of robust and generalizable solutions [27][3][7][25][29][49][50][51][61][64]
- **Explainability Challenges:** Deep learning models are often viewed as "black boxes," which limits clinical trust and regulatory acceptance. There is a compelling need for domain-specific explainable AI (XAI) approaches that provide clinically meaningful interpretations while maintaining high predictive performance [13][25][26][29][48][49][50]
- **Multi-objective Optimization:** The application of multi-objective evolutionary algorithms (MOEAs) to chores such as biomarker selection remains largely unexplored. Existing approaches rarely incorporate biologically grounded objective functions, uncertainty estimates, or automated decision-support mechanisms, limiting their practical utility in biomedical research [42].

Information and Information Systems Lapses

- **Lack of Dataset Diversity:** The availability of high-quality biomedical datasets remains limited, heavily skewed toward oncology, which restricts model robustness and generalizability across diverse populations and disease domains [3] [38].
- **Poor Standardisation and Interoperability**: conflicting terminology, ontologies, and data formats present significant barriers to data sharing and integration across biomedical institutions, particularly in clinical natural language processing (NLP) and the omics sciences. In the chemical domain, NLP challenges arise from out-of-order punctuation and complex chemical expressions. Existing tools, such as OSCAR4 and Chemical Tagger, rely essentially on rule- and dictionary-based methods that perform well for chemical entity recognition but are less effective for general text tokenisation. While machine learning approaches have shown promise in chemical NLP—owing to the relatively distinct structure of chemical expressions—general interoperability issues persist.
  Furthermore, terminological imprecision in biomedical language further complicates standardisation. For example, "tumour" or "neoplasm" mainly denotes abnormal tissue growth, whereas "cancer" refers specifically to malignant neoplasms with invasive and metastatic potential. While all cancers are tumours, not all tumours are cancers, and the commonly used clinical descriptors "benign" and "malignant" are probabilistic rather than categorical. Such ambivalences underscore the need for carefully standardised vocabularies and reproducible, collaborative research frameworks to ensure consistency, reliability, and effective knowledge exchange across studies [31] [38] [49].
- **Inadequate Cost-Benefit Evidence:** There are still few rigorous economic analyses evaluating the viability, applicability, and practicality of using expensive AI systems in therapeutic contexts. This limits informed decision-making regarding large-scale clinical adoption and sustainability [3] [4][17][28][41][49][64].
- **Knowledge Graphs and Translational Informatics:** State-of-the-art translational information processing systems require further enhancement through the integration of knowledge graphs (e.g., Biolink). Such frameworks can enable finer-grained semantic representations and evidence-weighted reasoning over multimodal biomedical data, thereby improving inference and translational utility [11][47]

Clinical, Ethical, and Legal Gaps

- **Impact on Physician–Patient Relationships:** The effects of AI-assisted clinical decision-making on trust, communication, and informed consent remain unresolved. Empirical studies are needed to better understand how AI integration affects patient outcomes and clinician behaviour in real-world practice [3] [5] [13].
- **Legal Liability and Accountability:** The unified frameworks defining legal responsibility in cases of AI-related medical errors are currently lacking, creating significant uncertainty regarding liability among clinicians, institutions, and AI developers [38][48].
- **Sensitive Data Governance:** Strong, ethical, and GDPR-compliant biomedical data governance systems are still lacking. Methods such as federated learning, synthetic data generation, and privacy-preserving AI require further refinement and standardization to enable secure and responsible data use [28].
- **Disease-Area Imbalance:** AI research remains markedly focused on oncology, with critical areas such as cardiovascular, neurological, and rare diseases receiving relatively limited attention. This imbalance risks biasing innovation and limiting equitable healthcare outcomes [6] [9][14][16].
- **Healthcare Access Inequities:** Unequal access to AI-enabled services based on geographic location or socioeconomic status may exacerbate healthcare delivery disparities.
- **Algorithmic Bias ("Coded Inequity"):** AI systems can sustain or amplify existing biases when trained on incomplete or unrepresentative datasets. Underrepresentation of specific demographic groups may result in diminished diagnostic accuracy and suboptimal treatment recommendations. Active bias detection, mitigation, and fairness-aware model design must be extensively integrated into feature engineering and model development pipelines [2][26][31][36][38][46][48][49][51][63][64][72][73].

Emergence of Trustworthy AI

Toward Authoritative and Trustworthy AI: There is a growing need for formal, measurable frameworks that explicitly model trade-offs among core dimensions of trustworthiness—such as accuracy, explainability, privacy, and fairness. This includes the development of:

- Measurable instruments to measure relationships among competing objectives;
- Standardized metrics for evaluating the comprehensibility and usefulness of explainable AI (XAI) across diverse stakeholders [13][25][29][48][50][51].

Established Applications of Deep Learning in Diagnostics

Radiomics (Medical Imaging Analysis): Convolutional Neural Networks (CNNs) have demonstrated strong performance in automated lesion detection, segmentation, micro-calcification analysis, and objective breast density assessment. Additionally, AI-driven radiogenomic studies progressively link imaging phenotypes with underlying molecular drivers [3] [13][16].

Digital Pathology (Whole-Slide Image Analysis): AI systems excel in high-precision counting and classification tasks, including detection of localized micrometastases, automated tumour grading, and analysis of spatial interactions within the tumour microenvironment [7][17].

Advanced AI Architectures in Diagnostic Systems include the following:

- **Graph Neural Networks (GNNs):** GNNs model interactions among neighbouring nuclei, tissue regions, or multimedia imaging features, making them particularly valuable for complex, multi-label diagnostic tasks where spatial and topological relationships are critical [10][11][15][58].
- **Neural Architecture Search (NAS):** NAS automates the discovery of optimal CNN or GNN architectures, accelerating model development and often achieving superior performance in classification and segmentation tasks while lessen reliance on manual design expertise [58].

- **The Interpretability Imperative (XAI):** For successful clinical deployment, AI systems must move beyond opaque "black-box" models. Explainability methods that offer measurable and visual insights into model decision-making include SHAP, LIME, and saliency maps. Regulatory approval, clinical validation, and long-term clinician trust all depend on the interpretability of the models.

3. Research Methodology

The study technique is based on artificial intelligence (machine learning and deep learning) models, which have often been used in breast cancer research. These models were then trained on a breast cancer dataset from Kaggle (https://www.kaggle.com/datasets/yasserh/breast-cancer-dataset?select=breast-cancer.csv), and a comparison of their performance was conducted. The models' classification performance was evaluated using common metrics such as accuracy, precision, recall, F1 score, etc., in order to identify which model performed best.

- Biomedical dataset contains sensitive personal details which indicates possible disease/diagnose, in the case of breast cancer , Malignant(M) and Benign(B) cancer making it suitable for supervised learning [3][8][10][17][18][26][27][28][30][31][38][41][42][44][45][46][50]. However, the study adapts it for unsupervised learning.

Key data preparation steps include:

- Checking for and handling missing values.
- Dropping attributes with only one unique entry.
- Transforming categorical variables into numerical values.
- Addressing the data imbalance (a common issue in biomedical datasets) by oversampling with SMOTE (Synthetic Minority Oversampling Technique) to prevent under-fitting and overfitting before splitting the dataset.
- For unsupervised learning models, a 569-data-point dataset with 33 attributes, named 'Breast-cancer.csv', was used to train models.
- The core of the system is a Hybrid learning (machine learning and deep learning models) anomaly detection model trained on biomedical data 'breast-cancer.csv'.
- **Training Phase:** The first half of a dataset is used as "normal" data for training a Hybrid learning model.
- **Prediction and Error Calculation:** The trained hybrid model predicts values across the entire dataset (both training and testing halves). The Mean Absolute Error (MAE) is calculated of the expected and actual outputs.
- **Anomaly Thresholding:** A threshold for prediction error is set using the 95th percentile (top 5%) of the training prediction errors. Any error exceeding this boundary is flagged as an anomaly.
- **Anomaly Detection:** The established threshold is then applied to the testing data. Anomalies are defined as errors in the testing set that exceed this threshold.
- This hybrid approach works because training only on "normal" data teaches the model to recognize typical patterns. When it encounters an anomaly in unseen data, its forecast will significantly alter, resulting in a high error that exceeds the predefined threshold.

Software Environment and Libraries:

The project utilizes Python 3.12.4 (Anaconda distribution) and several key libraries:

- Pandas for data manipulation.
- Scikit-learn for machine learning and pre-processing.
- NumPy for numerical computations.
- TensorFlow 2.19.0 for deep learning.

- Additional libraries like Statsmodels, Seaborn, Plotly, SciPy, Keras, Lime, shap and Matplotlib for data analysis.
- The studies also emphasize rigorous processing of real-world, imperfect data, including handling missing values, outliers, and incorrect entries[5][3][8][15][25][26][29][32][36][6].

Dataset Splitting and Model Evaluation:

The data set is methodically separated into:

- **Training Set (60-80%):** Used for model learning.
- **Testing Set (10-20%):** For evaluating performance on unseen data (generalization ability).
- **Validation Set (10-20%):** An optional subset from the training split, used for experimentation, prototyping, and detecting overfitting during iterative model training. The model must never train on validation or test sets to ensure unbiased evaluation.

Machine Learning Paradigms

The machine learning paradigms are:

- **Supervised Learning:** Models (such as classification and regression) learn from labelled data [4] [68] [73].
- **Unsupervised Learning**: Models (such as Dimensionality Reduction and Clustering) identify patterns in unlabelled data.
- **Semi-supervised Learning:** Combining elements of supervised and unsupervised learning is known as semi-supervised learning. [62].

Model Evaluation Metrics

The importance of detailed model evaluation beyond simple accuracy is emphasized, especially for imbalanced datasets [5] [44] [65].

Loss Functions: Common loss functions include:

- For classification, use log loss or cross-entropy loss.
- •For regression, the mean-squared error (MSE) is used [31].

Confusion Matrix: Provides a detailed view of performance by categorizing predictions into:

- True Positive (TP)[65]
- False Positive (FP) (Type I error/false alarm)[65]
- True Negative (TN)[65]
- False Negative (FN) (Type II error/missed detection)[66]

Key Classification Metrics:

- Precision: The percentage of expected positives that turn out to be actual positives [18] [67].
- Recall (Sensitivity): The percentage of true positives that were accurately identified [17] [68].
- F1-measure (F1-Score): The precision and recall harmonic means [69] [70].
- True Negative Rate (Specificity): The percentage of real negatives that are accurately detected [53] [64].
- Precision-Recall Trade-off and ROC AUC: Talks about how to balance precision and recall, as well as how the Receiver Operating Characteristic (ROC) curve uses Area Under the Curve (AUC) to distinguish across classes overall [18].
- Regression Metrics: Unlike classification, regression models predict continuous, real-valued outputs. The primary metrics measure the deviation between predicted and true values:
- The Mean Squared Error (MSE) is susceptible to breast cancer anomalies.
- The Mean Absolute Error (MAE) is more resilient to breast cancer anomalies.
- Mean Absolute Percentage Error (MAPE) is beneficial for data that is quite variable [80].

Overfitting and Bias-Variance Trade-off

- A model is said to be overfit when it performs poorly on unknown data (low bias, high variance) because it has learned the training data, including noise, too well. [3][71] [72].
- A model that performs poorly on both training and test data (high bias) is said to be underfitting if it is either too simplistic or not sufficiently trained. Generalization is achieved when both datasets exhibit identical performance [31].

Preventing Overfitting and Enhancing Generalization

Strategies include:

- Loss Penalty (Regularization)[66]
- Reducing Complexity
- Data Processing
- Hyperparameter Tuning (e.g., using Grid Search)[74]
- Adding Training Data
- Validation Strategies: It is emphasized that models must never train on validation or test sets and that hold-out validation and K-Fold Cross-Validation are essential for avoiding overfitting, model selection, and hyperparameter tweaking [44][73].

Machine Learning Concepts and LightGBM Implementation

LightGBM (Light Gradient Boosting Machine) is a machine learning model categorised as an ensemble method [64]. It is a gradient-boosting framework that uses decision trees as its base learners. LightGBM builds a strong predictive model by combining multiple decision tree models iteratively, correcting the errors of the previous trees [17].

Key characteristics that distinguish LightGBM as a machine learning model:

- **Tree-based learning:** It builds a decision tree, a key element of conventional machine learning.
- **Gradient Boosting:** lightGBM uses the gradient boosting technique iteratively to improve predictions.
- **No neural networks:** Unlike deep learning models, LightGBM does not use neural networks or their layered architectures [17] [18][64][74].

LightGBM is designed for efficiency, scalability and high accuracy particularly with large datasets [18]. It uses decision trees that grow efficiently by minimizing memory usage and optimizing training time. Table 1 shows the result of implementing several models, including the LightGBM model using the breast-cancer.csv dataset [68].

Table 1: Performance comparison of the different Anomaly Detection Algorithms on the Breast cancer dataset

Classifier	Accuracy	F1 Score	ROC-AUC	Sensitivity	Specificity
CatBoost	0.9737	0.9630	0.9997	0.9286	1.0000
LightGBM	0.9912	0.9880	0.9990	0.9762	1.0000
Stacking Ensemble	0.9649	0.9500	0.9977	0.9048	1.0000
Logistic Regression	0.9649	0.9512	0.9960	0.9286	0.9861
XGBoost	0.9737	0.9630	0.9950	0.9286	1.0000
SVM	0.9649	0.9512	0.9947	0.9286	0.9861
DNN	0.9737	0.9630	0.9947	0.9286	1.0000
Random Forest	0.9737	0.9630	0.9931	0.9286	1.0000
CNN-LSTM	0.9474	0.9286	0.9901	0.9286	0.9583
Naive Bayes	0.9211	0.8889	0.9894	0.8571	0.9583
KNN	0.9561	0.9383	0.9828	0.9048	0.9861
Decision Tree	0.9211	0.8916	0.9127	0.8810	0.9444

Model valuations:

Future research in tumour detection and classification should prioritise more sensitive and precise diagnostic tools and the standardisation of diagnostic criteria to improve accuracy and consistency.

A major direction is the development of advanced imaging techniques —such as contrast-enhanced ultrasound and PET/MRI—that can significantly enhance both sensitivity (early and accurate detection of true cancers) and specificity (correctly identifying benign lesions)[73]. At the same time, machine learning and AI will play an increasingly central role by enabling algorithms that more reliably localise tumours and differentiate malignant from benign cases, directly improving diagnostic performance.

Sensitivity and specificity remain essential metrics in this context.

- Sensitivity: (TP / (TP + FN)) reflects a model's ability to detect true cancer cases and minimise missed diagnoses—critical in early cancer screening.
- Specificity: (TN / (TN + FP)) indicates how well a system avoids false alarms and unnecessary interventions [76].

Optimising the sensitivity–specificity trade-off is essential since accuracy alone can be deceptive, particularly with imbalanced datasets. This enables researchers to create diagnostic instruments that are suitable for therapeutic requirements.

The proposed system, utilising the LightGM model, was rigorously evaluated:

The LightGBM model was compared against eleven state-of-the-art methods (CatBoost, LightGBM, Stacking Ensemble, Logistic Regression, XGBoost, SVM, DNN, Random Forest, Naive Bayes,KNN and Decision Tree) . It achieved the highest accuracy, F1-Score, perfect specificity, very high sensitivity and nearly perfect ROC-AUC compared to other models. LightGBM is the most balanced and robust [76].

LightGBM Performance: The LightGBM model consistently showed one of the models with the lowest error values across all metrics when compared to other models. LightGBM model slightly outperformance or is at par with some machine learning models, as depicted in the comparison below (Figure 1).

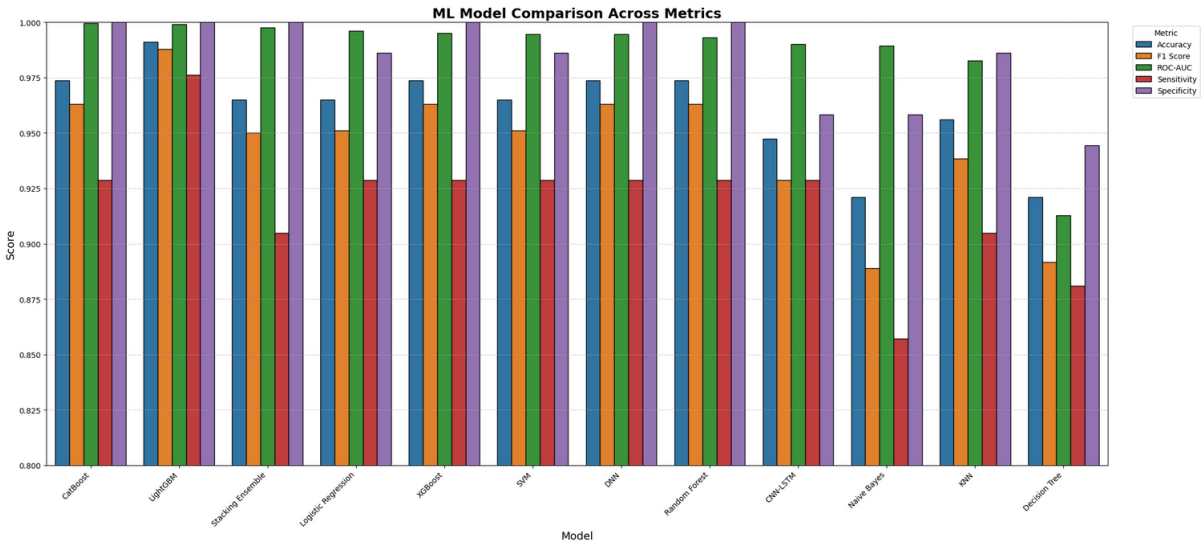

Figure 1: Bar chart comparing the accuracy of CatBoost, LigtGBM, Stacking Ensemble,Logistic Regression, XGBoost,SVM,DNN,Random Forest, CNN-LSTM, Naive Bayes, KNN and Decision Tree.

LightGBM Model Performance: The study introduces LightGBM model and compares it to CatBoost, CNN-LSTM, Stacking Ensemble, Logistic Regression, XGBoost, SVM, DNN, Random Forest, Naive Bayes, KNN and Decision Tree models. The LightGBM model showed superior prediction performance with lower error metrics on the same dataset. This is attributed to the model's responsiveness to the data, which gives it better adaptability to varied data distributions [76].

4. Results and Discussion

The evaluation of the anomaly detection system using three metrics: True Negative Rate (Specificity), Recall (Sensitivity), and the F1-Score. Precision measures the system's ability to correctly identify anomalies without mistaking normal points for anomalies, while Recall measures its ability to detect anomalies without missing any. The F1-score combines both Precision and Recall to provide a comprehensive measure of overall performance.

The research paper presents several key findings across three main experiments:

- Comparison with Other Methods: The LightGBM model was compared against eleven other state-of-the-art methods, including CatBoost, CNN-LSTM, Stacking Ensemble, Logistic Regression, XGBoost, SVM, DNN, Random Forest, Naive Bayes, KNN and Decision Tree models. The LightGBM method outperformed all others in recall and F1-score, demonstrating a stronger ability to detect anomalies in given biomedical data.
- **LightGBM Model Performance:** The study introduces a LightGBM model and compares it to CatBoost, Stacking Ensemble,Logistic Regression,XGBoost, SVM,DNN, Random Forest, CNN-LSTM, Naive Bayes, KNN,and Decision Tree. The LightGBM model showed superior prediction performance with lower error metrics on the same dataset. This is attributed to the model responsive to the data, giving it better adaptability for varied data distributions.

5. Conclusion

Breast cancer remains a major global health burden, driven by substantial biological heterogeneity and rising incidence worldwide. This study bridged clinical challenges with the capabilities of Artificial Intelligence (AI) and Deep Learning (DL), evaluating their effectiveness for breast cancer anomaly detection [81]. Using a hybrid anomaly detection framework, we compared twelve ML and DL models on the breast-cancer.csv dataset. LightGBM achieved the strongest overall performance, with an accuracy of 0.9912, F1- Score of 0.9880, Specificity of 1.0000, Sensitivity of 0.9762, ROC-AUC of 0.9990. LightGBM outperformed both traditional Machine learning algorithms and deep neural architectures, demonstrating exceptional balance between sensitivity and specificity, qualities essential for reliable clinical diagnosis.

These findings highlight the promise of advanced ensemble methods in developing high-precision diagnostic tools. Nevertheless, translating such models into real-world practice requires addressing key challenges of explainability, interpretability, unbiased data, ethical consideration and the patient's privacy to enhancing interpretation of patients' results and to support clinician trust and regulatory acceptance. Improvement dataset representation and standardization to reduce bias and bolster generalizability should be rigorous pursued. Expanding research beyond conventional DL (e.g., CNNs) to include GNNs, reinforcement learning, and generative models should be pursued in future research efforts [60]. Future progress will depend on continuous benchmarking across diverse clinical datasets, stronger ethical frameworks, and rigorous validation to ensure AI-driven systems support equitable, personalized breast cancer care.

Conflict of Interest

The authors declare no conflict of interest.

References

[1] Z. Mirza, M. S. Ansari, M. S. Iqbal, N. Ahmad, N. Alganmi, H. Banjar, M. H. Al-Qahtani, and S. Karim, "Identification of novel diagnostic and prognostic gene signature biomarkers for breast cancer using artificial intelligence and machine learning assisted transcriptomics analysis," *Cancers*, vol. 15, Art. no. 3237, 2023, doi: 10.3390/cancers15123237.

[2] C. Katsura, I. Ogunmwonyi, H. K. N. Kankam, and S. Saha, "Breast cancer: Presentation, investigation and management," *British Journal of Hospital Medicine*, 2022, doi: 10.12968/hmed.2021.0459.

[3] N. H. Singh, U. Köse, and S. P. Gochhayat, Eds., *Deep Learning in Biomedical Signal and Medical Imaging*. Boca Raton, FL, USA: CRC Press, Taylor & Francis, 2023.

[4] H. Xiao, Y. Zou, J. Wang, and S. Wan, "A review of artificial intelligence-based protein subcellular localization," *Biomolecules*, vol. 14, Art. no. 409, 2024, doi: 10.3390/biom14040409.

[5] J. Wang, Z. Zhang, and Y. Wang, "Utilizing feature selection techniques for AI-driven tumor subtype classification," *Biomolecules*, vol. 15, Art. no. 81, 2025, doi: 10.3390/biom15010081.

[6] J. U. Kazi, *Python Essentials for Biomedical Data Analysis*. Cham, Switzerland: Springer Nature, 2025, doi: 10.1007/978-3-031-85600-6.

[7] X. Luo, J. Y. Chen, M. Ataei, and A. Lee, "Microfluidic compartmentalization platforms for single-cell analysis," *Biosensors*, vol. 12, Art. no. 58, 2022, doi: 10.3390/bios12020058.

[8] C. Hayford *et al*., "Microfluidics-free single-cell genomics with templated emulsification," *Nature Biotechnology*, vol. 41, pp. 1557–1566, Nov. 2023, doi: 10.1038/s41587-023-01685-z.

[9] W. M. Zhou *et al*., "Microfluidics applications for high-throughput single-cell sequencing," *Journal of Nanobiotechnology*, vol. 19, Art. no. 312, 2021, doi: 10.1186/s12951-021-01045-6.

[10] R. Nema, A. Kumar, and D. K. Saini, Eds., *Advances in Cancer Detection, Prediction, and Prognosis Using Artificial Intelligence and Machine Learning*. Singapore: Springer Nature, 2025, doi: 10.1007/978-981-96-9346-7.

[11] M. Ossandon, B. Prickril, and A. Rasooly, Eds., *Cancer Detection and Diagnosis: A Handbook of Emerging Technologies*. Boca Raton, FL, USA: CRC Press, 2025, doi: 10.1201/9781003449942.

[12] T. T. Ogunjobi *et al*., "Bioinformatics applications in chronic diseases," *Medinformatics*, vol. 1, no. 1, pp. 1–18, 2024, doi: 10.47852/bonviewMEDIN42022335.

[13] J. Lötsch, D. Kringel, and A. Ultsch, "Explainable artificial intelligence in biomedicine," *BioMedInformatics*, vol. 2, pp. 1–17, 2022, doi: 10.3390/biomedinformatics2010001.

[14] T. Hulsen, "Literature analysis of artificial intelligence in biomedicine," *Annals of Translational Medicine*, vol. 10, no. 23, Art. no. 1284, 2022, doi: 10.21037/atm-2022-50.

[15] M. Liu, G. Srivastava, J. Ramanujam, and M. Brylinski, "SynerGNet: A graph neural network model to predict anticancer drug synergy," *Biomolecules*, vol. 14, Art. no. 253, 2024, doi: 10.3390/biom14030253.

[16] N. Q. K. Le, "Redefining biomedicine: Artificial intelligence at the forefront of discovery," *Biomolecules*, vol. 14, Art. no. 1597, 2024, doi: 10.3390/biom14121597.

[17] X. Wang, L. Yang, and R. Wang, "mRCat: A CatBoost predictor for mRNA subcellular localization," *Biomolecules*, vol. 14, Art. no. 767, 2024, doi: 10.3390/biom14070767.

[18] Z. Zhang, R. Zhang, K. Xiao, and X. Sun, "G4Beacon: An in vivo G4 prediction method," *Biomolecules*, vol. 13, Art. no. 292, 2023, doi: 10.3390/biom13020292.

[19] F. Dilnawaz and A. K. Behura, *Artificial Intelligence-Based Cancer Nanomedicine*. Sharjah, UAE: Bentham Science, 2022, doi: 10.2174/9789815050561122010l.

[20] A. Mitsala *et al*., "Artificial intelligence in colorectal cancer screening, diagnosis and treatment," *Current Oncology*, vol. 28, no. 3, pp. 1581–1607, 2021, doi: 10.3390/curroncol28030149.

[21] I. N. Weerarathna, A. R. Kamble, and A. Luharia, "Artificial intelligence applications for biomedical cancer research," *Cureus*, vol. 15, no. 11, Art. no. e48307, 2023, doi: 10.7759/cureus.48307.

[22] Z. Dlamini *et al*., "Artificial intelligence and big data in cancer and precision oncology," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 2300–2311, 2020, doi: 10.1016/j.csbj.2020.08.019.

[23] D. Zheng, X. He, and J. Jing, "Overview of artificial intelligence in breast cancer medical imaging," *Journal of Clinical Medicine*, vol. 12, Art. no. 419, 2023, doi: 10.3390/jcm12020419.

[24] S. B. Johnson *et al*., "Using ChatGPT to evaluate cancer myths and misconceptions," *JNCI Cancer Spectrum*, vol. 7, no. 2, Art. no. pkad015, 2023, doi: 10.1093/jncics/pkad015.

[25] Z. H. Chen *et al*., "Artificial intelligence for assisting cancer diagnosis and treatment in the era of precision medicine," *Cancer Communications*, vol. 41, no. 11, pp. 1100–1115, 2021, doi: 10.1002/cac2.12215.

[26] J. Liao *et al*., "Artificial intelligence assists precision medicine in cancer treatment," *Frontiers in Oncology*, vol. 12, Art. no. 998222, 2023, doi: 10.3389/fonc.2022.998222.

[27] M. J. Iqbal *et al*., "Clinical applications of artificial intelligence and machine learning in cancer diagnosis," *Cancer Cell International*, vol. 21, Art. no. 270, 2021, doi: 10.1186/s12935-021-01981-1.

[28] A. M. Sebastian and D. Peter, "Artificial intelligence in cancer research: Trends, challenges and future directions," *Life*, vol. 12, Art. no. 1991, 2022, doi: 10.3390/life12121991.

[29] D. M. Koh *et al*., "Artificial intelligence and machine learning in cancer imaging," *Communications Medicine*, 2022, doi: 10.1038/s43856-022-00199-0.

[30] B. Hunter, S. Hindocha, and R. W. Lee, "The role of artificial intelligence in early cancer diagnosis," *Cancers*, vol. 14, Art. no. 1524, 2022, doi: 10.3390/cancers14061524.

[31] Y. V. Pathak, S. Saikia, S. Pathak, J. Patel, and B. G. Prajapati, Eds., *Artificial Intelligence in Bioinformatics and Chemoinformatics*. Boca Raton, FL, USA: CRC Press, 2024.

[32] P. Gentile, "Breast cancer therapy: The potential role of mesenchymal stem cells," *Biomedicines*, vol. 10, Art. no. 1179, 2022, doi: 10.3390/biomedicines10051179.

[33] J. Dombi and O. Csiszár, *Explainable Neural Networks Based on Fuzzy Logic and Multi-Criteria Decision Tools*. Cham, Switzerland: Springer Nature, 2021, doi: 10.1007/978-3-030-72280-7.

[34] S. Wang *et al*., "A review of 3D printing technology in pharmaceutics," *Pharmaceutics*, vol. 15, Art. no. 416, 2023, doi: 10.3390/pharmaceutics15020416.

[35] S. M. Badr-Eldin *et al*., "Three-dimensional in vitro cell culture models for efficient drug discovery," *Pharmaceuticals*, vol. 15, Art. no. 926, 2022, doi: 10.3390/ph15080926.

[36] T. Antao, *Bioinformatics with Python Cookbook*, 3rd ed. Birmingham, U.K.: Packt Publishing, 2022.

[37] M. Domb, S. Joshi, and A. Khan, "Anomaly detection in IoT," in *Artificial Intelligence*, IntechOpen, 2024, doi: 10.5772/intechopen.111944.

[38] D. Tarin, *Understanding Cancer: The Molecular Mechanisms, Biology, Pathology and Clinical Implications*. Cham, Switzerland: Springer Nature, 2023, doi: 10.1007/978-3-030-97393-3.

[39] Z. Zhang *et al*., "Protein language models learn evolutionary statistics of interacting sequence motifs," *Proceedings of the National Academy of Sciences*, vol. 121, no. 45, Art. no. e2406285121, 2024, doi: 10.1073/pnas.2406285121.

[40] C. Wang *et al*., "Microfluidic biochips for single-cell analysis of multiomics," *Advanced Science*, vol. 11, no. 28, Art. no. e2401263, 2024, doi: 10.1002/advs.202401263.

[41] Y. You *et al*., "Artificial intelligence in cancer target identification and drug discovery," *Signal Transduction and Targeted Therapy*, vol. 7, Art. no. 156, 2022, doi: 10.1038/s41392-022-00994-0.

[42] A. Saini *et al*., "Cancer causes and treatments," *International Journal of Pharmaceutical Sciences and Research*, vol. 11, no. 7, pp. 3121–3134, 2020, doi: 10.13040/IJPSR.0975-8232.11(7).3121-34.

[43] K. P. T. Kathryn and S. E. H. Cokenakes, "Breast cancer treatment," *American Family Physician*, vol. 104, no. 2, Aug. 2021.

[44] P. K. Das, H. K. Tripathy, and S. A. M. Yusof, Eds., *Privacy and Security Issues in Big Data*. Singapore: Springer Nature, 2021, doi: 10.1007/978-981-16-1007-3.

[45] A. F. M. Gavriilidou *et al*., "High-throughput native mass spectrometry screening in drug discovery," *Frontiers in Molecular Biosciences*, vol. 9, Art. no. 837901, 2022, doi: 10.3389/fmolb.2022.837901.

[46] J. Baker-Brunnbauer, *Trustworthy Artificial Intelligence Implementation*. Cham, Switzerland: Springer, 2023, doi: 10.1007/978-3-031-18275-4.

[47] N. J. Ayon, "High-throughput screening for antibacterial drug discovery," *Metabolites*, vol. 13, Art. no. 625, 2023, doi: 10.3390/metabo13050625.

[48] F. A. Batarseh and L. J. Freeman, *AI Assurance: Towards Trustworthy, Explainable, Safe, and Ethical AI*. London, U.K.: Academic Press, Elsevier, 2023.

[49] B. Ammanath, *Trustworthy AI: A Business Guide*. Hoboken, NJ, USA: Wiley, 2022.

[50] A. Dingli and D. Farrugia, *Neuro-Symbolic AI*. Birmingham, U.K.: Packt Publishing, 2023.

[51] L. Alzubaidi *et al*., "Towards risk-free trustworthy artificial intelligence," *International Journal of Intelligent Systems*, vol. 2023, Art. no. 4459198, 2023, doi: 10.1155/2023/4459198.

[52] R. W. McGee, "Using Chinese herbal medicine to treat cancer patients," *Biomedical Journal of Scientific & Technical Research*, vol. 56, no. 5, 2024.

[53] D. Chen *et al*., "Integrated machine learning and bioinformatics analyses for cancer prognosis," *International Journal of Biological Sciences*, vol. 18, no. 1, pp. 360–373, 2022, doi: 10.7150/ijbs.66913.

[54] G. Liang *et al*., "The emerging roles of artificial intelligence in cancer drug development," *Biomedicine & Pharmacotherapy*, vol. 128, Art. no. 110255, 2020, doi: 10.1016/j.biopha.2020.110255.

[55] G. Dileep and S. G. Gianchandani Gyani, "Artificial intelligence in breast cancer screening," *Cureus*, vol. 14, no. 10, Art. no. e30318, 2022, doi: 10.7759/cureus.30318.

[56] X. Hou *et al*., "Artificial intelligence in cervical cancer screening," *Frontiers in Oncology*, vol. 12, Art. no. 851367, 2022, doi: 10.3389/fonc.2022.851367.

[57] F. Chollet, *Deep Learning with Python*, 2nd ed. Shelter Island, NY, USA: Manning Publications, 2021.

[58] W. Liu *et al*., *Graph Neural Network Methods and Applications in Scene Understanding*. Singapore: Springer Nature, 2024, doi: 10.1007/978-981-97-9933-6.

[59] L. Gianfagna and A. Di Cecco, *Explainable AI with Python*. Cham, Switzerland: Springer Nature, 2021, doi: 10.1007/978-3-030-68640-6.

[60] A. Munir, J. Kong, and M. A. Qureshi, *Accelerators for Convolutional Neural Networks*. Hoboken, NJ, USA: Wiley, 2024.

[61] S. K. Niazi and Z. Mariam, "Computer-aided drug design and drug discovery," *Pharmaceuticals*, vol. 17, Art. no. 22, 2024, doi: 10.3390/ph17010022.

[62] A. K. N. Neelam, "Advancing drug discovery through computer-aided design," *Discover Pharmaceutical Sciences*, vol. 1, Art. no. 8, 2025, doi: 10.1007/s44395-025-00008-2.

[63] A. B. Gurung *et al*., "An updated review of computer-aided drug design," *BioMed Research International*, vol. 2021, Art. no. 8853056, 2021, doi: 10.1155/2021/8853056.

[64] L. K. Vora *et al*., "Artificial intelligence in pharmaceutical technology," *Pharmaceutics*, vol. 15, Art. no. 1916, 2023, doi: 10.3390/pharmaceutics15071916.

[65] S. Nasim *et al*., "A novel approach for PCOS prediction using machine learning," *IEEE Access*, vol. 10, pp. 97610–97624, 2022, doi: 10.1109/ACCESS.2022.3205587.

[66] N. Goel and R. Kumar Yadav, Eds., *Internet of Things Enabled Machine Learning for Biomedical Applications*. Boca Raton, FL, USA: CRC Press, 2025.

[67] J. Zyla *et al*., "Combining radiomics and metabolomics for lung cancer diagnosis," *Biomolecules*, vol. 14, Art. no. 44, 2024, doi: 10.3390/biom14010044.

[68] S. M. Khade and R. G. Mishra, Eds., *Future of AI in Biomedicine and Biotechnology*. Hershey, PA, USA: IGI Global, 2024.

[69] S. K. Jha *et al*., Eds., *Computational Advances in Bio and Medical Sciences*. Cham, Switzerland: Springer Nature, 2021, doi: 10.1007/978-3-030-79290-9.

[70] H. S. Madhusudhan *et al*., Eds., *Artificial Intelligence and Cloud Computing Applications in Biomedical Engineering*. Boca Raton, FL, USA: CRC Press, 2025.

[71] C. Cerchia and A. Lavecchia, "New avenues in AI-assisted drug discovery," *Drug Discovery Today*, vol. 28, no. 4, Apr. 2023.

[72] A. S. Walker and J. Clardy, "A machine learning bioinformatics method to predict biological activity," *Journal of Chemical Information and Modeling*, doi: 10.1021/acs.jcim.0c01304.

[73] S. Srivastava *et al*., *Bio-Inspired Optimization for Medical Data Mining*. Beverly, MA, USA: Scrivener Publishing, 2024.

[74] W. Xie *et al*., "Transformer-based multimodal data fusion for COPD classification," *Biomolecules*, vol. 13, Art. no. 1391, 2023, doi: 10.3390/biom13091391.

[75] K. Athanasopoulou *et al*., "Artificial intelligence: The milestone in modern biomedical research," *BioMedInformatics*, vol. 2, pp. 727–744, 2022, doi: 10.3390/biomedinformatics2040049.

[76] A. R. Khan and T. Saba, Eds., *Explainable Artificial Intelligence in Medical Imaging*. Boca Raton, FL, USA: Auerbach Publications, 2025.

[77] S. P. Yadav, S. Yadav, and V. H. C. de Albuquerque, Eds., *Advances in Fuzzy-Based Internet of Medical Things*. Beverly, MA, USA: Scrivener Publishing, 2024.

[78] A. Gogoi and N. Mazumder, Eds., *Biological and Medical Physics, Biomedical Engineering*. Singapore: Springer Nature, 2024, doi: 10.1007/978-981-97-5345-1.

[79] A. Mukhopadhyay *et al*., *Multiobjective Optimization Algorithms for Bioinformatics*. Singapore: Springer Nature, 2024, doi: 10.1007/978-981-97-1631-9.

[80] L. Wang, "Mammography with deep learning for breast cancer detection," *Frontiers in Oncology*, vol. 14, Art. no. 1281922, 2024, doi: 10.3389/fonc.2024.1281922.

[81] R. Buyya *et al*., *Security and Privacy Issues in Internet of Medical Things*. Cambridge, MA, USA: Elsevier Academic Press, 2023.

[82] M. Torrente *et al*., "An AI-based tool for prognosis in cancer patients," *Cancers*, vol. 14, Art. no. 4041, 2022, doi: 10.3390/cancers14164041.

[83] W. Y. Lee *et al*., "Machine learning for recommending herbal formulae," *Biomolecules*, vol. 12, Art. no. 1604, 2022, doi: 10.3390/biom12111604.

[84] X. Wang, L. Yang, and R. Wang, "DRpred: A deep learning-based predictor for multi-label mRNA subcellular localization," *Biomolecules*, vol. 14, Art. no. 1067, 2024, doi: 10.3390/biom14091067.

[85] S. N. Shivhare and N. Kumar, "Brain tumor detection using manifold ranking in FLAIR MRI," in *Proc. 2019 Int. Conf. Emerging Trends in Information Technology (ICETIT)*, Lecture Notes in Electrical Engineering, vol. 605, P. Singh, B. Panigrahi, N. Suryadevara, S. Sharma, and A. Singh, Eds. Cham, Switzerland: Springer, 2020, pp. 271–279, doi: 10.1007/978-3-030-30577-2_25.