OPEN ACCESS JOURNALS

Indiana Journal of Multidisciplinary Research

Abbriviate Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.12667327

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Research Article

Volume-04|Issue-03|2024

Empowering Early Cancer Detection Through Machine Learning and Generative Models

Srivishnu M V*[1], Shreyas S[2], S Akash[3], Srikanth A[4], Mallanagouda Patil[5]

[1,2,4,3,5]Department of Computer Science and Engineering, RV Institute of Technology and Management, Bengaluru, India.

Article History
Received: 20.05.2024
Accepted: 05.06.2024
Published: 30.06.2024

Citation
Srivishnu, M. V., Shreyas, S. Akash, S., Srikanth, A. & Patil, M. (2024). Empowering Early Cancer Detection Through Machine Learning and Generative Models. *Indiana Journal of Multidisciplinary Research*, *4*(3), 43-48.

Abstract: Abstract- The project titled "Empowering Early Cancer Detection through Machine Learning and Generative Models" aims to revolutionize cancer diagnostics by leveraging advanced machine learning techniques and generative models. This research endeavours to develop sophisticated system capable of detecting early signs of cancer through the analysis of diverse medical data sets. By harnessing the power of machine learning algorithms, the project seeks to enhance the accuracy and efficiency of cancer diagnosis, enabling timely intervention and improving patient outcomes. In this study, we propose a novel approach for breast cancer detection leveraging machine learning models, specifically XGBoost, trained on synthetic data generated by a Generative Adversarial Network (GAN). Our results demonstrate a significant improvement in classification accuracy, with XGBoost achieving 95% accuracy prior to GAN training, and 99.04% accuracy thereafter. This enhancement underscores the efficacy of GAN-generated synthetic data in augmenting the training set, facilitating better model generalization and performance. GANs contribute to improved accuracy by generating diverse and representative data samples, thereby addressing class imbalance and enhancing the model's ability to capture complex patterns inherent in breast cancer datasets.
Keywords: Generative Models, Generative Adversial Networks, Synthetic data, Data augmentation.

INTRODUCTION

Global Cancer Challenge: Addressing the increasing cancer burden globally, our project focuses on innovative machine learning solutions for early detection. Data-Driven Precision: Employing advanced algorithms, our goal is to analyse extensive datasets, accurately identifying early-stage cancer patterns. Generative AI Augmentation: Integrating generative AI, we enhance our model's robustness by creating synthetic data, improving its generalization capabilities. Tailored Risk Assessment: Beyond traditional methods, our approach includes personalized risk stratification, providing nuanced predictions tailored to individual patient profiles. Ethical AI in Healthcare: Emphasizing transparency and ethical considerations, our project ensures trustworthy and accountable machine learning applications in critical healthcare scenarios. The below section discusses the literature survey.

LITERATURE SURVEY

Several studies have incorporated state-of-the-art AI-based approaches for the early detection of breast cancer by examining the morphology of breast tumours obtained thermos medical For breast cancer diagnostic purposes, many AI researchers have considered the UCI Wisconsin Breast Cancer Diagnostic dataset as a benchmark datasetfor evaluating machine-learning-based researches. Notably, Lahour et al. incorporated Extreme Learning Machine, avariant of Artificial Neural Network, on the Wisconsin Breast Cancer Diagnostic Dataset to build a cloud computing-based architecture for remote diagnosis of breast cancer. Kumar et al. Studies such as [1] emphasize the application of machine learning techniques for early detection, leveraging features extracted from medical imaging and patient data. Additionally, investigations by Cruz and Wishart [2] underscore the pivotal role of machine learning in cancer prediction and prognosis, shedding light on its potential in improving patient outcomes. In [3] the Support Vector Machine (SVM) algorithm, as the best performing machine-learning algorithm for breast cancer diagnosis. In recent, to overcome the limitations of a single machine-learning algorithm and increase the accuracy in the diagnosis of breast cancer, machine-learning researchers have focused on the ensemble or hybridization of machine-learning algorithms. In [4] heterogeneous ensemble machine-learning model by stacking KNN (K Nearest Neighbour), SVM (Support Vector Machine), and DT (Decision Tree) for the prediction of breast cancer was proposed by Nanglia et al. In [5] toensure a robust and efficient classification of malignant and benign or recurrent and non-recurrent breast cancers, this study incorporates a state-of-the-art attention-based interpretable deep architecture and compared the performance with other state-of-the-art machine learning and deep-learning algorithms. Moreover, research efforts such as those by Hüsemann et al. [6] delve into the biological mechanisms underlying cancer progression, emphasizing the importance of early detection in mitigating systemic spread and improving treatment outcomes. Globally, initiatives like the GLOBOCAN project [7] contribute to estimating cancer burden and mortality, providing valuable data

***Corresponding Author:** Srivishnu M V

for informing public health policies and interventions. The below section explains the problem statement of the project.

Problem Statement

Despite significant advancements in medical technology, early cancer detection remains a critical challenge with profound implications for patient outcomes and healthcare systems worldwide. Current diagnostic methods often relyon invasive procedures and lack the sensitivity to detect cancer at its nascent stages, leading to delayed diagnoses and suboptimal treatment outcomes. To address these challenges, there is a pressing need to leverage cutting-edge technologies such as machine learning (ML) and generative models like generative adversarial networks (GANs) in conjunction with various classification algorithms. In this research we have come up with the machine learning models for detecting cancer cells and compared these models with accuracy. Next section we will discuss about the proposedapproach

Proposed Approach

Our project employs a machine learning-based approach for the early diagnosis of breast cancer, leveraging computational intelligence techniques to analyze diverse datasets encompassing clinical, genomic, and imaging data. The proposed framework begins with comprehensive data preprocessing, including feature extraction and normalization, to ensure optimal input for subsequent model training. We utilize state-of-the-art machine learning algorithms, such as support vector machines (SVM), decision trees, and XGBoost, to develop robust classifiers capable of accurately distinguishing between malignant and benign breast lesions. Additionally, we integrate data augmentation techniques, such as generative adversarial networks (GANs), to enhance the diversity and representativeness of our training data, thereby improving model generalization and performance. Furthermore, our approach emphasizes interpretability and transparency in model decision-making, enabling clinicians to gain insights into the underlying factors driving classification outcomes. We adopt a multidisciplinary approach, incorporating insights from epidemiology, oncology, and artificial intelligence research, to develop a holistic framework for early cancer detection. By harnessing the power of computational intelligence and interdisciplinary collaboration, our project aims to advance the field of oncology and contribute to the development of effective screening and diagnostic tools for improving patient outcomes and quality of life.

METHODOLOGY

Following are the sub modules in proposed work

Data Collection and Preprocessing: Gather diverse datasets comprising clinical, genomic, and imaging data related to breast cancer patients from reputable sources such as medical repositories, research databases, and healthcare institutions. Conduct comprehensive data preprocessing steps including cleaning, missing value imputation, feature selection, and normalization to ensure high-quality input data for subsequent analysis.

Feature Engineering: Employ advanced feature engineering techniques to extract relevant features from theraw data, leveraging domain knowledge and machine learning algorithms to identify discriminative features associated with breast cancer diagnosis.

Model Development and Training: Develop machine learning models using various algorithms such as support vector machines (SVM), decision trees, random forests, and gradient boosting methods like XGBoost, tailored to the specific characteristics of the dataset. Train the machine learning models on the preprocessed dataset using appropriate training techniques such as cross-validation to optimize model parameters and improve generalization performance.

Model Evaluation and Comparison: Evaluate the performance of the trained models using standard evaluation metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC) to assess their effectiveness in breast cancer diagnosis. Compare the performance of different machine learning models to identify the most suitable approach for early breast cancer detection based on the defined evaluation criteria.

Data Augmentation: Explore data augmentation techniques such as generative adversarial networks (GANs) or synthetic minority oversampling technique (SMOTE) to enhance the diversity and representativeness of the dataset, potentially improving model performance as shown in fig 1.

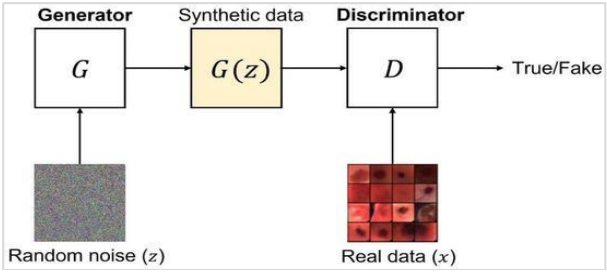

Figure 1

Interpretability Analysis: Conduct interpretability analysis to gain insights into the decision-making process of the developed models, elucidating the key features driving the classification outcomes and facilitating clinical interpretation.

Validation and Deployment: Validate the trained models using independent datasets or through clinical validation studies to ensure their reliability and effectiveness in real-world scenarios. Deploy the

***Corresponding Author:** Srivishnu M V

validated models as a web app or as decision support tools in clinical settings for assisting healthcare professionals in early breast cancer diagnosis.

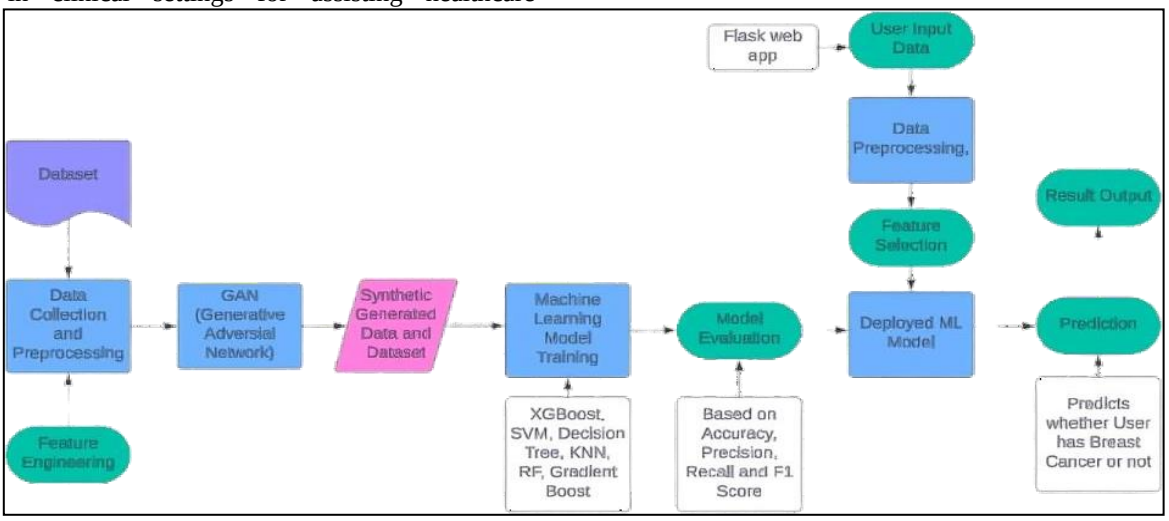

Figure 2: Flowchart of project

The above flowchart describes the architecture of the project which generates synthetic data using GAN and merges The generated data with the dataset and trains the machine learning model, in this case the XGBoost classifier to predict the cancer results. Next section discusses the mathematical model for the proposed approach.

MATHEMATICAL MODEL

To delineate the process in a computational view. Here we use the Euclidean distance based on the formula:

G(x) = T (Decompress (Compress(x))) Here, in Eq.(l G G) represents generated synthetic data of breast tumours, represents real data instance of breast tumours, T is the tabular variational autoencoder function that takes c as input and generates G (c). Moreover, the Compress function is the Encoder of the XGBoost model which learns latent distribution from real data and GANs use SoftMax and tanh on the output to generate a mix of discrete and continuouscolumns at the same time. Clinical data like breast cancer data is highly sensitive. Imbalance in the minority class labels and also categorical columns is a common issue in such sensitive medical data which can create severe modal collapse issues. To prevent such modal collapse issues, GAN-based framework with 10 data samples in each pac. Mathematically, F1-Score: It is delineated as a term that represents a trade-off between recall and precision. Mathematically, True Positive (TP): The number of correctly predicted positive cases (cancer cases).

True Negative (TN): The number of correctly predicted negative cases (non-cancer cases). False Positive (FP): The number of incorrectly predicted positive cases (non-cancer cases predicted as cancer). False Negative (FN): The number of incorrectly predicted negative cases (cancer cases predicted as non-cancer). Accuracy (ACC) is defined asthe ratio of correct predictions to the total number of predictions. Mathematically, it can be expressed as:

$$\text{Accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}} \quad (1)$$

Logistic Regression Function:

$$P(y=1|\mathbf{x};\mathbf{w}) = \frac{1}{1+e^{-(\mathbf{w}\cdot\mathbf{x}+b)}} \quad (2)$$

This formula represents the logistic regression function used for binary classification tasks, where x is the input feature vector, w are the weights, and b is the bias term as stated in equation 3)

Support Vector Machine (SVM) Decision Function:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}\cdot\mathbf{x}+b) \quad (3)$$

This formula represents the decision function of a support vector machine, where x is the input feature vector, w arethe weights, and b is the bias term as stated in equation 4)

XGBoost Objective Function:

$$\text{Objective} = \sum_{i=1}^{n} \text{loss}(\hat{y}_i, y_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (4)$$

This formula represents the objective function of the XGBoost algorithm, where θ represents the set of model parameters. n is the number of training instances. L (yi, y^i) is the loss function that measures the difference between the true label yi and the predicted label $y^{\wedge}i.\Omega(fk$ is the regularization term that penalizes the complexity of the model by adding up the scores of all the leaves in each tree fk. K is the number of trees in the ensembleich consists of the sum of the individual losses and a regularization term Ω(fk)

***Corresponding Author:** Srivishnu M V

Decision Tree Splitting Criterion (Gini Impurity):

$$G(t) = 1 - \sum_{i=1}^{c} p(i|t)^2 \quad (5)$$

This formula calculates the Gini impurity for a node t in a decision tree, where c is the number of classes and p(i/t) isthe probability of class i at node t

Area Under the ROC Curve (AUC):

$$\text{AUC} = \int_0^1 \text{TPR}(fpr)\, d(fpr) \quad (6)$$

This formula calculates the area under the ROC curve where TPR represents the True Positive Rate (Sensitivity) at a specific FPR threshold. FPR represents the False Positive Rate.

TPR can be derived as

$$TPR = \frac{TP}{TP + FN}00 \quad (7)$$

Where TP represents the True positives which is the number of correctly predicted positive instances. FN represents the False Negatives which is the number of incorrectly predicted negative instances.

FPR can be derived as

$$FPR = \frac{FP}{FP + TN} \quad (8)$$

Where FP represents the False positives which is the number of incorrectly predicted positive instances. TNrepresents the True negatives which is the number of correctly predicted negative instances.

Above mathematical model has been used in proposed approach to do the result analysis which is discussed nextsection.

RESULT ANALYSIS

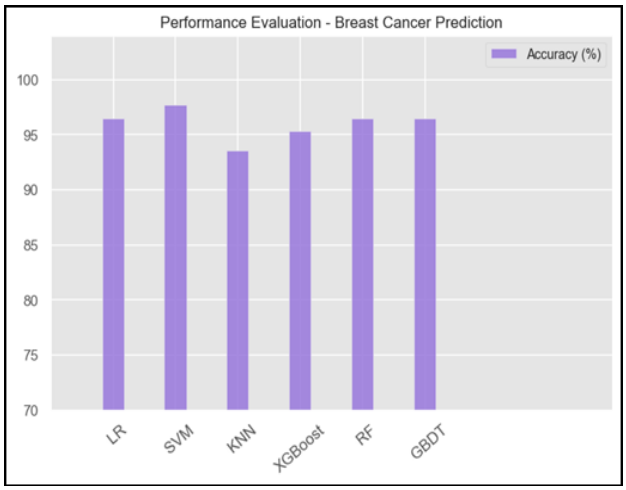

Figure 3: Before Gen AI

	Model	Score
2	SVM	97.66
0	Logistic Regression	96.49
3	Random Forest Classifier	96.49
4	Gradient Boosting Classifier	96.49
5	XgBoost	95.32
1	KNN	93.57

Figure 4: Scores of different models before Gen AI

The above figures visualizes and describes the comparison of the various classification models used in the project and listsout the Scores of these models. We can observe that SVM performs better with an accuracy of 97.66 % without being Trained on the synthetic data generated using the GAN model. The graph in Fig 3 compares the scores of the classification models used which helps identify and evaluate the model based on accuracy.

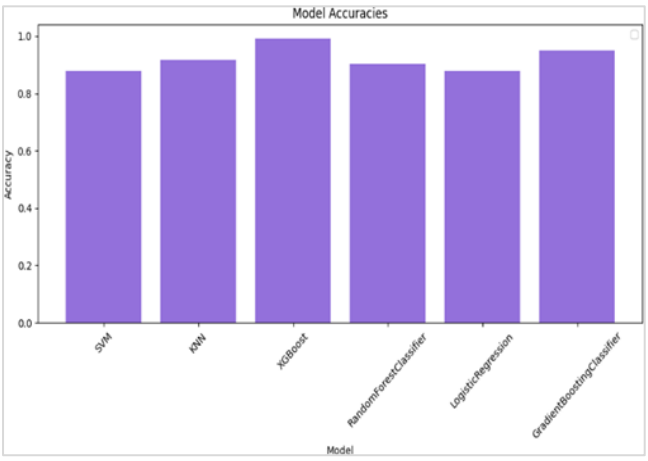

Figure 5: After Gen AI

	Model	Score
5	XgBoost	99.04
4	Gradient Boosting Classifier	91.40
1	KNN	91.08
3	Random Forest Classifier	87.26
2	SVM	85.35
0	Logistic Regression	84.71

Figure 6: Scores of different models after Gen AI

***Corresponding Author:** Srivishnu M V

The above figures visualizes and describes the comparision of the classification models after being trained on the data generated using the GAN. These models are now tested under high load of data and evaluated based on accuracy. We can observe that XGBoost classification model which is known for performing better with high loads of data performs exceedingly well with an accuracy of 99.04% as compared to 95.32% before the generation of synthetic data.

The graph in Fig 5 compares the scores of the classification models after being trained on the data generated by the GAN model. This graph enables us to identify and evaluate the best performing model in terms of accuracy.

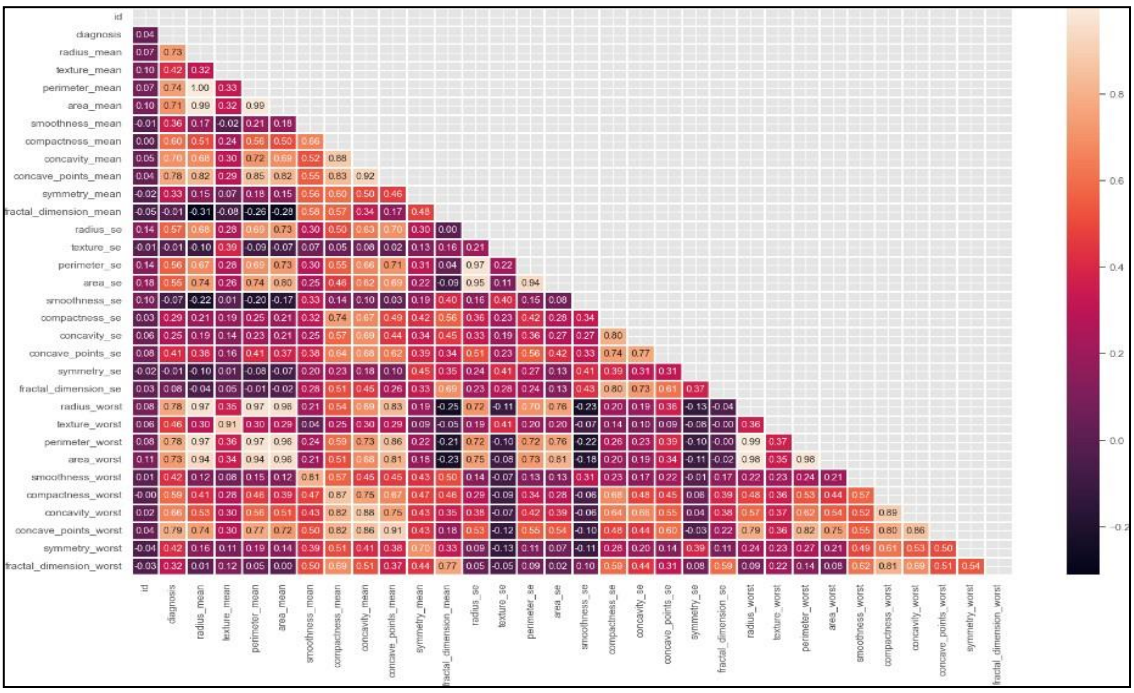

Figure 7: Heatmap

The heatmap in the figure represents the correlation between different features and their importance in predicting cancer diagnosis. It provides a color-coded matrix where each cell's color intensity indicates the strength of the correlation between corresponding features. By analyzing the heatmap, researchers can identify significant features that contribute to accurate cancer detection and prioritize them for further analysis or model development.

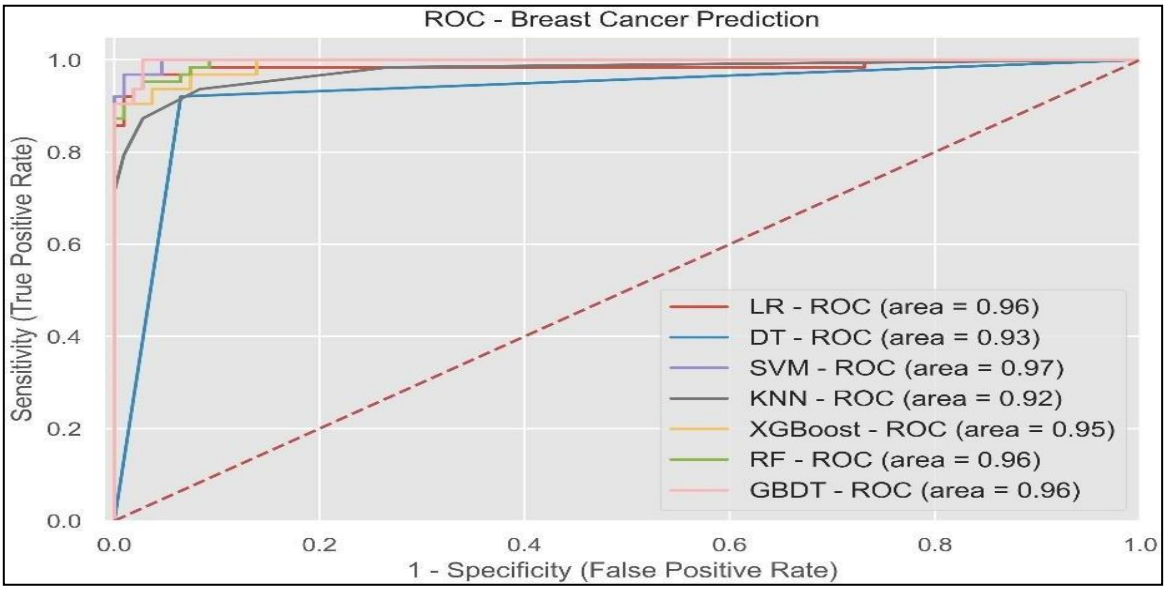

Figure 8: ROC – Breast Cancer Prediction

***Corresponding Author:** Srivishnu M V

The ROC (Receiver Operating Characteristic) illustrates the trade-off between sensitivity and specificity of a predictive model across various threshold values. It plots the true positive rate (sensitivity) against the false positive rate (1-specificity), showing how well the model distinguishes between positive and negative instances. A higher ROC curve indicates better performance of the model in accurately classifying cancer cases, with values closer to 1 representing ideal classification.

CONCLUSION

Cancer is one of the greatest threats to humankind in today's world. It takes away lives that are invaluable, creating hindrances to social and global development. In the era of industrial revolution 4.0, in which AI is the new oil, it is necessary to ameliorate cancer research with the integration of AI. The scarcity of available data should not stop researchers from developing high-quality automated AI models to fight cancers including breast cancer. This study has highlighted the possibility of generating breast cancer data synthetically to support high-quality breast cancer research. After the rigorous performance comparison against GAN-generated data This study experimented with several state- of-the-art machine learning and deep-learning classifiers trained on the synthetically generated data followed by performance evaluation with real data.The proposed integrated model is successfully able to generate high-quality synthetic data for the breast cancer domain, which can be utilized for building high-quality AI models for early diagnosis and proper prognosis of breast cancer. In the future, we will work with more challenging breast cancer datasets to improve the proposed integrated architecture from every aspect of breast cancer research to develop high-quality interpretable AI models to aid cancer research for the welfare of mankind. We can conclude that XGBoost classifier outperformed all other classification models with an astonishing accuracy of 99.04 after being trained by the synthetic data generated by the implemented GAN model.

REFERENCES

1. Raza, K. (2022). Computational intelligence in oncology: ML-based approach for early diagnosis of breast cancer. In *Computational Intelligence in Oncology* (pp. 285–306).
2. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Informatics*, *2*, 59–77.
3. Wild, S. H., Fischbacher, C. M., Brock, A., Griffiths, C., & Bhopal, R. (2006). Mortality from all cancers and lung, colorectal, breast, and prostate cancer by country of birth in England and Wales, 2001–2003. *British Journal of Cancer, 94*(7), 79–85.
4. Hippisley-Cox, J., et al. (2003). The electronic patient record in primary care: Regression or progression? A cross-sectional study. *Journal*, 39–43.
5. Sloggett, A., Young, H., & Grundy, E. (2007). The association of cancer survival with four socioeconomic indicators: A longitudinal study of the older population of England and Wales. *Journal*, 67–89.
6. Hüsemann, Y., Geigl, J. B., Schubert, F., Musiani, P., Meyer, M., Burghart, E., Forni, G., Eils, R., Fehm, T., Riethmüller, G., & Klein, C. A. (2008). Systematic detection of malignant cells in breast cancer. *Journal*, 58-68.
7. Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., Znaor, A., & Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International Journal of Cancer*, *144*(8), 1941-1953.
8. S., Yang, J., Fong, S., & Zhao, Q. (2021). Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges. *Cancer Letters*, *471*, 61-71.
9. Ginsburg, O., Yip, C. H., Brooks, A., Cabanes, A., Caleffi, M., Dunstan Yataco, J. A., Gyawali, B., McCormack, V., McLaughlin de Anderson, M., & Mehrotra, R. (2021). Breast cancer early detection: A phased approach to implementation. *Cancer*, *126*, 2379-2393.
10. Lan, K., Wang, D. T., Fong, S., Liu, L. S., Wong, K. K., & Dey, N. (2018). A survey of data mining and deep learning in bioinformatics. *Journal of Medical Systems*, *42*, 1-20.
11. Toğaçar, M., & Ergen, B. (2018). Deep learning approach for classification of breast cancer. In *2018 International Conference on Artificial Intelligence and Data Processing (IDAP)* (pp. 1-5).
12. Kumar S., Akshita Thapliyal S., Bhatt S., Negi N. (2021) Performance analysis of machine learning-based breast cancer detection algorithms Bajpai M.K., Kumar Singh K., Giakos G. (Eds.), Machine vision and augmented intelligence Theory and applications, Springer Singapore, Singapore, pp. 145-155.

***Corresponding Author:** Srivishnu M V