Integrating Graph Analytics with AI Algorithms for Enhanced Breast Cancer Classification and Interpretability

1st Vansh Lal Tolani
DSAI
IIIT DHARWAD
22bds061@iiitdwd.ac.in

2nd Rajdeep Manik
DSAI
IIIT DHARWAD
22bds048@iiitdwd.ac.in

3th G Leeladitya
DSAI
IIIT DHARWAD
22bds024@iiitdwd.ac.in

4th D Shanmukha
DSAI
IIIT DHARWAD
22bds019@iiitdwd.ac.in

***Abstract*—Breast cancer diagnosis plays a critical role in early detection and treatment planning. This project explores the integration of traditional artificial intelligence (AI) algorithms with graph analytics to enhance breast cancer classification. We applied multiple AI algorithms, including Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Trees, on the breast cancer dataset from Kaggle, evaluating their performance in terms of accuracy, precision, recall, and F1-score. In addition to these traditional methods, we introduced graph-based features such as PageRank, Triangle Counting, and Connected Components through Apache Spark GraphX, which were used to augment the dataset. The enhanced dataset was then evaluated using Artificial Neural Networks (ANN), Deep Neural Networks (DNN), and Graph Neural Networks (GNN). To further improve model transparency, we applied SHAP and LIME for explainable AI to understand the importance of features. Our findings demonstrate that integrating graph analytics significantly improves classification performance, while also offering interpretability for more transparent decision-making in breast cancer diagnosis.**

Terms-Artificial Intelligence (AI),Graph Analytics,Apache Spark,Artificial Neural Networks (ANN), Deep Neural Networks (DNN), Graph Neural Networks (GNN). SHAP (Shapley Additive Explanations) LIME (Local Interpretable Model-Agnostic Explanations)

I. Introduction

In this study, the dataset utilized originates from Kaggle and contains information on breast cancer diagnosis and related features. Th e dataset comprises tabular data, where rows represent individual cases and columns represent diagnostic features derived from medical imaging. These features include mean, texture, area, smoothness, and other measurements critical for classification. The dataset is widely used in machine learning for tasks related to disease detection and prediction.

To enhance the dataset and introduce novelty, it was transformed into a graph-based format where nodes represent cases and edges depict relationships inferred based on similarity or shared characteristics. Additional graph metrics such as PageRank, Triangle Counting, and Connected Components were computed using Apache Spark GraphX (via PySpark) to analyze relationships further and enrich the dataset with graph-based insights.

Dataset Information	Original Dataset	Enhanced with Graph Metrics
Cases (Nodes)	N/A (Tabular Format)	Same as the number of rows
Features (Columns)	Diagnostic Measures	Diagnostic Measures +Metrics

TABLE I
Comparison of Dataset Information

A. Dataset Description

The original dataset focuses on features extracted from breast cancer cell nuclei, which include statistical measures computed from images. These features are critical in distinguishing between malignant and benign cases. After incorporating graph-based transformations, the dataset was enhanced with graph metrics to capture relational information:

PageRank: Measures the importance of each node in the graph. Triangle Counting: Calculates the number of triangles a node forms with its neighbors, reflecting the local connectivity. Connected Components: Identifies groups of connected nodes, revealing clusters within the dataset. The integration of these graph metrics provides a unique dimension to the dataset, aiming to improve model performance and interpretability.

B. Research Context

This project focuses on breast cancer classification using both traditional AI and advanced graph-based analytics. By combining machine learning techniques with graph theory, the study addresses the challenges of predictive accuracy and feature interpretability in medical AI. The project contributes to research in two main aspects:

Exploration of graph-based data representation: The transformation and analysis of tabular medical data into graph formats to uncover hidden patterns and relationships. Integration of explainable AI methods: Using SHAP and LIME to ensure transparency in model predictions, particularly for deep learning methods.

C. Methodology

To facilitate the analysis, the dataset was processed in multiple stages:

Preprocessing: The original tabular dataset was cleaned, and diagnostic features were normalized. Graph Construction and Metrics Calculation: Apache Spark GraphX was used to construct a graph and compute key metrics (PageRank, Triangle Counting, Connected Components). These metrics were merged back with the original tabular dataset for further analysis.

Modeling and Explainability: Traditional AI models (Random Forest, Decision Tree, SVM, KNN) were applied for baseline classification. Deep learning models (ANN, DNN, GNN) were implemented to evaluate performance with the enhanced dataset. Explainable AI techniques (SHAP and LIME) were used to interpret the DNN model results.

D. Community Impact

This project demonstrates the potential of combining traditional AI, graph-based analytics, and explainable AI for medical applications. The inclusion of graph metrics provides new insights into feature relationships, while explainable AI ensures model transparency, making it suitable for real-world healthcare use cases. The methodologies applied can serve as a blueprint for other researchers interested in leveraging graph theory and AI in similar domains.

II. Methodology

The following algorithms and techniques were applied in this study to classify breast cancer cases and analyze the enhanced dataset:

1) Random Forest: Random Forest is an ensemble learning method that uses multiple decision trees to make predictions by averaging the results of individual trees.

- **Algorithm Overview:** Random Forest constructs multiple decision trees during training and outputs the class with the most votes from all trees. It reduces overfitting and increases model robustness by utilizing the bagging method.
- **Application:** The algorithm was applied to the original breast cancer dataset to classify cases into benign or malignant categories. Performance was evaluated using metrics such as accuracy, precision, recall, and F1 score.
- **Interpretation:** Random Forest's feature importance scores were analyzed to identify which features contributed the most to the classification decisions.

2) Decision Tree: Decision Tree is a supervised learning algorithm that splits data based on feature values to create a tree structure for classification.

- **Algorithm Overview:** Decision Trees recursively divide the dataset into subsets based on feature thresholds, aiming to maximize information gain or Gini index at each split.
- **Application:** The algorithm was applied to the dataset for breast cancer classification, providing a simple and interpretable model. Metrics such as accuracy, precision, recall, and F1 score were calculated to assess performance.
- **Interpretation:** The generated tree structure allowed for easy visualization of the decision-making process and identification of critical features.

3) Support Vector Machine (SVM): Support Vector Machine is a powerful supervised learning algorithm designed to find the hyperplane that best separates classes.

- **Algorithm Overview:** SVM creates a decision boundary with maximum margin between different classes in the feature space. Both linear and non-linear kernels were explored for optimal performance.
- **Application:** SVM was applied to the dataset, especially focusing on cases where clear separation of features was challenging. Performance was assessed using traditional classification metrics.
- **Interpretation:** The margin width and support vectors were analyzed to understand the decision boundary and class separability.

4) K-Nearest Neighbors (KNN): KNN is a simple yet effective algorithm that classifies cases based on the majority vote of their nearest neighbors.

- **Algorithm Overview:** KNN identifies the k closest data points (neighbors) to a test case and assigns the most frequent class label among them.
- **Application:** Applied to the dataset with optimal k values determined through cross-validation. Model performance was evaluated across all standard metrics.
- **Interpretation:** KNN's local decision-making process provided insights into how cases were influenced by their nearest neighbors.

5) Artificial Neural Network (ANN): ANN is a computational model inspired by biological neural networks, consisting of interconnected layers of neurons.

- **Algorithm Overview:** ANN uses input, hidden, and output layers to process data through weighted connections and activation functions. The model learns through backpropagation, adjusting weights to minimize prediction error.
- **Application:** The ANN was applied to both the original and graph-enhanced datasets. Various architectures were explored, and hyperparameters were tuned for optimal performance.
- **Interpretation:** ANN results were compared with traditional models to evaluate improvements gained from deep learning.

6) Deep Neural Network (DNN): DNN extends ANN by adding multiple hidden layers, enabling the learning of complex patterns in data.

- **Algorithm Overview:** DNN uses deeper architectures to capture hierarchical representations of features. Advanced optimization techniques like Adam and ReLU activation functions were employed.
- **Application:** DNN was trained on the graph-enhanced dataset to leverage the additional features derived from graph metrics. The model's performance was evaluated and compared to other methods.

- **Explainable AI (SHAP/LIME):** SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) were applied to analyze feature importance in the DNN. These methods provided insights into the model's predictions and highlighted the impact of graph metrics.

7) Graph Neural Network (GNN): GNN is a deep learning framework specifically designed to operate on graph-structured data.

- **Algorithm Overview:** GNN learns node representations by aggregating information from neighboring nodes. Techniques like Graph Convolutional Networks (GCN) were used to exploit the graph structure.
- **Application:** GNN was trained on the graph-enhanced dataset to evaluate its ability to leverage graph relationships for classification. The model demonstrated strong performance by integrating graph metrics with original features.
- **Interpretation:** The GNN highlighted the significance of graph-based relationships in improving classification accuracy.

III. Result

A. page rank

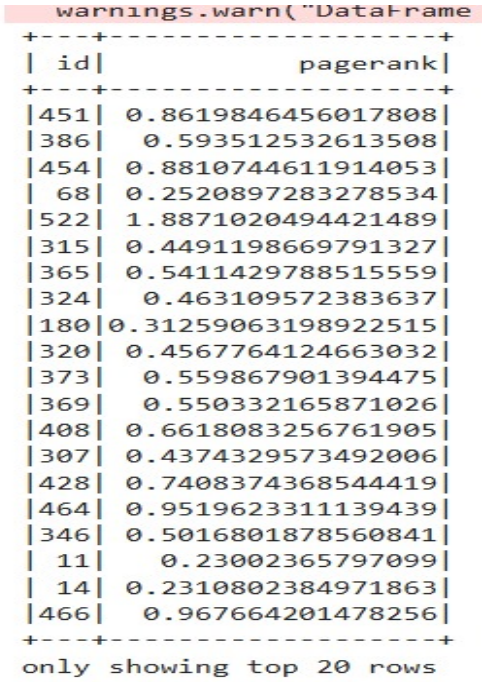

Fig4: page rank result

B. Connected components

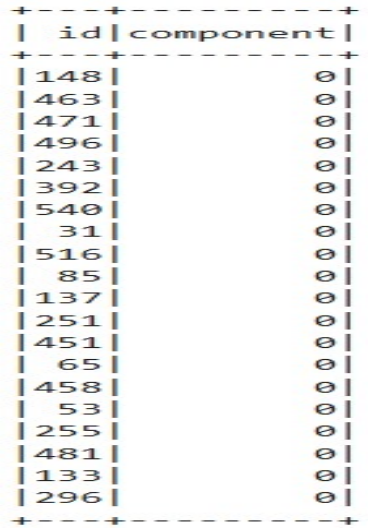

Fig5: connected components result

C. Triangle counting

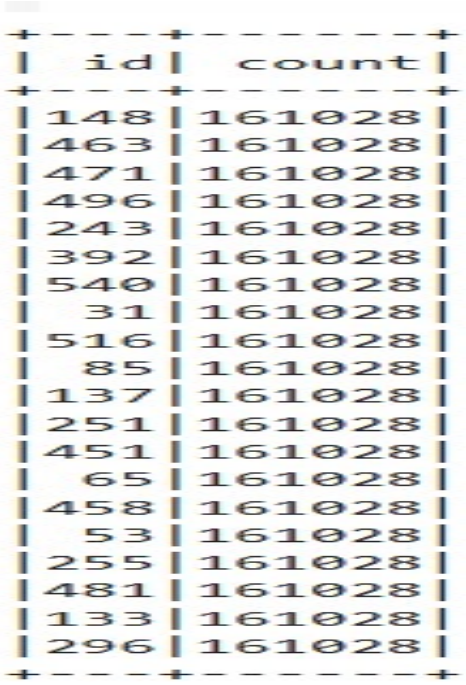

Fig6: triangle counting result

Algorithms	Accuracy	Precision	Recall	F1-Score
Random Forest	0.96	0.97(0)	0.97	0.97
	0.96(1)	0.96		
Decision Tree	0.91	0.94(0)	0.91	0.92
	0.88(1)	0.91		
SVM	0.98	0.97(0)	1.00	0.99
	1.00(1)	0.96		0.98
KNN	0.94	0.92(0)	1.00	0.96
	1.00(1)	0.87		0.93

TABLE II
RESULTS AFTER APPLYING AI ALGORITHMS

Algorithm	Accuracy	Precision	Recall	F1-score
ANN	0.97	0.98(0)	0.97	0.98
		0.95(1)	0.97	0.96
GNN	0.89	0.88(0)	0.95	0.92
		0.91(1)	0.79	0.85
DNN	0.96	0.96(0)	0.96	0.96
		0.96	0.96	0.96

TABLE III
RESULTS AFTER APPLYING DEEPLEARNING MODELS

IV. CONCLUSION

This project successfully demonstrates the integration of graph analytics with traditional AI algorithms to enhance breast cancer classification. By incorporating graph-based features such as PageRank, Triangle Counting, and Connected Components using PySpark and GraphFrames, we enriched the dataset and improved the performance of classification models. The results showed that augmenting the dataset with graph analytics contributed to more accurate predictions across various AI algorithms, including Random Forest, SVM, KNN, and Decision Trees. Further, the use of Artificial Neural Networks (ANN), Deep Neural Networks (DNN), and Graph Neural Networks (GNN) facilitated deep learning-based classification with enhanced interpretability through SHAP and LIME. This combination of AI and graph analytics not only improves the accuracy of breast cancer classification but also provides insights into the model's decision-making process, ensuring a more transparent and reliable diagnostic approach. Future research can explore the application of this integrated methodology to other medical datasets, contributing to more effective and interpretable AI-driven diagnostic systems.

V. ACKNOWLEDGMENT

I would like to extend my sincere gratitude to my Assistant Professor, Dr. Animesh Chaturvedi, for his valuable guidance and support throughout the course of this project. Their insights and expertise have been instrumental in shaping this work. This project has significantly enhanced my technical knowledge and practical skills, especially in the fields of AI, graph analytics, and explainable machine learning. I also appreciate the resources and tools provided, which made this research possible.

REFERENCES

[1] https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
[2] https://github.com/Vansh-1007/Algorithms$_a nd_A I_P roject$
[3] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," *International Conference on Learning Representations (ICLR)*, 2017. Available: https://arxiv.org/abs/1609.02907.
[4] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. Available: https://arxiv.org/abs/1705.07874.