IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

GraphX-Net: A Graph Neural Network-based Shapley Values for Predicting Breast Cancer Occurrence

ABDULLAH BASAAD[1], SHADI BASURRA[1], EDLIRA VAKAJ, MOHAMMED ALESKANDARANY[2], MOHAMMED M. ABDELSAMEA[3]

[1]School of Computing and Digital Technology, Birmingham City University, Birmingham B4 7XG, U.K. (e-mail: abdullah.basaad@mail.bcu.ac.uk, Shadi shadi.basurra@bcu.ac.uk, edlira.vakaj@bcu.ac.uk)
[2]School of Human Sciences, University of Derby, Derby DE22 3AW, U.K. (e-mail: m.aleskandarany@derby.ac.uk)
[3]Computer science Department, University of Exeter, North Park Road Exeter EX4 4QF, U.K. (e-mail: m.abdelsamea@exeter.ac.uk)

Corresponding author: Mohammed M. Abdelsamea (e-mail: m.abdelsamea@exeter.ac.uk).

ABSTRACT Breast cancer is a major health problem worldwide, and accurate prediction of its recurrence is crucial to early detection of recurrence and personalised treatment. In recent years, various AI techniques have been applied to predict cancer recurrence with increasingly high accuracy. Graph Neural Networks (GNNs) have emerged as powerful tools for analyzing structured data, including knowledge networks. In this study, we explore the application of GNN-based node classification to predict the occurrence of relapse status in breast cancer patients. We propose a novel model, namely GraphX-Net, a Shapley value-based graph neural network. GraphX-Net leverages a graph representation by analysing public breast cancer data, where nodes represent patients and edges capture relationships between them based on various factors such as tumour cellularity, histological subtype, hormone therapy, patient vital status, primary tumour laterality, type of breast surgery, and other clinicopathological parameters. Our model utilises graph convolutional layers to generate informative node embeddings. The model uniquely calculates node feature contributions using Shapley values, sets nodes' thresholds, and considers the total node neighboring effects to form the relationships between nodes. These innovations enable GraphX-Net to achieve state-of-the-art performance in predicting breast cancer recurrence, highlighting its potential as a powerful tool for improving patient outcomes.

INDEX TERMS Graph Neural Networks, XAI, Node Classification, Breast Cancer, Occurrence Prediction.

I. INTRODUCTION

THE use of node classification graphs to predict the occurrence of relapse status in breast cancer has emerged as a promising avenue in the field of cancer research. Relapse or recurrence of breast cancer after treatment continues to be a major concern for both patients and healthcare providers. It has long been reported that morbidity and mortality observed in patients with breast cancer primarily result from disease progression through local-regional or systemic dissemination, as well as the occurrence of recurrences post-treatment, or a combination of these factors [1]. Upon development of metastatic deposits outside the primary location, the prognosis of patients is dramatically worsened, and cure would be unlikely [2]. The ability to accurately predict the probability of relapse can greatly help in treatment planning, monitoring, and ultimately improving patient outcomes [3],[4]. Researchers have achieved notable advancements in the development of predictive models of relapse status in breast cancer through the effective use of node classification graphs [5]. One particularly interesting research study was explored the predictive potential of K-Banhatti and Zagreb type degree-based topological indices in quantitative structure–property relationship (QSPR) analysis for medications used to treat type-I and type-II diabetes. These indices were computed for 14 anti-diabetes drug molecules using edge and vertex partitioning techniques. By utilizing these topological indices, researchers developed QSPR regression models to predict the physicochemical properties of the drugs under study [6]. Another interesting study discussed the use of topological indices (TIs) to predict the physical and biological properties of drugs used to treat Alzheimer's disease. Degree-based TIs were generated using edge partitioning for drugs like Tacrine, Donepezil, and Rivastigmine. A QSPR model was developed using linear regression to predict characteristics such as boiling point, flash point, molar volume, molecular weight, complexity, and polarizability. The findings suggest that TIs can be valuable tools in drug discovery and design for Alzheimer's disease treatment [7].

This study focuses on a node classification graph representing the interconnectivity between nodes, where each node corresponds to an individual patient. This representation encapsulates the intricate relationships and interdependencies among various factors that contribute to relapse, encompassing clinical variables, histopathological features, and treatment regimens. Through the utilisation of machine learning algorithms on these graphs, researchers can discern patterns and features that possess predictive value for relapse occurrence. A notable advantage of employing node classification graphs lies in their capacity to accommodate the heterogeneous nature of breast cancer data [8]. Breast cancer is a multifaceted ailment influenced by diverse variables that interact with one another [9]. By representing these variables as features of the nodes and their relationships as edges, node classification graphs offer a comprehensive framework for modelling and analysing these intricate connections.

In recent years, several studies have been dedicated to improving the precision of predicting relapse recurrence in breast cancer through the use of various techniques. One such approach involves the integration of Multi-Omics Data, which encompasses the incorporation of multiple types of omics data, including genomics, transcriptomics, proteomics, and epigenomics [10]. By amalgamating information from various molecular levels, researchers endeavour to identify biomarkers and molecular signatures that are associated with relapse. This comprehensive integration facilitates the development of models that predict the occurrence of relapses. Another avenue for providing valuable information on tumour characteristics and improving relapse prediction is the incorporation of imaging and radiationomics. Using medical imaging techniques such as mammography, magnetic resonance imaging, or PET-CT scans, researchers can gain pertinent information on tumour properties [11]. Radiomic models, which involve the extraction of a multitude of features from the imaging data, enable the capture of subtle patterns and textures that are indicative of relapse. These extracted features are subsequently employed to train machine learning models, thereby enhancing the accuracy of relapse prediction. Furthermore, the inclusion of clinical and pathological information [12] and the analysis of longitudinal data are instrumental in elucidating the dynamic changes associated with relapse [13]. These additional factors contribute valuable information to advance our overall understanding of relapse recurrence in breast cancer.

This study aims to tackle the challenges associated with understanding the relational nature of patient data and determining the contribution of inputs within a neural network framework. To address these challenges, we propose a novel approach based on Graph Neural Networks (GNNs), we called GraphX-Net, that harnesses Shapley (SHapley Additive exPlanations) values to evaluate the significance of features and generate an explainable graph representation. This is achieved by employing Explainable Artificial Intelligence (XAI) techniques to enhance transparency and interpretability in predicting relapse. Specifically, we used the XGBClassifier algorithm, known for its interpretability, to train a machine learning model. To gain insight at both local and global levels, we incorporate Shapley values as a means of quantifying the contribution of each feature to relapse prediction for individual patients, as well as the entire dataset. To classify unlabelled nodes and establish edges between nodes, we adopt a weighted approach that considers the contributions of features to measure probabilities. Furthermore, we initialised two graph convolutional layers as message passing layers to aggregate, transform, and update node representations, incorporating the information learned from the underlying graph structure.

Here are the main contributions of GraphX-Net:

- The GraphX-Net utilizes GNNs, which are well-suited for capturing complex relationships and dependencies within graph-structured data. In the context of breast cancer, representing patients and their relevant features as nodes in a graph, with the edges representing the relationships between patients based on similarity.
- GraphX-Net can effectively capture local and global information about patients and their features, offering more accurate predictions compared to traditional methods.
- GraphX-Net incorporates Shapley values to provide the interpretability and explainability of the predictions. Using Shapley values, it can identify the importance of different features in the prediction of relapse, allowing clinicians and researchers to understand the underlying factors driving the predictions.
- GraphX-Net provides a powerful tool for clinicians and researchers in the field of breast cancer by combining GNNs with Shapley values. It offers accurate predictions of relapse risk while also providing interpretable insights into the underlying factors driving those predictions.

II. RELATED WORK

Breast cancer constitutes a multifaceted disease influenced by a number of factors influencing its course and likelihood of recurrence. Traditional machine learning methodologies frequently disregard the inherent interdependencies and associations between individual patients, thus restricting their predictive efficacy [14]. On the contrary, graph neural networks (GNN) present a compelling avenue to harness the intricate network of relationships between patients and their associated attributes, thus improving the precision of the prediction of breast cancer relapse [15]. In this section, we explore the application of GNNs in predicting breast cancer relapse, highlighting their potential to improve prognostic accuracy and help in personalized treatment strategies. Recent studies have explored deep learning's ability to automatically learn intricate data representations, proving its effectiveness in extracting essential features for classification tasks. For example, A deep Neural Network (DNN) with a DBSCAN clustering algorithm was used in the second stage to reveal morphological features of cancerous regions [18].

The graph convolutional neural network (GCNN) was used to showcase 20 gene signatures to predict the likelihood of relapse in patients with breast cancer (BRCA). The prognostic and diagnostic capabilities of these genes were rigorously tested against other established algorithms and biomarkers, establishing the superior performance of the GCNN genes. [19].

Some machine learning (ML) algorithms were constructed and assessed, including Random Forest, Boosting, and Stacking. These models were built to compare their performance with the Graph Neural Network model. Prior to evaluating the models, the data was split and SMOTE (Synthetic Minority Oversampling Technique) was utilized to address the class imbalance. Additionally, a feature selection technique, such as using a forest-based method [20], was applied to identify important features. These steps were carried out before proceeding with the evaluation of each model to assess its performance.

Random Forest is a popular ensemble learning method that is used for both classification and regression tasks in machine learning. It is a combination of multiple decision trees, where each tree is trained on a random subset of training data and features. The final prediction is determined by aggregating the predictions of all individual trees [21]. Boosting is an ensemble learning technique in machine learning where multiple weak learners (often decision trees) are combined to create a strong learner. Unlike random forests, which train multiple trees independently, boosting trains weak learners sequentially, with each subsequent learner focusing on the mistakes made by the previous ones. The key idea behind boosting is to iteratively improve the overall model by giving more weight to the misclassified or difficult examples [22]. Stacking, also known as stacked generalization, is a machine learning ensemble technique that combines multiple models to improve predictive performance. It involves training multiple base models on a dataset and then training a meta-model that takes the predictions of the base models as input to make the final prediction [23].

Having discussed related work, here we introduce a novel Graph Neural Network model, called GraphX-Net, that leverages Shapley values to construct an efficient neural graph tailored to the data. This neural graph is designed to fit entirely in memory and represents the graph data using graph information components, including node features and edges. In the context of relapse breast cancer data, where edge weights are unavailable, we utilise these values effectively to connect patient nodes and create distinct clusters based on their Shapley value contributions. Given the subtle differences in contributions, we introduce two additional parameters, the node threshold and node connections, to manage the relationships among nodes. The integration of Graph Neural Networks and Shapley values enhances the accuracy and interpretability of relapse prediction for breast cancer patients, addressing the critical need for accurate and explainable models in the field of breast cancer prognosis. GraphX-Net not only achieves high prediction accuracy but also offers valuable insights into the specific features and interactions within the patient's treatment graph that contribute to the relapse risk. This unique contribution enables clinicians and researchers to understand the underlying mechanisms that drive relapse in breast cancer and to make informed decisions regarding treatment and follow-up strategies. Our extensive experiments and comparative analysis underscore the substantial progress achieved by the GraphX-Net model integrated with Shapley values in the realm of breast cancer prognosis. These findings pave the way for better patient care and personalized medicine.

III. GRAPHX-NET

The GraphX-Net model incorporates a fusion of machine learning algorithms and deep learning algorithms. We first transformed the data into a graph representation, where each patient is represented as a node within the graph, while the edges symbolise the interconnections between patients. The determination of node edges is based on the assessment of input contributions. By analysing the positive or negative contributions of each input, the corresponding node labels can be identified.

The utilisation of Shapley Values (see Fig. 1) enables quantification of each feature's contribution to the predictions made by GraphX-Net model. This computation involves iteratively training the model, while systematically excluding distinct features during each iteration. Through a comparison of prediction differences between the complete model and models with excluded features, Shapley values were attributed to individual features. These values accurately depict the incremental importance of each feature in the prediction process. In the SHAP summary plots, colors convey important information about the value of the features and their impact on the model's predictions. Specifically, the colors in a SHAP summary plot are used to indicate the magnitude and direction of the feature values. Here is a detailed explanation of what

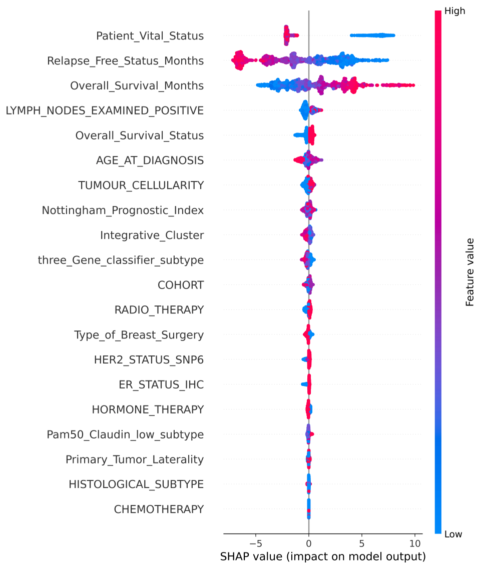


FIGURE 1: The important features according to their Shapley values contributions that play a critical role in understanding and interpreting the impact of features on a machine learning model's output.

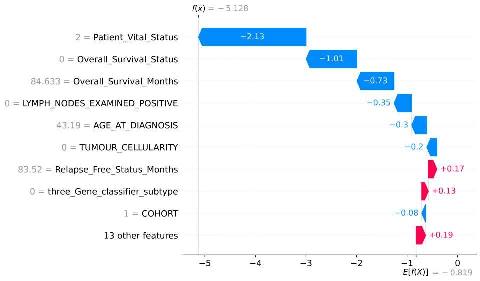


FIGURE 2: Waterfall SHAP values illustrate how SHAP values impact individual predictions in GraphX-Net model.

the colors represent [24]:

- **Feature Values**:
  - **Red/Hot Colors**: Indicate high values of the feature.
  - **Blue/Cool Colors**: Indicate low values of the feature.
- **Impact on Model Prediction**:
  - The position of the points along the x-axis represents the SHAP value. Positive SHAP values (to the right) increase the predicted outcome, while negative SHAP values (to the left) decrease it.
- **Combination of Color and Position**:
  - By combining color and position, we can infer how different values of a feature affect the prediction.

The SHAP summary plot plays a crucial role in discerning the most significant features and their respective range of effects throughout the dataset. By examining the Shapley Values within the context of the GraphX-Net model, valuable insights and interpretations could be obtained. It aids in the identification of influential features and facilitates an understanding of how these features influence the predictions generated by the GNN. Through the visualisation and analysis of the Shapley Values (see Fig. 2), a deeper comprehension of the underlying relationships and dependencies within the graph can be attained. Furthermore, this process helped identify potential biases or limitations present within the model [24].

- f(x): is the prediction after considering all features.
- E[f(x)]: is the mean prediction.
- The blue bar shows how much a particular feature decreases the value of the prediction.
- The red bar shows how much a particular feature increases the value of the prediction.

While GNN architecture diagram in itself does not directly depict the Shapley Values, incorporating these values into the analysis and interpretation of the GraphX-Net model greatly enhanced comprehension. It provides valuable insights into the relative importance of different features within the prediction process.

1) Graph Construction

Graph construction process involves the establishment of edges between nodes, following predefined criteria. This key task is of paramount importance, as it forms the basis for the creation of pertinent connections between patients within the network. In this endeavour, we utilised the collective contributions of all input attributes encapsulated within each node to discern the intricate relationships that govern the adjacency's between the focal node and its counterparts in the graph.

$$f(N_c) = \sum_{i=0}^{n}(l_1sv, l_2sv, l_3sv, \ldots, l_nsv) \qquad (1)$$

This equation calculates the Shapley value scores for specific features associated with each node i from 0 to n. Here, l_isv represents the Shapley value for feature i, and the sum aggregates these values to quantify the contribution of all relevant features across the nodes.

Through the computation of contributions for all inputs within each observation, accounting for both positive and negative contributions, we can determine the probabilities assigned to each node's classification as either 0 (indicating "No") or 1 (indicating "Yes"). Moreover, by considering the average contribution of adjacent nodes, we can assess their respective adjacencies.

By performing calculations on the contributions of all inputs for each observation, encompassing both positive and negative contributions, we derived the

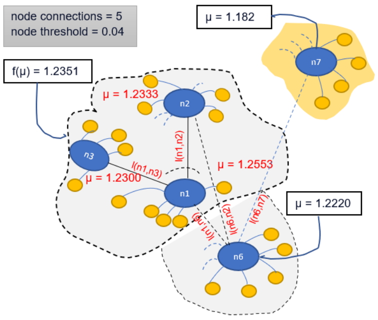


FIGURE 3: This diagram depicts the process of connecting a node by considering the information from its neighboring nodes. The figure highlights that the information of clusters is influenced by both the node threshold value and the requirement that the number of nodes in each cluster not exceed the node connections parameter.

probabilities associated with the classification of each node as 0 (representing "No") or 1 (representing "Yes"). Additionally, we evaluated the adjacencies of neighboring nodes based on their respective contributions. For example, it suggests that the node should be connected to a group where the average contribution is either less than or equal to a predefined threshold. This threshold serves as an integer parameter that is incorporated into the grouping of nodes. To determine the connectivity of a new node to a group, the difference between the average contributions of the group and the contribution of the new node was assessed against the threshold. If this difference was less than or equal to the threshold, and the total number of nodes in the group remains within the permitted group size, the new node should be connected to the group as depicted in Fig. 3. The figure shows different colors to indicate clusters of nodes. Nodes within the gray area are all belonging to a single cluster, where the mean contributions meet the node threshold value. In contrast, the node in the yellow area, although connected to the gray cluster, belongs to a different cluster. This is because its mean contribution does not align with the overall mean of the gray cluster.

2) Feature Extraction

Here we used XGBoost method for calculating the relative contribution of each feature to the model. This is based on the gain metric to evaluate the importance of a feature by considering its contribution between individual trees within the model [25]. A higher gain value, relative to other features, signifies greater significance in generating predictions. Consequently, the gain metric served as a pivotal attribute for interpreting the relative importance of each feature. By scrutinizing the gain values, we can effectively pinpoint the most influential features within the model as shown in Table 2.

Algorithm 1 Calculate the total contributions of each node

```
Input: lst_shap_values                    ▷ List of all Shapley values
Input: df                                 ▷ Data frame
Output: df                                ▷ Data frame updated
for each $i$, row ∈ df do
    pos_cont ⇐ 0
    neg_cont ⇐ 0
    for each $c$ ∈ df.cols do
        f_cont ⇐ lst_shap_values[$c$].values[$i$]
        if f_cont > 0 then
            pos_cont ⇐ f_cont + pos_cont
        end if
        if f_cont < 0 then
            neg_cont ⇐ f_cont + neg_cont
        end if
    end for
    df.at[$i$, 'pos_cont'] ⇐ pos_cont
    df.at[$i$, 'neg_cont'] ⇐ neg_cont
end for
return df
```

Algorithm 2 Create Graph's Adjacency Matrix

```
Input: df                                 ▷ Data frame
Output: adjacency Matrix
for each row in df do
    neg_cont ⇐ row.neg_cont
    pos_cont ⇐ row.pos_cont
    classi ⇐ row.relapse_free_status
    patient_id ⇐ row.patient_id
    node_ad ⇐ get_node_adjacents(neg_cont, pos_cont, classi, patient_id)
    adjacency matrix ⇐ node_ad
end for
return adjacency Matrix
```

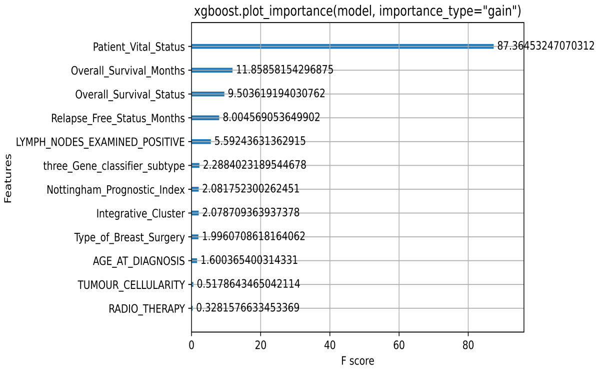


FIGURE 4: Visualizing Feature Importance with XGBoost Gain.

Patient attributes and clinical data are of significant importance in the prediction of relapse in breast cancer cases. Various features, including patient demographics, medical history, genetic information, and treatment records, are carefully extracted and asso-

Algorithm 3 Get Node Adjacents

```
Input: node_connections              ▷ Cluster nodes = 5
Input: node_threshold     ▷ Node's threshold parameter = 0.04
Input: neg_cont               ▷ Node negative contributions
Input: pos_cont               ▷ Node positive contributions
Input: classi                        ▷ Node classification
Output: edge_indexes
for each row in df do
    if (classi = 0 and row.relapse_free_status = 0) then
        Add node, negative contributions into an array
    end if
    if (classi = 1 and row.relapse_free_status = 1) then
        Add node, positive contributions into an array
    end if
end for
if (classi == 0) then
    Sorting negative contributions
    for each $i, v$ in neg.dic do
        if (neg_cont < $v$ and $i \neq$ patient_id) then
            if (adjacents $\geq$ node_connections) or (count > 1 and ($v$ - (sum/count)) > node_threshold) then
                Exit Loop
            end if
        end if
        adjacentsArray $\Leftarrow$ node
        count $\Leftarrow$ count + 1
        sum $\Leftarrow$ sum + $v$
    end for
end if
if (classi == 1) then
    Sorting positive contributions
    for each $i, v$ in pos.dic do
        if (pos_cont < $v$ and $i \neq$ patient_id) then
            if (adjacents $\geq$ node_connections) or (count > 1 and ($v$ - (sum/count)) > node_threshold) then
                Exit Loop
            end if
        end if
        adjacentsArray $\Leftarrow$ node
        count $\Leftarrow$ count + 1
        sum $\Leftarrow$ sum + $v$
    end for
end if
return adjacentsArray
```

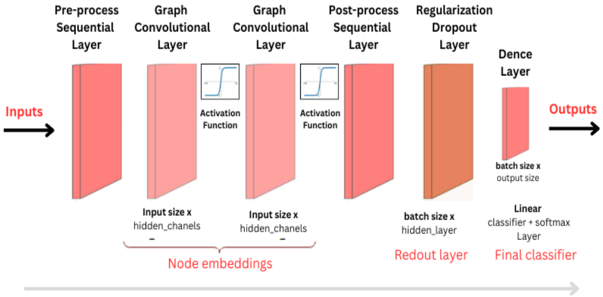

Pre-process Sequential Layer
Graph Convolutional Layer
Graph Convolutional Layer
Post-process Sequential Layer
Regularization Dropout Layer
Dence Layer
Inputs
Activation Function
Activation Function
Outputs
batch size x output size
Input size x hidden_chanels
Input size x hidden_chanels
batch size x hidden_layer
Linear classifier + softmax Layer
Node embeddings
Redout layer
Final classifier

FIGURE 5: GraphX-Net model architecture. The model applies preprocessing using feed- forward network to the node features to generate initial node representations. Also applies two graph convolutional layers, with skip connections, to the node representation to produce node embeddings. And finally feed the node embeddings in a Softmax layer to predict the node class. Each graph convolutional layer added captures information from a further level of neighbours.

ciated with each patient node. These features were obtained using the gain metric, which allows for the identification of relevant attributes to predict relapse. These features are incorporated as node features in the graph representation. This integration enabled a thorough analysis of the breast cancer relapse graph by leveraging the insights gained from the XGBoost model. The presence of these features as attributes associated with each node offers valuable information for various graph-based tasks, such as node classification, anomaly detection, and graph clustering.

By incorporating the insights obtained by the XGBoost model within the graph context, a deeper understanding of the underlying patterns and relationships within the breast cancer relapse graph can be achieved. Each node within the graph corresponds to a patient, with the colour assigned to each node indicating the occurrence of relapse for that particular patient. It is worth mentioning that due to the size of the data set, only a sample of patients is displayed in this visualisation.

3) GNN Model Architecture

The selection of an appropriate architecture for the neural network graph model is of paramount importance to achieve accurate predictions for breast cancer relapse. Models such as Graph Convolutional Networks and other GNN variants are specifically designed to leverage the inherent connectivity patterns within the graph data. These models excel at learning meaningful representations by both the features of individual nodes and the information from their neighbouring nodes. Through iterative information aggregation, these models can effectively capture local and global dependencies, enabling robust predictions for breast cancer relapse.

The graph data is represented by the graph_info tuple, which consists of the following three elements:

- node features: This is a [num nodes, num features] NumPy array that includes the node features. In this dataset, the nodes are the patient IDs, and the node features are the features that we have extracted in the feature extraction section.
- edges: This is [num edges, num edges] NumPy array representing a sparse adjacency matrix of the links between the nodes. In this example, the links are the citations between the papers.
- edge weights (Optional): No weight relationships in the graph between patients with relapsed breast cancer.

4) Implementation

The model architecture has been implemented as in Fig. 5. The graph convolutional layers perform the following steps:

- **preparation**: the input node representations were processed using a feed-forward Network to generate a message. To simplify the processing, a linear transformation was applied to the node representations. It includes three layers: BatchNormalization, Dropout, and Dense with the activation function Gelu.
  1) BatchNormalization: it normalizes the input data along the dimension of the features. It helps stabilize and accelerate training by reducing the internal covariate shift. The input here should be a tensor type which is a multi-dimensional array of numbers. The formula for BatchNormalization can be expressed as:

  $$y = \frac{x - \text{mean}(x)}{\sqrt{\text{var}(x) + \epsilon}} \times \gamma + \beta \tag{2}$$

  As proved by [26]. Where:
     ✸ x is the input tensor.
     ✸ $\text{mean}(x)$ and $\text{var}(x)$ are the mean and variance of x over the batch. Variance means how far x is from the mean.
     ✸ γ and β are learnable scaling and shifting parameters, respectively.
     ✸ $\in$ is a small constant to avoid division by zero.
  2) Dropout: Dropout is a regularization technique that randomly sets a fraction of the inputs to zero during training. This helps prevent overfitting and enhances generalization. The formula for Dropout is simple, where ρ is the dropout rate:

  $$\text{Dropout}(x) = x \times \text{mask} \tag{3}$$

  $$\text{mask} \sim \text{Bernoulli}(1 - p)$$

  In this equation:
     ✸ x represents the input tensor.
     ✸ mask is a random binary mask with the same shape as x that is drawn from a Bernoulli distribution with a probability of $1 - p$, where p is the dropout rate.
  3) Dense with Gelu Activation: The Dense layer performs a linear transformation on the input data and applies the Gelu activation function. Gelu is a variant of the ReLU activation and is defined as follows:

  $$\text{Gelu}(x) = 0.5x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}}\left(x + 0.044715x^3\right)\right)\right) \tag{4}$$

  As proved by [27]

- **Aggregate**: The messages of the neighbours of each node are aggregated concerning the edge weights using a permutation invariant pooling operation, such as sum, mean, and max, to prepare a single aggregated message for each node. In the message-passing framework of GNNs, node features are update d by aggregating messages from neighboring nodes. Using permutation invariant pooling operations ensures that the aggregated message is the same regardless of the order in which neighbors are considered.
  - Sum-Pooling: $(x_1, x_2, \ldots, x_n) = \sum_{i=1}^{n} x_i$
  - Mean-Pooling: $(x_1, x_2, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Max-Pooling: $(x_1, x_2, .., x_n) = \text{max}_i x_i$

  $$\text{Aggregate}(\text{messages}) = \sum_{i \in \text{neighbors}} \text{messages}_i \tag{5}$$

  This equation represents the aggregation process, where messages are a set of messages prepared for a given node, and the sum iterates over the neighbouring nodes of the current node to aggregate their messages.
- **Update**: The node representations and aggregated messages both of shape [num nodes, representation dim] are combined and processed to produce the new state of the node representations (node embeddings). The node representations and aggregated messages are concatenated to create a sequence and then processed by Concatenating the node representations and aggregated messages.

  $$\begin{aligned} &\text{Update}(\text{node embedding, aggregated messages}) = \\ &\text{Concatenate}(\text{node embedding, aggregated messages}) \end{aligned} \tag{6}$$

  This equation represents the update process, where the node embedding is updated by adding the aggregated messages from neighbouring nodes and passing the result through the sigmoid activation function.

$$\text{Reset gate: } z_t = \sigma(W_z \cdot x_t + U_z \cdot h_{h-1} + b_z)$$

$$\text{Update gate: } r_t = \sigma(W_r \cdot x_t + U_r \cdot h_{h-1} + b_r)$$

$$\text{Candidate hidden state: } \overline{h}_t = \tanh(W_h \cdot x_t + U_h \cdot (r_t \odot h_{t-1}) + b_h)$$

$$\text{Hidden state: } h_t = (1 - z_t) \odot (h_{t-1} + z_t \odot \overline{h}_t)$$

As proved by [28]. In these equations:

- x_t represents the input at time step t.
- h_{t-1} represents the hidden state (output) from the previous time step $t-1$
- z_t is the reset gate, controlling which parts of the previous state should be ignored.
- r_t is the update gate, controlling how much of the previous state to keep.
- $\overline{h}_t$ is the candidate hidden state that combines the input and the reset-gated previous state.
- h_t is the updated hidden state at time step t.
- W_z, W_r,W_h are learnable weight matrices.

GRU (Gated Recurrent Unit) has been used in the GraphX-Net model. Combining different layers, including Graph Convolutional Layers, Pooling Layers, and Recurrent Layers such as GRU. We can typically feed the node features and graph structure (e.g., adjacency matrix or graph Laplacian) into the GraphX-Net model. The GRU layer can be used to capture temporal dependencies among neighbouring nodes and iteratively update the node embeddings during the message passing process [29].

$$h_n^{(k+1)} = \text{UPDATE}_u^{(k)}\left(h_u^{(k)}, \text{AGGREGATE}_u^{(k)}\left(\{h_v^{(k)}, \forall v \in N(u)\}\right)\right) \tag{7}$$

As proved by [30]. In this equation:

- $h_n^{(k+1)}$ represents the node embedding of node n at the $(k+1)-th$ layer.
- $h_u^{(k)}$ represents the node embedding of node u at the $k-th$ layer.
- $h_v^{(k)}$ represents the node embedding of neighboring node v of node u at the $k-th$ layer.
- UPDATE$_u^{(k)}$ is the update function that takes the current node embedding $h_u^{(k)}$ and the aggregated messages from neighbouring nodes as inputs and updates the node embedding of node u to $h_v^{(k)}$ for the $(k+1)-th$ layer.
- AGGREGATE$_u^{(k)}$ is the aggregation function that takes the node embeddings of neighbouring nodes $h_v^{(k)}, \forall v \in N(u)$ and combines them to obtain aggregated messages for node u at the $k-th$ layer.

$$h_u^{(k+1)} = \text{UPDATE}_u^{(k)}(h_u^{(k)}, m_{N(u)}^{(k)}) \tag{8}$$

As proved by [30]. In this equation:

- $h_u^{(k)}$ represents the node embedding of node u at the $k-th$ layer.
- $m_{N(u)}^{(k)}$represents the aggregated messages from the neighboring nodes $N(u)$ of node u at the $k-th$ layer.
- UPDATE$_u^{(k)}$ is the update function that takes the current node embedding ($h_u^{(k)}$ and the aggregated messages $m_{N(u)}^{(k)}$ as inputs and updates the node embedding to $h_u^{(k+1)}$ for the $(k+1)-th$ layer.

In the context of message-passing iterations in a Graph Neural Network, we utilise the terms UPDATE and AGGREGATE to refer to arbitrary differentiable functions, typically implemented as neural networks. These functions are responsible for updating the node embeddings and aggregating information from the graph neighbourhood. The aggregated messages from the neighbouring nodes $N(u)$ of node u at the k-th layer are denoted as $m_{N(u)}^{(k)}$.

In the message-passing process of a Graph Neural Network, the update function UPDATE combines the message $m_{N(u)}^{(k)}$ with the previous embedding $h_u^{(k-1)}$ of node u to generate the updated embedding $h_u^{(k)}$. At the initial iteration ($k = 0$), the embeddings are set to the input features for all nodes, i.e., $h_u^{(0)} = x_u, \forall u \in V$, where x_u represents the input features of node u. After running K iterations of the GNN message passing, we can use the output of the final layer to define the embeddings for each node, i.e.,

$$z_u = h_u^{(k)}, \forall u \in V \tag{9}$$

This is proved by [31]. In this equation:

- z_u represents the embedding vector for node u.
- $h_u^{(k)}$ represents the node embedding of node u at the $k-th$ layer.
- $\forall u \in V$ means the equation applies to all nodes u in the set of nodes V.

The superscripts, such as k, were used to differentiate the embeddings and functions at different iterations of the message-passing process. This enabled us to track and distinguish the evolving embeddings and functions throughout the message-passing iterations in the GNN models.[32]

The main steps of our GraphX-Net model can be described as:

- Apply reprocessing using feed-forward to the node features to generate initial node representations.
- Apply one or more graph convolutional layers, with skip connections, to the node representation to produce node embeddings.
- Apply post-processing using feed-forward network to the node embeddings to generate the final node embeddings.
- Feed the node embeddings in a Softmax layer to predict the node class.

With each additional graph convolutional layer, the model was able to capture information from an extended range of neighbouring nodes.

- Each graph convolutional layer in the model captured and integrated information from neighbouring nodes.
- The layer applies a convolutional operation on the node representations, taking into account the connections and relationships defined by the graph structure.
- The convolutional operation involved aggregating and combining information from neighbouring nodes to update the representations of each node (Eq. 9).

IV. PATIENTS' COHORT AND EXPERIMENTAL SETUP

This study utilised a large cohort of invasive breast cancer with long-term clinical follow-up and complete clinicopathological data. This dataset consists of1980 breast cancer cases within the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort. The METABRIC study protocol, detailing the molecular profiling methodology in a cohort of 1980 breast cancer samples is described by [33]. Patient demographics are summarised in Appendix A. Data was downloaded from the cBioportal data repository and it include of clinical records, patient demographics, tumor characteristics, treatment details, and follow-up data sourced from trusted repositories [34].

A. DATA PRE-PROCESSING

Here the original data has undergone multiple pre-processing steps to ensure its suitability for analysis. Initially, the data was presented in a tabular format, requiring various data-wrangling operations. One of the challenges encountered was dealing with missing values, which could not be uniformly imputed across the dataset. Consequently, these missing values were excluded from the analysis to ensure the compatibility of the data. In addition, certain parameters were categorised to enhance their utility in subsequent analyses. Inconsistencies were reviewed and rectified to maintain data integrity.

Following data cleaning, we further performed pre-processing tasks were performed, including feature selection and normalisation. Feature selection helped identify the most informative variables for relapse prediction, reducing dimensionality and improving model performance. Normalization ensured that features were on a common scale, preventing bias due to varying magnitudes. These cleaning and pre-processing steps formed a crucial foundation for our subsequent analysis and modelling efforts, ensuring reliable and meaningful results.

The processing of the METABRIC dataset includes crucial steps to facilitate effective analysis and modelling. Initially, the selection of the target variable, especially the 'Relapse free status' plays a pivotal role. Feature selection techniques are then applied to iden-

Illustration	**METABRIC cross-sectional data**
Size of the training set	935
Size of the test set	584
Total class count	2
Total number of features in the dataset	24
Number of trainings data features after correlation	10
Number of features in the optimal set	9
Learning rate	1.2 x 10^{-2}
Hidden units	[32,32]
Dropout rate	0.5
Epochs	300
Batch size	128

TABLE 1: Experiment findings on feature selection and other metrics

#	**Feature**
1.	Patient's Vital Status
2.	AGE_AT_DIAGNOSIS
3.	Relapse Free Status (Months)
4.	Overall Survival (Months)
5.	LYMPH_NODES_EXAMINED_POSITIVE
6.	Overall Survival Status
7.	TUMOUR CELLULARITY
8.	Nottingham Prognostic Index
9.	Integrative Cluster

TABLE 2: The most significant 9 input features in the Relapse breast cancer dataset.

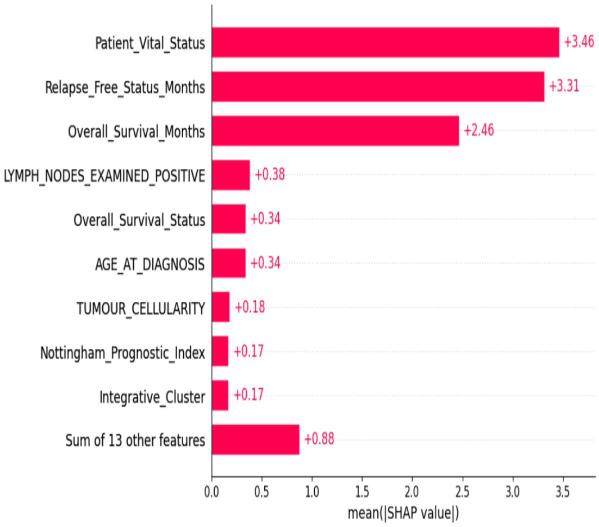


FIGURE 6: An illustrative SHAP bar plot is employed to evaluate global feature importance, offering a comprehensive assessment of the influence of each feature on the model's predictions. The plot showcases how each feature contributes to the model's output, enabling a deeper understanding of their impact and aiding in model interpretation, thereby enhancing the transparency and reliability of the predictive framework.

tify the most influential attributes affecting this clinical outcome. To mitigate class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was employed, creating synthetic samples for the minority class. The ExtractTreeClassifier class can extract valuable features from the dataset, providing insights beyond raw attributes. Following feature selection and extraction, the dataset was partitioned into training and test sets. Subsequently, the importance of the features was determined using methods such as the XGBoost classifier and the Shapley values, offering a clear understanding of the impact of each feature. Classical machine learning models, including Boosting, Random forest, and Stacking algorithms, were constructed and assessed for predictive accuracy. To explore feature interactions in depth, a graph-based approach was adopted, where features are represented as nodes and their relationships as edges. Graph neural networks were used for node classification in this feature graph, providing a holistic perspective of feature interactions and their influence on the target variable. This comprehensive approach improves our understanding of breast cancer prognosis and treatment decisions.

In summary, METABRIC dataset processing involves meticulous target selection, feature engineering, SMOTE for class balance, advanced feature analysis with XGBoost and Shapley values, classical ML modelling, and GNN-based feature interaction exploration. These steps collectively empowered robust predictions and a deeper understanding of breast cancer patient outcomes.

B. TRAINING SETTINGS

For the classification of nodes, the GraphX-Net model was trained using GNN. Specifically, we focus on the architecture of the model, which includes preprocess and post-process layers to enhance the model's performance Fig. 5.

C. GRAPHX-NET TRAINING

1) Pre-process Layer: to prepare the input data and set the stage for effective learning, which was constructed using three distinct layers: BatchNormalisation, Dropout, and Dense with Gelu activation.
    a) BatchNormalisation: to normalise the node features during training, thereby reducing the internal covariate shift. This layer helps stabilise the learning process and accelerates convergence, ultimately improving the model's performance.
    b) Dropout: The Dropout layer introduces regularisation by randomly dropping out a fraction of the nodes during each training epoch. This is to prevent overfitting and to ensure that the model learns robust representations that generalise well to unseen data.
    c) Dense with Gelu activation: Apply linear transformation followed by a Gelu activation function to introduce nonlinearity. Gelu is a Gaussian Error Linear Unit activation that is known to perform well in deep learning models and has a smooth gradient, allowing for more stable training [35].
2) GraphConvLayers Module:
   The backbone of our node classification model lies in the two graph convolutional layers. Graph convolutional layers have been designed specifically to handle graph-structured data and can capture local and global dependencies among nodes. It aggregates information from neighbouring nodes to update the node embeddings. It leverages the graph's adjacency matrix to weigh the importance of each neighboring node, capturing the graph's structural dependencies. Each layer refines the node representations, allowing the model to gradually understand the underlying graph topology.
3) Post-process Layer:
   The postprocess layer was the final part of our GNN-based node classification model, which is responsible for refining the learned representations and generating the final node predictions. Similar to the pre-processing layer, the postprocess layer consisted of three layers: BatchNormalisation, Dropout, and Dense with Gelu activation. The model outputs a probability distribution for each node, indicating the likelihood of belonging to each class.

Training a GNN model for node classification using GraphConvLayer module involved a thoughtful design of the model architecture. The incorporation of pre-process and post-process layers, each consisting of BatchNormalisation, Dropout, and Dense with Gelu activation, enhanced the model's ability to capture meaningful features and structural dependencies present in the graph. With this powerful approach, we can achieve accurate node classification results in various applications, enabling us to harness the potential of GNNs in solving real-world problems effectively.

D. PERFORMANCE EVALUATION

To evaluate the GraphX-Net model, its performance was thoroughly assessed to assess its ability to be generalised effectively to unseen data. Several evaluation techniques are commonly employed for this purpose. The key techniques utilised in GraphX-Net evaluation included:

- **Train-Validation-Test Split**: The data was split into train 0.07 and 0.15 for validation and test data respectively. The GraphX-Net model

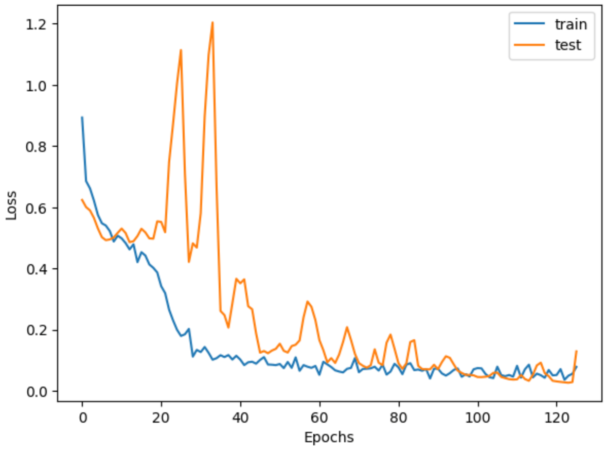


FIGURE 7: Analysing the Reduction in Loss Across Epochs to Optimise the Performance of GraphX-Net in Breast Cancer Relapse Prediction

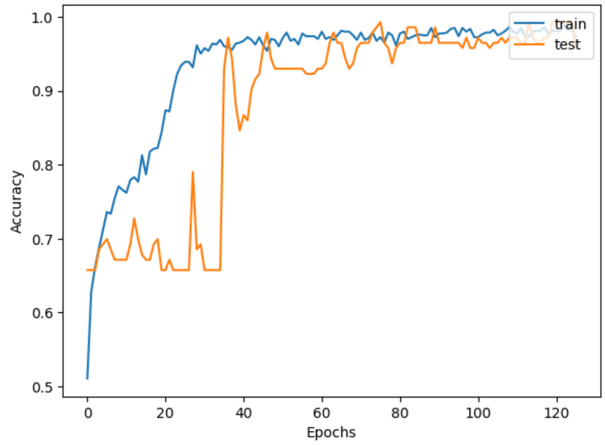


FIGURE 8: Improving Accuracy Across Training Epochs: Optimization of GraphX-Net in Breast Cancer Relapse Prediction

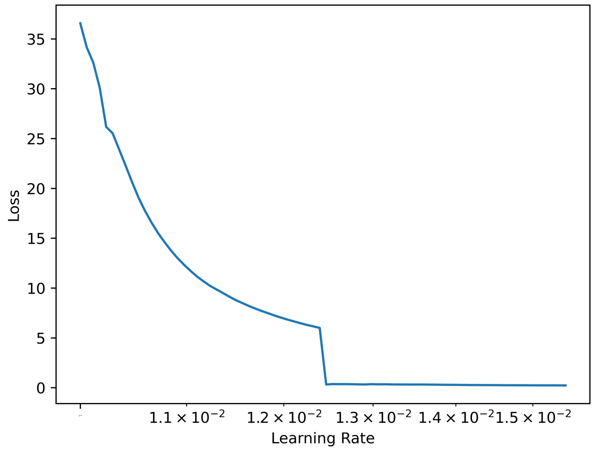


FIGURE 9: Exploring the Impact of Learning Rate on Model Training: An In-Depth Analysis of Learning Rate vs. Loss

Models	Metrics			
	AUC score	F1 score	Balanced accuracy	Cross-validation
RF	76.15	76.47	75.96	83.30
Boosting	75.82	75.88	75.68	82.24
Stacking	75.72	75.20	75.68	80.74

TABLE 3: Classical machine learning model evaluation metrics. K-fold cross-validation across all models = 3.

showed a good performance with a test accuracy of 98.90%.

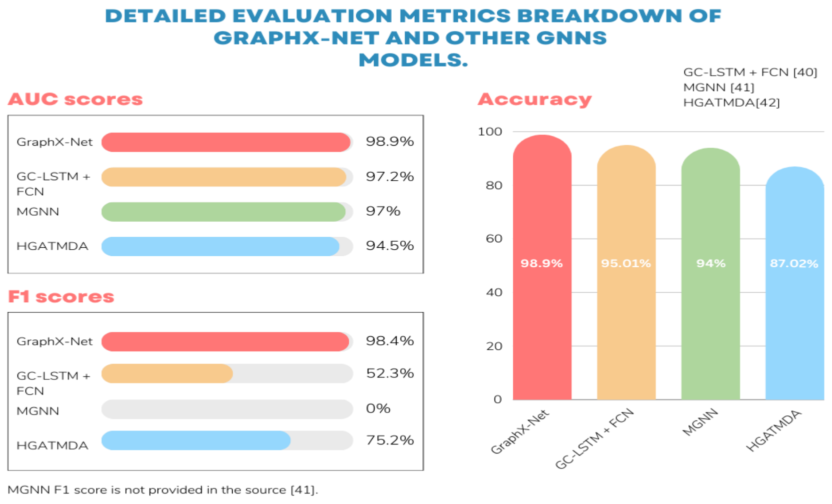
- **Learning Rate Adjustment:** To detect the best learning rate value, Lambda Callback package has been used to calculate the best LR value. With callback actions, we recorded all loss values.
  By analysing the graph, we can identify the optimal learning rate, which is typically the learning rate where the loss decreases the fastest before diverging or oscillating. Here in our case, the graph shows that the best LR value is between 1.2 x 10^{-2} and 1.3 x 10^{-2}
- **Cross-Validation**: Performed k-fold cross-validation to obtain more robust performance estimates.
- **Task-Specific Metrics**: As shown in Table 1.

E. MULTIPART FIGURES

Figures compiled of more than one sub-figure presented side-by-side, or stacked. If a multipart figure is made up of multiple figure types (one part is lineart, and another is grayscale or color) the figure should meet the stricter guidelines.

Due to the extensive process involved in implementing both Graph Neural Network (GNN) and classical machine learning (ML) approaches, as well as the subsequent evaluation of these models, it is crucial to present the outcomes in a clear and organized manner. The complexity of these methodologies requires a thorough examination and comparison to ensure a comprehensive understanding of their performance. To facilitate this, the results have been systematically compiled and are illustrated in the Table 4. This table provides a detailed comparison of various aspects, allowing for an easy and effective assessment of the strengths and limitations of each approach.

Classical machine learning techniques often fail to provide reliable results for complex problems such as breast cancer recurrence prediction. A study on the WPBC dataset, consisting of 198 patients (151 non-recurring and 47 recurring cases), reported accuracy rates of 78.5%, 73.8%, and 67.2% for Support Vector Machine (SVM), Random Forest, and Decision Tree, respectively [42]. Similarly, our study as shown in Table 3 used Random Forest, Boosting, and Stacking yielded accuracies of 75.96%, 75.68%, and 75.68%. In contrast, our GraphX-Net model as in Fig.10, which utilizes Graph Neural Networks (GNNs), achieved a



MGNN F1 score is not provided in the source [41].

FIGURE 10: Detailed evaluation metrics breakdown of GraphX-Net and other GNNs models.

Aspect	GraphX-Net Model	Classical Machine Learning Approaches
Dataset	Same dataset	Relapse breast cancer dataset
Ease of Implementation	Moderate	High
Interpretability	Moderate	High
Performance with Graph-Structured Data	Superior	Limited
Accuracy and Performance Metrics	Displayed in Fig. 10	Displayed in Table 3
Hierarchical Understanding	Leverages hierarchical nature of GNNs for deeper understanding of structural patterns	Lacks ability to capture hierarchical nature of data
Capability to Capture Complex Interactions	High	Limited
Predictive Capabilities	Higher, with powerful predictive capabilities	Lower compared to GraphX-Net
Scalability	Scales better for large and complex datasets	Scales well for small to moderately large datasets
Flexibility in Feature Engineering	Automatically captures relevant features through graph structure	Requires extensive manual feature engineering
Computational Complexity	Higher due to the complexity of graph computations	Generally lower, dependent on the simplicity of the algorithm
Training Time	Longer, due to the complex nature of graph computations	Typically shorter, depending on the algorithm
Handling Missing Data	More robust to missing data due to the interconnected nature of graphs	Requires imputation or preprocessing
Robustness to Noisy Data	More robust due to the ability to capture and utilize relationships within the graph	Varies with the algorithm, often sensitive to noise
Examples of Algorithms Used	Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), GraphSAGE	Decision Trees, Random Forests, Support Vector Machines (SVM), Logistic Regression
Insight into Data Trends	Provides valuable insights by uncovering complex patterns and relationships within the data	Limited to predefined features and relationships
Use Cases Beyond Current Study	Increasingly adopted in areas requiring analysis of relational data, such as social networks, biology, and chemistry	Widely used in various fields such as finance, healthcare, and marketing

TABLE 4: Comparison between GraphX-Net Model and Classical Machine Learning Approaches

significantly higher accuracy of 98.9%, showcasing its superior ability to predict breast cancer recurrence accurately by effectively extracting crucial information from medical data.

V. DISCUSSION AND CONCLUSION

Herein, we introduce a novel Graph Neural Network model, called GraphX-Net, that leverages Shapley values to construct an efficient neural graph tailored to

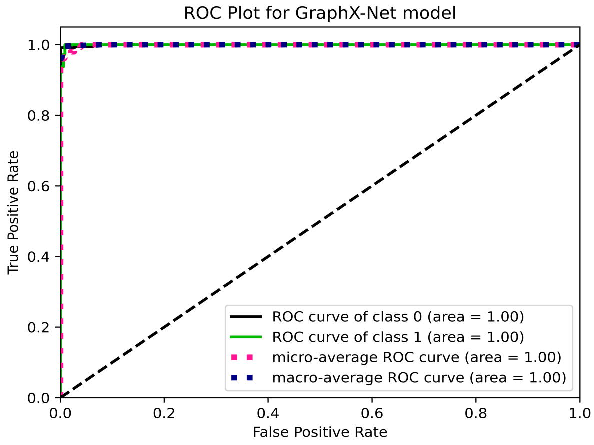


FIGURE 11: ROC Analysis: GraphX-Net Performance in Discriminating Positive and Negative Cases for the model. The classifier's curve goes straight up to (0, 1) and then straight to the right to (1, 1) which is the perfect classifier output.

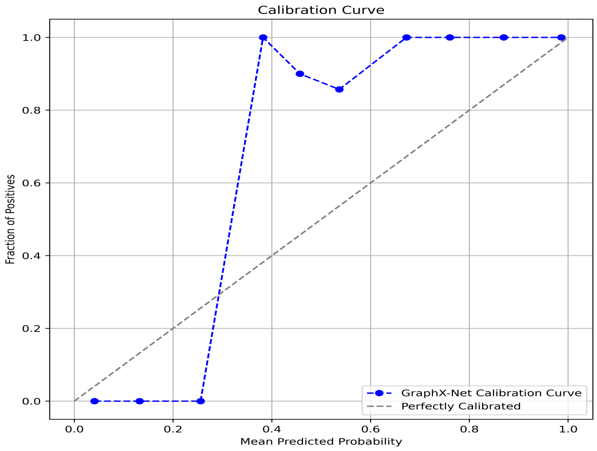


FIGURE 12: GraphX-Net Model Calibration Analysis: Predicted Probabilities vs. Actual Outcomes. The predicted probabilities are divided into several bins or intervals. For each bin, the actual observed event rate is calculated. This is the fraction of true positive cases within that bin.

the data. This neural graph is designed to fit entirely in memory and represents the graph data using graph information components, including node features and edges. In the context of relapse breast cancer data, where edge weights are unavailable, we utilise these values effectively to connect patient nodes and create distinct clusters based on their Shapley value contributions. Given the subtle differences in contributions, we introduce two additional parameters, node threshold and node connections, to manage the relationships among nodes. The integration of Graph Neural Networks and Shapley values enhances the accuracy and interpretability of relapse prediction for breast cancer patients, addressing the critical need for accurate and explainable models in the field of breast cancer prognosis. GraphX- Net not only achieves high prediction accuracy but also offers valuable insights into the specific features and interactions within the patient's treatment graph that contribute to the relapse risk. This unique contribution enables clinicians and researchers to understand the underlying mechanisms that drive relapse in breast cancer and make informed decisions regarding treatment and follow-up strategies. Through rigorous experimentation and comparison with existing approaches, our results demonstrate the significant advancements of GraphX-Net model with Shapley values brings to the field of breast cancer prognosis, opening up new avenues for improved patient care and personalised medicine.

Our study introduces the GraphX-Net model, a novel approach leveraging Graph Neural Networks (GNNs) for predicting breast cancer relapse. Demonstrating its effectiveness in utilising graph structures for classification, the GraphX-Net model offers valuable insights into graph connectivity, node clustering, and layout visualisation, facilitating the conversion of traditional tabular datasets into graphs. By incorporating Shapley values, a deeper understanding of feature importance and underlying factors influencing predictions could be gained, revealing biological mechanisms and potential risk factors associated with breast cancer relapse. The model's interpretability proves vital for clinical decision-making and personalised treatment strategies. Its accurate relapse predictions highlight the potential for clinical translation, enabling real-world healthcare applications and risk assessment for improved patient outcomes. Overall, our research advances precision medicine, fostering tailored and targeted breast cancer therapies worldwide.

The proposed GraphX-Net model holds significant practical and theoretical implications in the fields of computational biology and precision medicine. From a computational perspective, the utilization of Graph Neural Networks (GNNs) allows for the effective handling of structured data, capturing intricate dependencies that traditional models might overlook. This advanced capability ensures more accurate predictions, which are crucial for early intervention and personalized treatment strategies. By achieving state-of-the-art performance metrics, GraphX-Net demonstrates its potential to be a valuable tool in clinical decision-making processes, ultimately contributing to improved patient outcomes. The findings of this study not only advance the field of bioinformatics but also pave the way for future research in leveraging GNNs for various predictive modeling tasks in healthcare and beyond.

One promising future research direction involves the integration of longitudinal patient data to enhance the predictive power of the model. By incorporating temporal information, such as changes in gene expression or treatment responses over time, the model could capture dynamic patterns and provide even more accurate predictions of breast cancer recurrence. Additionally, exploring the application of GraphX-Net to other types of cancer or diseases with similar multifactorial etiologies could further validate its utility and extend its benefits to a broader range of medical conditions.

VI. CODE AVAILABILITY

The code is publicly available at
https://github.com/abdullahbasaad/GraphX-Net.git.
.

APPENDIX A SUPPLEMENTARY TABLE S1:

Variables	**N (%)**
Age at diagnosis[Median (range)]	61.8 (21.93-96.29)
Tumour size [Median (range)]	23 (1, 182)
NPI [Median (95% CI)]	4.46 (4.41-4.51)
Survival [Median (Months, 95% Cl)]	149 (141-159)
Axillary lymph nodes status	
0	1035
1	337
2	171
3	114
>3	314
Axillary lymph nodes status	
Positive	1497
Negative	438
Null	42
PAM50 subtype	
Basal	330
HER2	238
Luminal A	715
Luminal B	489
Normal-like	199
Not classified	6
Adjuvant systemic therapy (AT)	
No AT	305
Hormone therapy (HT)	1216
Chemotherapy	416
Hormone + chemotherapy	192

TABLE 5: Clinicopathological characteristics in the METABRIC cohort.

REFERENCES

[1] Rabbani, S. & Mazar, A. Evaluating distant metastases in breast cancer: from biology to outcomes. Cancer Metastasis Rev.. **26**, 663-674 (2007,12)

[2] Nicolini, A., Giardino, R., Carpi, A., Ferrari, P., Anselmi, L., Colosimo, S., Conte, M., Fini, M., Giavaresi, G., Berti, P. & Miccoli, P. Metastatic breast cancer: an updating. Biomed. Pharmacother.. **60**, 548-556 (2006,11)

[3] Riggio, A., Varley, K. & Welm, A. The lingering mysteries of metastatic recurrence in breast cancer. British Journal Of Cancer. **124** pp. 13 - 26 (2020)

[4] Hutchinson, L. & Devita, V. Focus issue on biomarkers. Nature Reviews Clinical Oncology. **7** pp. 295-295 (2010)

[5] Chen, L., Hodskins, J., Chokshi, S., Croley, J., Stevens, M., Pasley, G., Huller, K., Reynolds, J., Weiss, H. & Massarweh, S. P5-13-24: A Predictive Model of Early Systemic Disease Relapse after Standard Adjuvant Therapy for Breast Cancer.. Cancer Research. **71** (2011)

[6] Ullah, A., Jabeen, S., Zaman, S., Hamraz, A. & Meherban, S. Predictive potential of K-Banhatti and Zagreb type molecular descriptors in structure–property relationship analysis of some novel drug molecules. Journal Of The Chinese Chemical Society. (2024)

[7] Ahmed, W., Ali, K., Zaman, S. & Raza, A. Molecular insights into anti-Alzheimer's drugs through predictive modeling using linear regression and QSPR analysis. Modern Physics Letters B. pp. 2450260 (2024)

[8] Won-Song, Huang, T., Yoo, S., Lee, E., Zhao, Y., Wang, L., Tu, Z., Dai, X., Irie, H., Zhu, J. & Zhang, B. Abstract 363: Planar filtered gene regulatory networks in breast cancer. Cancer Research. **74** pp. 363-363 (2014)

[9] Testa, U., Castelli, G. & Pelosi, E. Breast cancer: A molecularly heterogenous disease needing subtype-specific treatments. Med. Sci. (Basel). **8**, 18 (2020,3)

[10] Yuan, L., Guo, L., Chang-Yuan, Zhang, Y., Han, K., Nandi, A., Honig, B. & De-Huang Integration of Multi-Omics Data for Gene Regulatory Network Inference and Application to Breast Cancer. IEEE/ACM Transactions On Computational Biology And Bioinformatics. **16** pp. 782-791 (2019)

[11] Luo, W., Qing-Huang, Xiao-Huang, Hang-Hu, Fu-Zeng & Wang, W. Predicting Breast Cancer in Breast Imaging Reporting and Data System (BI-RADS) Ultrasound Category 4 or 5 Lesions: A Nomogram Combining Radiomics and BI-RADS. Scientific Reports. **9** (2019)

[12] Acharya, C., Hsu, D., Anders, C., Anguiano, A., Salter, K., Walters, K., Redman, R., Tuchman, S., Moylan, C., Mukherjee, S., Barry, W., Dressman, H., Ginsburg, G., Marcom, K., Garman, K., Lyman, G., Nevins, J. & Potti, A. Gene expression signatures, clinicopathological features, and individualized therapy in breast cancer.. JAMA. **299**

13 pp. 1574-87 (2008)

[13] Zhu, Y., Tzoras, E., Matikas, A., Bergh, J., Valachis, A., Zerdes, I. & Foukakis, T. Expression patterns and prognostic implications of tumor-infiltrating lymphocytes dynamics in early breast cancer patients receiving neoadjuvant therapy: A systematic review and meta-analysis. Frontiers In Oncology. **12** (2022)

[14] Lafourcade, A., His, M., Baglietto, L., Boutron-Ruault, M., Dossus, L. & Rondeau, V. Factors associated with breast cancer recurrences or mortality and dynamic prediction of death using history of cancer recurrences: the French E3N cohort. BMC Cancer. **18**, 171 (2018,2)

[15] Bode, A. & Dong, Z. Precision oncology- the future of personalized cancer medicine?. NPJ Precision Oncology. **1** (2017), https://api.semanticscholar.org/CorpusID:3279548

[16] Zuo, C., Qian, J., Feng, S., Yin, W., Li, Y., Fan, P., Han, J., Qian, K. and Chen, Q. Deep learning in optical metrology: a review. Light: Science and Applications. **11**, 39 (2022,2)

[17] Zhang, Y., Satapathy, S., Guttery, D., Górriz, J. and Wang, S. Improved Breast Cancer Classification Through Combining Graph Convolutional Network and Convolutional Neural Network. Information Processing and Management. **58**, 102439 (2021,3)

[18] Vulli, A., Srinivasu, P., Sashank, M., Shafi, J., Choi, J. and Ijaz, M. Fine-tuned DenseNet-169 for breast cancer metastasis prediction using FastAI and 1-cycle policy. Sensors. **22**, 2988 (2022)

[19] Jha, A., Verma, G., Khan, Y., Mehmood, Q., Rebholz-Schuhmann, D. and Sahay, R. Deep Convolution Neural Network Model to Predict Relapse in Breast Cancer. 2018 17th IEEE International Conference On Machine Learning And Applications (ICMLA). pp. 351-358 (2018)

[20] Ollech, D. & Webel, K. A random forest-based approach to identifying the most informative seasonality tests. SSRN Electron. J.. (2020)

[21] Rigatti, S. Random Forest.. Journal Of Insurance Medicine. **47 1** pp. 31-39 (2017)

[22] De-Feng, Liu, Z., Wang, X., Chen, Y., Chang, J., Wei, D. & Jiang, Z. Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach. Construction And Building Materials. **230** pp. 117000 (2020)

[23] Ting, K. & Witten, I. Stacked Generalizations: When Does It Work?. International Joint Conference On Artificial Intelligence. (1997)

[24] Lundberg, S. & Others SHAP (SHapley Additive exPlanations) Documentation. (2023), https://shap.readthedocs.io/en/latest/index.html, Accessed: 2024-05-27

[25] Ben Jabeur, S., Stef, N. & Carmona, P. Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. Comput. Econ.. **61**, 715-741 (2023,2)

[26] Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. Proceedings Of The 32nd International Conference On Machine Learning. **37** pp. 448-456 (2015,7,7), https://proceedings.mlr.press/v37/ioffe15.html

[27] Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus). ArXiv Preprint ArXiv:1606.08415. (2016)

[28] Ravanelli, M., Brakel, P., Omologo, M. & Bengio, Y. Light gated recurrent units for speech recognition. IEEE Transactions On Emerging Topics In Computational Intelligence. **2**, 92-102 (2018)

[29] Ben Jabeur, S., Stef, N. & Carmona, P. Bankruptcy prediction using the XGBoost algorithm and variable importance feature engineering. Comput. Econ.. **61**, 715-741 (2023,2)

[30] Scarselli, F., Gori, M., Tsoi, A., Hagenbuchner, M. & Monfardini, G. The graph neural network model. IEEE Transactions On Neural Networks. **20**, 61-80 (2008)

[31] Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. Advances In Neural Information Processing Systems. **30** (2017)

[32] Lingfei Wu, Peng Cui, Jian Pei, and Liang Zhao. Graph Neural Networks: Foundations, Frontiers, and Applications.Springer, 1st edition, Jan. 2023, Singapore, Singapore.

[33] Curtis, C., Shah, S., Chin, S., Turashvili, G., Rueda, O., Dunning, M., Speed, D., Lynch, A., Samarajiwa, S., Yuan, Y., Graf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., Mckinney, S., Group, M., Langerod, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Borresen-Dale, A., Brenton, J., Tavare, S., Caldas, C. & Aparicio, S. The genomic and transcriptomic architecture of 2000 breast tumours reveals novel subgroups. Nature. **486** pp. 346-352 (2012)

[34] Gao, J., Aksoy, B., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C. & Schultz, N. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal.. **6**, 11 (2013,4)

[35] Lee, M. GELU activation function in deep learning: A comprehensive mathematical analysis and performance. (2023,5)

[36] Wang, M., Zheng, D., Ye, Z., Gan, Q., Li, M., Song, X., Zhou, J., Ma, C., Yu, L., Gai, Y., Xiao, T., He, T., Karypis, G., Li, J. & Zhang, Z. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks. ArXiv Preprint ArXiv:1909.01315. (2019)

[37] Jha, A., Verma, G., Khan, Y., Mehmood, Q., Rebholz-Schuhmann, D. & Sahay, R. Deep Convolution Neural Network Model to Predict Relapse in Breast Cancer. 2018 17th IEEE International Conference On Machine Learning And Applications (ICMLA). pp. 351-358 (2018)

[38] Akensert, K. Graph attention network (GAT) for node classification. (2021), https://keras.io/examples/graph/gat_node_classification/

[39] Presekal, A., Ştefanov, A., Rajkumar, V. & Palensky, P. Attack Graph Model for Cyber-Physical Power Systems Using Hybrid Deep Learning. IEEE Transactions On Smart Grid. **14**, 4007-4020 (2023)

[40] Gao, J., Lyu, T., Xiong, F., Wang, J., Ke, W. & Li, Z. MGNN: A multimodal graph neural network for predicting the survival of cancer patients. Proceedings Of The 43rd International ACM SIGIR Conference On Research And Development In Information Retrieval. pp. 1697-1700 (2020)

[41] Ji, C., Wang, Y., Ni, J., Zheng, C. & Su, Y. Predicting miRNA-disease associations based on heterogeneous graph attention networks. Frontiers In Genetics. **12** pp. 727744 (2021)

[42] Gupta, S. Prediction time of breast cancer tumor recurrence using Machine Learning. Cancer Treatment And Research Communications. **32** pp. 100602 (2022).

Abdullah Basaad received the B.S. degree in computer and data science from Birmingham City University, U.K. in 2021, and the P.G.Cert. in research practice at Birmingham City University, U.K. in 2022. His B.S. project is undergraduate project's platform. He is currently a Ph.D. researcher in the Department of Computing and Data Science at Birmingham City University, U.K, and lectures in introduction into data science, software analyst and design and data structures and algorithms. His Ph.D. is in Enhancing Interpretability and Explainability Modality in Deep Learning - Graph Neural Network Approach and Large Language Model. His areas of expertise include machine learning, artificial intelligence, graph neural network, software engineering and analytics.

Shadi Basurra received the B.Sc. degree (Hons.) in computer science from Exeter University, U.K., the M.Sc. degree in distributed systems and networks from Kent University, Canterbury, U.K, and the Ph.D. degree from the University of Bath in collaboration with Bristol University. He is currently a professor in computer science with Birmingham City University, UK. After his PhD degree, he was with Sony, where he was developing goal decision systems. He has taught postgraduate and undergraduate courses in computer science and networking. He has published a number of peer-reviewed scientific articles in international conferences and journals. His research interests include multiagent systems, game theory, multi-objective optimization, machine learning in the Internet of Things, energy efficiency in smart buildings, emulation of mobile ad hoc networks, nature-inspired computing, and social networks. He received The Yemen President National Science Prize, in 2010, the Best Presentation at Meeting of Minds Bath, in 2012, the MEX Scholarship, in 2013, the Ph.D. Scholarship from Toshiba Ltd, Great Western Research, and Yemen Government, in 2009, and various academic grants.

Dr Edlira Vakaj is leading the Natural Language Processing AI Lab and is associate professor at the Computing and Data Science department at Birmingham City University. She conducts research in multidisciplinary projects focusing on Semantic Web, Knowledge Graphs, AI, and Semantic Data Spaces. Edlira is the principal investigator of the ACCORD Horizon project and engaged in several European and UK-funded projects of various domains where Semantic Web Technologies are applied, such as Renew-able Energy (FP7 RENESENG), Industrialised Construction and Industry 4.0 (Innovate UK DfMA, Knowledge Transfer Partnership), Higher Education and Youth (Erasmus + Capacity Building, Learning mobility of individuals, Cooper-ation for innovation and the exchange of good practices action). She is acting as the Academic Lead of several Knowledge Transfer Partnership(KTP) projects, such as the 4Net KTP and Hadley Group KTP. She is a Professional Member of the British Computer Society, a Fellow and AURORA of Higher Education Academy UK, and a Marie Curie Experienced Researcher Fellow from the University of Surrey. Edlira is an active member of various communities, such as The Alan Turing Knowledge Graph Community, Linked Building Data, Knowledge, Graph Creation, and Common Action.

Dr Aleskandarany is a senior lecturer of biomedical science, within the School of Human Sciences, University of Derby. He currently lead the undergraduate programmes of BSc in Biomedical Science and BSc in Biomedical Health. In these programs, student-focused curricula are taught with a specific emphasis to enhance graduate employability. He is a Doctor of Medicine and a trained Diagnostic Surgical Pathologist and joined the University of Derby after completing more than two decades in higher education institutions within the UK and Egypt. His teaching within the modules he lead, and deliver is focused mainly on enhancing student engagement via active learning strategies, hands-on practical sessions, and research informed course content to improve students' learning outcomes. These have been furthered by his expertise in diagnostic surgical pathology and my postgraduate qualification in medical education.

Dr Mohammed Abdelsamea is a senior lecturer in computer science (machine learning and computer vision) at the University of Exeter, and a fellow of the British Higher Education Academy (HEA). Before joining Exeter University, he was a senior lecturer in data and information science at Birmingham City University, where he was the leading member of the computer vision research team. Dr Abdelsamea also worked for the School of Computer Science at Nottingham University, Mechanochemical Cell Biology at Warwick University, Nottingham Molecular Pathology Node (NMPN), and Division of Cancer and Stem Cells both at Nottingham Medical School, as a Research Fellow. He was awarded a PhD in Computer Science and Engineering (with a Doctor's Europaeus degree) from Scuola IMT Alti Studi Lucca, in Italy. Throughout his career, He had the privilege of collaborating with diverse teams of experts in fields ranging from biology and geology to entomology, pathology, engineering, and computer science. These enriching experiences have taken me to various corners of the world, including Egypt, Singapore, Italy, and the United Kingdom. His current research interests are concerned with the development of novel artificial intelligence (statistical machine learning and deep learning) solutions, with the overall ambition to assist human investigation in healthcare and data science applications. More precisely, Dr Abdelsamea is most interested in carrying out research on different theoretical foundations in computer vision and machine learning.

• • •