Graph Neural Network and its application in Inverse Participation Ratio calculation in Complex Networks

Shikharkya Deb* Chandrakala Meena†

Abstract

The network topologies can be widely categorized into three distinct types spanning across diffused or delocalized, weakly localized and strongly localized.The study in this paper revolves around investigating the steady state behavior of linear dynamical system .Neural networks have outperformed in the euclidian domain hence followed by such success the recent years have been into developing neural networks for graphs and data supported on graphs.At each layer of the neural network the graph is leveraged as a parameterization to capture detail at the node level with reduced number of paameters and computational complexity.Graph Neural Networks are information processing architectures tailored to these graph signals and made of stacked layers of graph convolutional filters with non linear activation functions.Graph convolutional endows GNNs with invariance to permutations of the graph nodes , labels. In this paper we dabble around building three different frameworks which include Graph Convolutional Neural Networks **GCNN**, Auto Regressive Moving Average **ARMA** and the Graph Convolutional Attention Networks **GCAT** which are dedicated to predict the inverse participation ratio values provided the actual ones are provided as input into the frameworks while the level of validation accuracy vary for each the paper establishes both numerically and analytically the superiority of both GCAT and ARMA over GCNN in ipr prediction.

Index Terms:- graph shift operator,auto regressive moving average, graph convolutional neural network, graph convolutional attention network, mean squared error, cross entropy loss, pooling

I Introduction

Diffused or delocalized, weakly localized and strongly localized are the three distinct states from information propagation in complex networks.Information condensed over a single component refer to strong localization while for few components refer to weak localization and for information diffusing evenly refer to the delocalization.In solid state physics localization and its presence or absence influences the properties of molecules or materials.The interactions within real world complex systems are often nonlinear[1].Complex Networks on the other hand have a wide range of application from providing accurate GPS services in satellites to epidemic spread, all these network models are primarily dependent on two factos - i) Network Structure and ii)Information propagation among its components.References can be drawn from the following sources[2–6].Development of understanding of epidemic spread awards one to draw insight into the development of strategies in reducing the pace of transmission and providing with ample time for vaccine development.[7–11] Network Centrality measure, spectral partitioning and development of approximation algorithms are the fundamental network problems.Investigation of the localization and delocalization behavior of complex networks is important for gaining insight into such fundamental network problems.In some cases non-linear systems can be solved by their transformation into linear counterparts near fixed points. Hence , understanding linear systems and their solutions is an essential first step toward comprehending more complex nonlinear dynamical systems.Numerous recent works suggests Graph Neural Networks(GNN) as means of translating to graphs followed by the success several convolutional and recurrent neural networks have attained at learning in time and space.GNNs at their core are first concretized followed by means of recursive neighboring label aggregations combined with pointwise non linearities[12–28].The study in this paper at its core is motivated by the work of professor priyoyuti prodhan in his study in predicting steady state behaviour in complex networks using GNN[**29**] and also by the work of Alejandro Ribeiro et al. in his study on edge varying graph neural networks [30].

In this study we develop Graph Neural Network (GNN) Architecture to identify the behavior of linear dynamical states on complex networks.Datasets for individual networks across topologies which include star,circle,wheel and scalefree and based on their feature weights training labels are derived from the inverse participation ratio (IPR) value of the principal eigen vector (PEV) of the network adjacency matrices. The GNN model takes in the network structure as input and predicts IPR values through a regression operation based on multi-layer perceptron (MLP), enabling the identification

*Harish Chandra Research Institute , HBNI
†Indian Institute of Science Education and Research , Pune

of graphs into their respective linear dynamical states.Our all the three models overall and also among themselves outperform each other in identifying different states and is particularly effective across varying sized networks.The key advantage of using GNN is its transferability, referring to its ability in getting trained on smaller networks and generalizing well to larger networks while testing.An analytical framework to understand the explainability of our model has been provided for establishment of the superiority of the models over each other.The models are also pretty well compatible and applicable to real world data sets.

II Objective

Consider a linear dynamical process M taking place on unknown network structures represented as $G = (V, E)$ where $V = (v_1, v_2, v_3, ..., v_n)$ is the set of vertices(nodes),$E = (v_i, v_j)|v_i, v_j \epsilon V$ is the set of edges(connections).The degree of a node adjacent to it, which is given by $\Sigma_{j=1}^{n} a_{ij}$ where a_{ij} is the adjacency matrix element.The links in G represent dynamic interactions whose nature depends on context.For instance,a_{ij} captures a interaction between individuals i and j. The linear dynamics of node i is represented by :-

$$\frac{dx_i(t)}{dt} = \gamma x_i(t) + \zeta \sum_{j=1}^{n} a_{ij} x_j(t) \tag{1}$$

where $x_i(t)$ is the dynamical term, the second term captures the neighboring interactions at time t and γ, ζ are the model parameters of the linear dynamical system.In matrix notation equation (1) can be expressed as :-

$$\frac{dx(t)}{dt} = Bx(t) \tag{2}$$

Where $x(t) = (x_1(t), x_2(t), x_3(t).....x_n(t))^T$, $B = \gamma I + \zeta A$ is the transition, **A** is the adjacency and **I** is the identity matrices, respectively, the steady state behavior of the (x^*) of the linear dynamical system can be found as :-

$$x(t) = e^{Bt} x(0) \rightarrow x^* \sim u_1^B \tag{3}$$

where u_1^B is the PEV of B. We consider $B = \mathbf{R}^{n \times n}$ is diagonalizable , $B = U\Lambda U^{-1}$ and $UU^{-1} = I$ where columns of U are the eigenvectors $(u_1^B, u_2^B, u_3^B, ..., u_n^B)$ of B and having n distinct eigenvalues which can be defined as $(\lambda_1^B, \lambda_2^B, \lambda_3^B, ..., \lambda_n^B)$.$x(0)$ being an initial state , it can be defined as a combination of the eigenvalues and eigenvectors of **B**, and therefore it can be written as :-

$$x(t) = e^{Bt}[c_1(0)u_1^B + c_2(0)u_2^B + c_3(0)u_3^B + ... + c_n(0)u_n^B]$$

$$x(t) = Ue^{\Lambda t}U^{-1}[c_1(0)u_1^B + c_2(0)u_2^B + ... + c_n(0)u_n^B] \tag{4}$$

$$x(t) = Ue^{\Lambda t}U^{-1}Uc(0) \tag{5}$$

$$x(t) = Ue^{\Lambda t}c(0) \tag{6}$$

$$x(t) = c_1(0)e^{\lambda_1^B t} + c_2(0)e^{\lambda_2^B t} + ... + c_n(0)e^{\lambda_n^B t} \tag{7}$$

$$x(t) = \Sigma_{i=1}^{n} c_i(0)e^{\lambda_i^M t u_i^M} \tag{8}$$

Here $c(0) = (c_1(0), c_2(0), c_3(0), ..., c_n(0))^T$ and the expression e^{Bt} is :-

$$e^{Bt} = I + Bt + \frac{(Bt)^2}{2} + \frac{(Bt)^3}{3} + ... \tag{9}$$

$$e^{Bt} = I + U\Lambda U^{-1}t + \frac{U\Lambda U^{-1}tU\Lambda U^{-1}t}{2!} + \frac{U\Lambda U^{-1}tU\Lambda U^{-1}tU\Lambda U^{-1}t}{3!} + ... \tag{10}$$

$$e^{Bt} = U[I + \Lambda t + \frac{(\Lambda t)^2}{2!} + \frac{(\Lambda t)^3}{3!} + ...]U^{-1} \tag{11}$$

$$e^{Bt} = Ue^{\Lambda t}U^{-1} \tag{12}$$

For $t \rightarrow \infty$, it can be approximated that equation (8) as :-

$$x^* \sim c_1(0)e^{\Lambda_1^{Bt}} \sim u_1^B \tag{13}$$

Since the largest eigenvalue dominates , λ_1 the principal eigenvector will decide the steady state behavior of the system.The localization and delocalization is quantified using the inverse participation ratio , which is the sum of the fourth power of state vector entries and is calculated using :-

$$y_{x*} = \frac{\Sigma_{i=1}^{n} x_i^{*4}}{[\Sigma_{i=1}^{n} x^{*2}]^2} \tag{14}$$

where x_i^* is the ith component of $x^* = (x_1^*, x_2^*, x_3^*...., x_n^*)^T$ and $\Sigma_{i=1}^{n} x_i^{*2}$ is the normalization term.

To understand the category of dynamical behavior of the states $x*$ a formulation can be devised for a threshold scheme for IPR identification within the boundary [1/n,1).

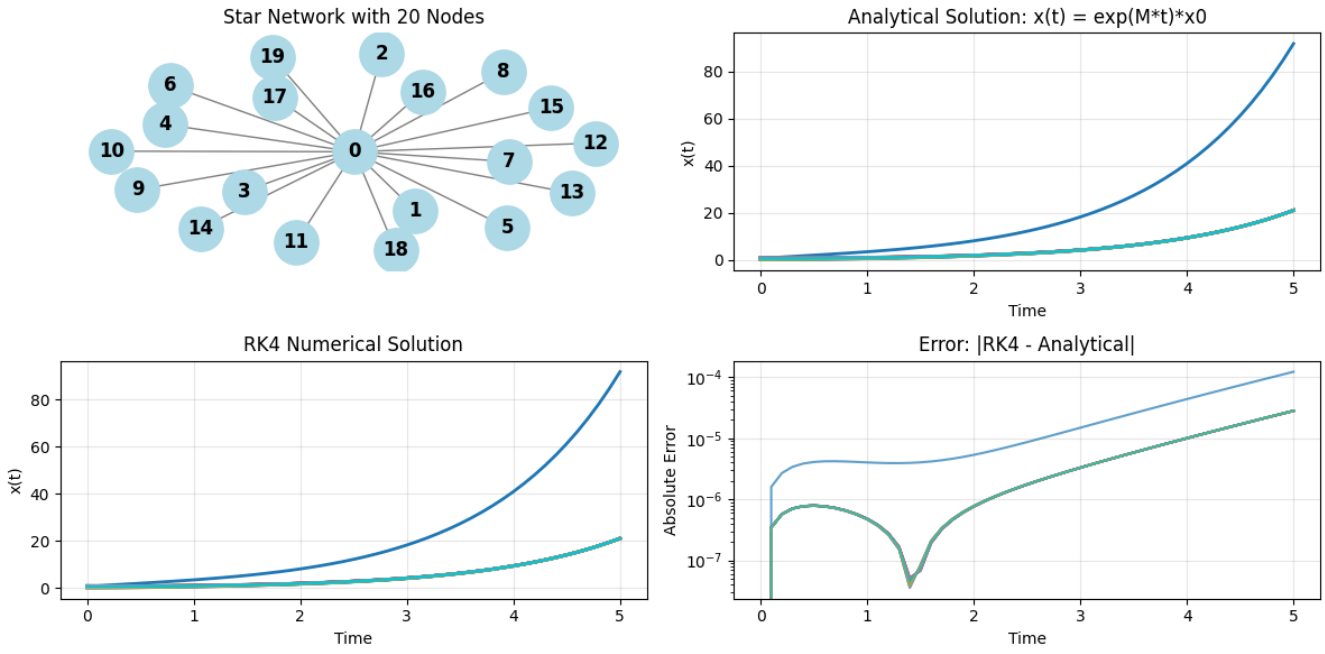

Figure 1: The time series analysis of dynamical equation for star network using RK-4 method

The defination includes two threshold values which are η_1 and η_2 with an additional parameter δ for flexibility in the boundaries.

Delocalized Region (r_1):-

$$r_1 = (y \quad \epsilon[1/n, 1) \leq \eta_1 - \delta) \tag{15}$$

Weakly localized region (r_2):-

$$r_2 = (y \quad \epsilon[1/n, 1) | \eta_1 - \delta < y < \eta_2 + \delta) \tag{16}$$

Strongly Localized Region (r_3):-

$$r_3 = (y \quad \epsilon[1/n, 1) | y \geq \eta_2 + \epsilon]) \tag{17}$$

The entire problem revolves around finding out a target value the IPR, the entire task is of graph regression for a set of graphs $(G_i)_{i=1}^N$ where each $G_i = (V_i, E_i)$ consists a set of nodes V_i and a set of edges E_i, where each graph is represented by its adjacency matrix A_i.For each graph G_i there is a target IPR value, $y^{(i)} \epsilon \mathbf{R}$ and for each node $v \epsilon V$ in G_i, there is an associated feature vector $h_j^{(i)} \epsilon \mathbf{R}^{|\mathbf{V_i}| \times \mathbf{d}}$ be the node feature matrix where $h_j^{(i)}$ is the jth row.The object is to learn a function $f : G \to \mathbf{R}$ such that $f(G_i) \approx y^{(i)}$.

III Architectures and Methodologies

The function to be studied is parameterized using a Graph Convolutional Network in our model. The prediction for IPR in each of our graphs from G_i is denoted by $y_i = f(G_i, \alpha, \beta)$, where the model parameters α and β are learned by minimizing a loss function between the predicted values $\hat{y}^{(i)}$ and the true target values $y^{(i)}$. For classification tasks, the Cross Entropy loss is used, while for IPR value regression, the Mean Squared Error loss function is employed.

Cross Entropy Loss: This loss function is particularly useful for classification problems with C classes. If provided, the optional argument **weight** should be a 1D Tensor assigning weight to each class, which is especially beneficial for handling unbalanced training sets[31].

The *input* is expected to contain unnormalized logits for each class (which do not need to be positive or sum to 1). The input can be a Tensor of size C for unbatched input, or $(minibatch, C)$ or $(minibatch, C, d_1, d_2, \ldots, d_K)$ with $K \geq 1$ for higher-dimensional inputs, such as per-pixel cross entropy loss computation in 2D images.

The target should contain class indices in the range $[0, C)$. If *ignore_index* is specified, this loss also accepts this class index. The unreduced loss (with reduction set to 'none') can be described as: For ($y_n \neq$ ignore_index)

$$l(x, y) = L = (l_1, \ldots, l_N)^T, \quad l_n = -w_n \log \frac{\exp(x_{n,y_n})}{\sum_{c=1}^{C} \exp(x_{n,c})} \cdot \mathbf{I} \tag{18}$$

where x is the input, y is the target, w is the weight, C

is the number of classes, and N spans the minibatch dimension as well as $d_1, \ldots, d_k$ for the K-dimensional case.

The MSE (Mean Squared Error) loss function is defined as:[32]

$$L(\alpha) = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})^2 \tag{19}$$

III.1 GCNN Architecture

Graph Convolutional Neural Networks have proven to be highly successful over time in learning representations for graph data with variants quite prominent as found in these references[16–18, 33].The GCNN appears in [15]where the convolutions are defined as pointwise operators in the laplacian's spectrum.Numerical instability and cost is a major factor in spectral decompositions, the reference in [16] performs approximations for the spectral convolutions with Chebyshev polynomial on the laplacian matrix.Meanwhile graph convolutional filters has been developed as polynomials on a matrix representations of a graph[34–40], besides the very nature of the convolutional filters forms the basis for our proof in later parts of paper to prove the superiority of other architectures compared to GCNN.Considering a weighted graph, G with the vertex set $V = (1, \ldots.., N)$, edge set being $E \subseteq V \times V$ composed of $|E| = M$ ordered pairs (i,j), and the weight function $W : E \to \mathbf{R}$.For each node i, define the neighborhood $N_i = (j : (j,i) \in E)$ as the set of each node connected to i and let $N_i = |N_i|$ denote the number of elements in this set.The graph G has a graph shift matrix associated with it $\mathbf{S} \in \mathbf{R^{N \times N}}$ whose sparsity pattern matches that of the edge set , i.e. entry $S_{ij} \neq 0$ when $(j,i) \in E$ or when $i = j$.The vertex set graph signals are $X = [x_1, x_2, \ldots...., x_n]^T \in \mathbf{R^N}$ in whcih component x_i is associated with node $i \in V$.
$\mathbf{S}$ captures the sparsity and the locality of the relationship between components of a signal x supported on G. It is then natural to take the shift operator as the basis for defining filters for graph signals. $\phi^{(0)}$ be a $N \times N$ diagonal matrix and $\phi^{(1)}, \ldots...., \phi^{(k)}$ be a collection of K matrices sharing the sparsity pattern $\mathbf{I_N} + \mathbf{S}$.Then for $k = 0, \ldots., K$ the sequence of signal $z^{(k)}$ as:

$$z^{(k)} = \Pi_{k'=0}^{k}\phi^{(k')}x = \phi^{(k:0)}x \tag{20}$$

GCNN have been quite successful in learning representations for graph data with prominent variants.All these variants written as GNN architectures in which edge varying components is fixed and given by powers of the shift operator matrix $\phi_l^{(k:0)} = S^k$

$$X_l = \sigma(\Sigma_{k=0}^{K} S^k X_{l-1} A_{lk}) \tag{21}$$

A GNN is defined by the recursive expression:-

$$x_l = \sigma(A_l(S)x_{l-1}) = \sigma(\Sigma_{k=0}^{K}\phi_l^{k:0}x_{l-1}) \tag{22}$$

here the x_0 is the input to the GNN and x_L is the output.To augment the representation power of GNNs,a multiple node feature hass to be added at each layer and this is done by defining matrices $X_l = [x_l^1, \ldots...., x_l^{F_l}] \in \mathbf{R^{N \times F_l}}$ in which each column x_l^f represents a different graph signal at layer l.The features are cascaded through layers and are processed with graph filters and composed with pointwise nonlinearities according to :-

$$X_l = \sigma(\Sigma_{k=0}^{K}\phi_l^{k:0}X_{l-1}A_{lk}) \tag{23}$$

where $A_{lk} \in \mathbf{R^{F_{l-1} \times F_l}}$ is a parameter matrix.It is seen that the above equation represents a stack of graph filters $A_l^{fg}(S)$ applied to a set of F_{l-1} input features x_{l-1}^g to produce a set of F_l output features x_l^f.Assuming $a_{lk}^{fg} = [A_{lk}]_{fg}$ denote the (f,g)th entry of A_{lk} the above equation produces a total of $F_{l-1}F_l$ intermediate features of the form:-

$$u_l^{fg} = A_l^{fg}(S)x_{l-1}^g = \Sigma_{k=0}^{K} a_{lk}^{fg}\phi_l^{fg,(k:0)}x_{l-1}^g \tag{24}$$

A comparison between equations (21) and (24) a particular restriction yields a tensor $A(S)$ with filters of the form:-

$$A(S) = \Sigma_{k=0}^{K} a_{lk}^{fg} S^k \tag{25}$$

For some order K and scalar parameters $a_{l0}^{fg}, a_{l1}^{fg}, \ldots...., a_{lK}^{fg}$.Simplification of the above expression by omitting layer and feature indices we have :-

$$A(S) = \Sigma_{k=0}^{K} a_k S^k \tag{26}$$

The filter above is of the form $\phi^{(0)} = a_0 I_N$ and $\phi^{k:0} = a_k S^k$ for $k \geq 1$.
The appeal for graph convolutional filters of the form (26) is due to the cause that they reduce the number of parameters from $K(M+N)+N$ of the edge varying filters to $K+1$;yielding a computational complexity of the order of $O(KM)$.While the number of parameters can be reduced in a lot of ways , the formulation in (26) is to be noted because it endows the resulting GNN with equivariance to permutation of the labels of the graph.

The **proof** for the above can be stated as follows Denoting the respective graph shift operator of the graphs G and G' as S and S'.For P being the a permutation matrix, S' and x' can be written as $S' = P^T S P$ and $x' = P^T x$ .Then, the output of the convolutional filter in (26) applied to x' is:-

$$u' = \Sigma_{k=0}^{K} a_k S'^k x' = \Sigma_{k=0}^{K} a_k (P^T S P) P^T x \tag{27}$$

By using the properties of the permutation matrix $P^k = P$ and $PP^T = I_N$, the output u' becomes :-

$$u' = P^T(\Sigma_{k=0}^{K} a_k S^k x) = P^T u \tag{28}$$

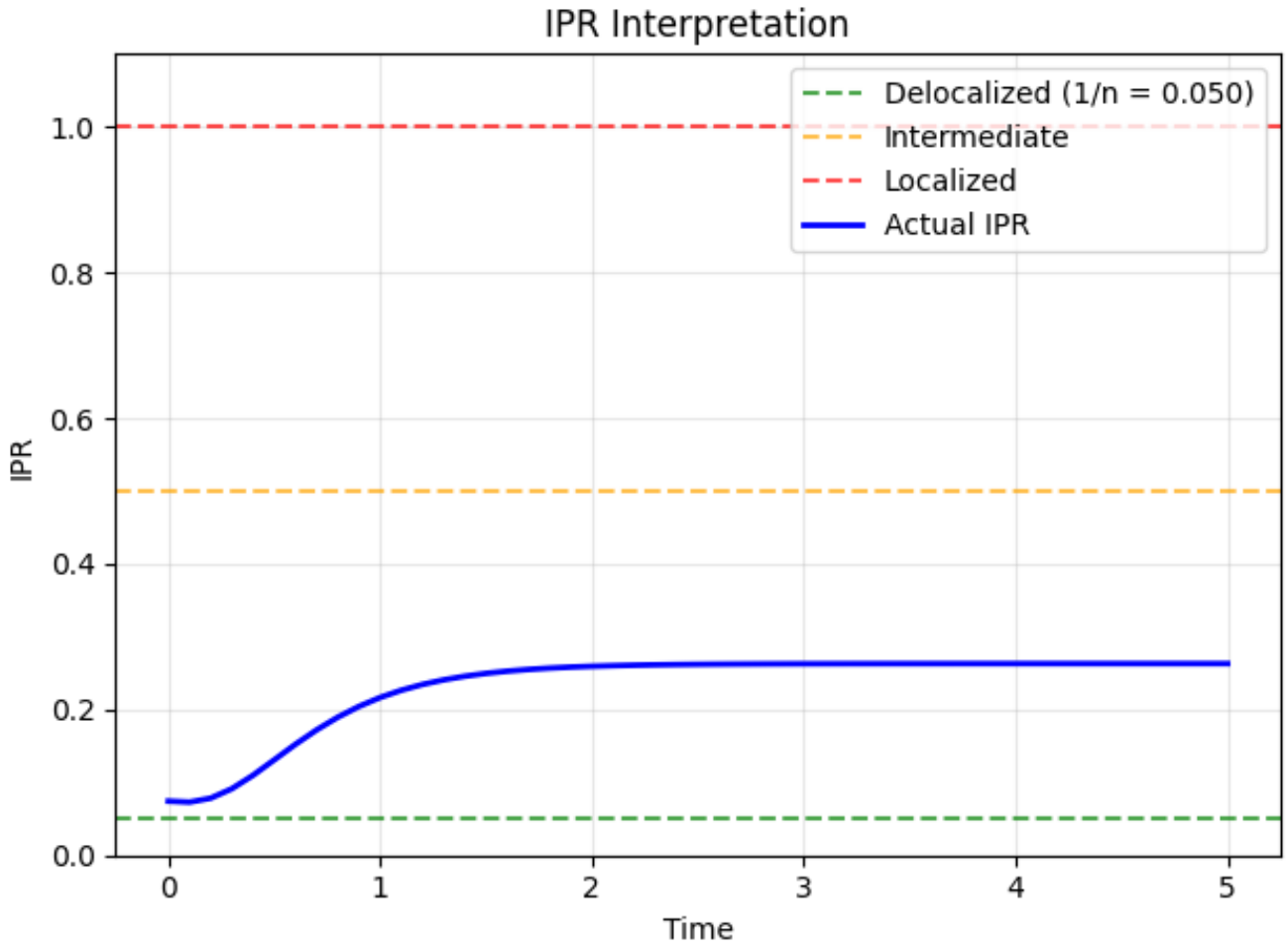

Figure 2: The IPR analysis and segregation into separate localized regions based on the property of the network input for scale free network

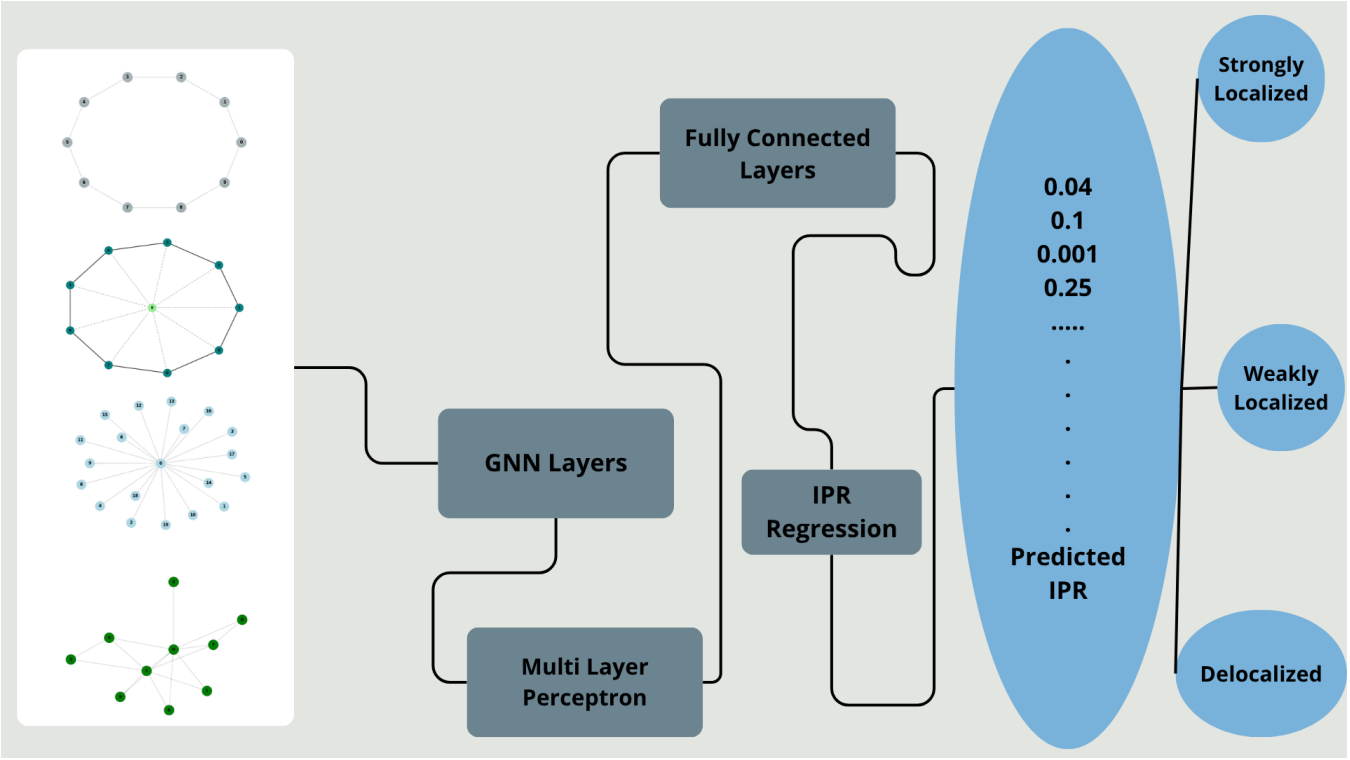

Figure 3: The General architecture of Graph Neural Networks for the regression task over the graphs in this literature

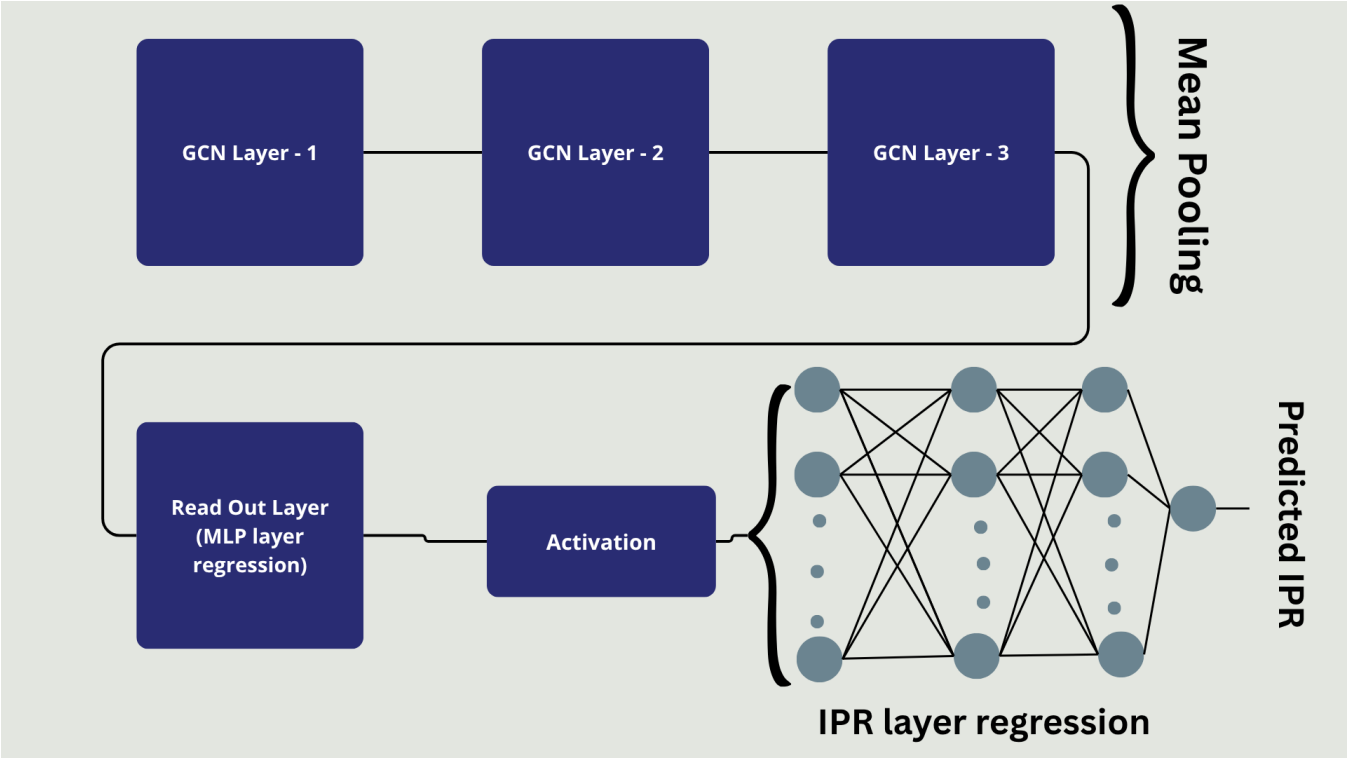

Figure 4: Architecture of the Graph Convolutional Neural Network model.Each node in the model is associated with its features.**1**The first layer of GCN is where the node aggregates and updates its features based on the intermediate neighbor and its features.**2**The second layer of GCN is where the node updates its features by aggregating the messages from neighbors to neighbors and its own. **3** The third layer of GCN aggregates the messages from neighbors to neighbors. **4** The final GCN layer performs the average pooling and then passes on the features as it is.**4** The readout layer or Multi Perceptron Layer performs the regression and classification losses are calculated through Cross Entropy Loss and finally it is passed on to **5** IPR Regression layer which performs the regression of over the IPR values and the error is predicted through the MSE loss calculating function.

which implies the filter output operating on the permuted graph G' with input x' is simply the permutation of the convolutional filter in (26) applied to x.Subsequently since the nonlinearities of each layer are pointwise they implicitly preserve permutation equivariance; hence the output of a GCNN layer is permuted likewise.These permutations will propagate in the cascade of different layers yielding the final permuted output.

The permutation invariance signifies the fact that the output of a GCNN is independent of node labelling.Hence, it also establishes the fact that GCNN exploit internal signal symmetries.For idnetical topological graphs and the nodes promoting identical signal values the GCNN yields identical outputs.References can be drawn from the following[20, 41, 42]

Input Layer:-

The input layer receives a normalized adjacency ($\hat{A}$) and the initial node feature $H^{(0)}$ matrices.The three convolutional layers are stacked together with an application of nonlinear activation function at the end of each layer.Each layer uses the node representation from the previous layer to compute the updated representation of the next layer.The mode of continuation of signal in GCN occurs in the following manner:-

$$H^{(l)} = \sigma(\hat{A}H^{l-1}W^{l-1}), \quad l = [1, 2, 3] \tag{29}$$

where $H^{(0)} \in \mathbf{R^{n \times d}}$ is the intial input feature matrix, and the $W^{(0)} \in \mathbf{R^{d \times k_0}}$, $W^{(1)} \in \mathbf{R^{K_0 \times k_1}}$, $W^2 \in \mathbf{R^{k_1 \times k_2}}$ are the weight matrices for the first, second and third layers, respectively.Hence, $H^{(1)} \in \mathbf{R^{n \times k_0}}$, and $H^{(2)} \in \mathbf{R^{n \times k_2}}$ are the intermediate node features matrices and after three layers of graph convolution, final output node features are represented as $H^{(3)} \in \mathbf{R^{n \times k_3}}$The pooling function is an averaging operation performed before passing it onto the readout layer.

Readout Layer :- The averaging operation is as follows:-

$$z = \frac{1}{n}\Sigma_{j=1}^{n} h_j^{(3)} \tag{30}$$

where $h_j^{(3)} \in \mathbf{R^{1 \times k_2}}$ is the j^{th} row of $H^{(3)} \in \mathbf{R^{n \times k_2}}$ and $z \in \mathbf{R^{k_2}}$, Finally it is past onto the basic neural network in multi layer perceptron to perform regression task over graphs followed by the ipr regression layer and output the predicted IPR value as :-

$$\hat{y} = zW^{(lin)} + b \tag{31}$$

Where $W^{lin} \in \mathbf{R^{k_2 \times 1}}$ is the weight matrix and $b \in \mathbf{R}$ is the bias value,

IV GCNN(Training and Testing Strategy)

IV.1 Training Strategy

The training process for this multi-task GNN is designed to simultaneously optimize for two distinct objectives: a classification task and a regression task. This is achieved through a combined loss function. The model is trained over a specified number of epochs, with each epoch consisting of an iterative loop over mini-batches of data from the training loader.

For each mini-batch, the model's outputs for both classification and regression are obtained. Two separate loss functions are then computed: **CrossEntropyLoss** for the classification task and **MSELoss** (Mean Squared Error) for the Inverse Participation Ratio (IPR) regression task. The total loss is calculated as the sum of these two individual losses, and this combined loss is used to drive the backpropagation and weight update process via the Adam optimizer.

To monitor the model's progress and detect potential overfitting, a validation phase is executed after each training epoch. The model is switched to evaluation mode (**model.eval()**), and its performance is assessed on a separate validation dataset. The validation losses for both tasks, along with the classification accuracy and mean absolute IPR error, are calculated and recorded. This continuous monitoring allows for a detailed view of how the model's ability to generalize to unseen data evolves throughout the training process.

IV.2 Evaluation Strategy

Once the training phase is complete, the model's final performance is evaluated on a completely independent test dataset that it has never seen before. The evaluation is conducted using the **evaluate_model** function, which runs the model in inference mode without any gradient calculations to ensure the assessment is unbiased.

The evaluation strategy provides a comprehensive set of metrics to quantify the model's effectiveness on both of its tasks. For the classification task, the performance is measured using **Accuracy** and the **F1 Score**. For the IPR regression task, a more detailed analysis is performed using multiple metrics, including **Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R^2 Score** (coefficient of determination), and the **Pearson Correlation coefficient**. These metrics collectively provide a robust and holistic understanding of how well the model predicts the true IPR values. The final test results are then printed to the console and visualized in a series of plots, including a scatter plot of predicted vs. actual IPR values, to provide a clear and definitive benchmark of the model's performance.

V Dataset Preparation for GCNN

V.1 Synthetic Star Network Dataset

The model is trained and evaluated on a custom-built synthetic dataset created by the StarNetworkDataset class. The dataset simulates a star-shaped graph where a central node is connected to all other peripheral nodes. This graph structure is defined by an adjacency matrix where the center node has a strong connection to others, while peripheral nodes are not directly connected to each other.

For each data sample, the dataset generates random node features, with a specific modification to the features of the central node to create a clear class distinction. The dataset is multi-task, meaning each sample has two targets:

1. A classification label: This is a binary label determined by the mean of the features on the central node.
2. An Inverse Participation Ratio (IPR) target: This is a real-valued number calculated from the first feature dimension of the nodes. The IPR is a metric used in physics to quantify the spatial localization of a wave function and serves as the regression target for the model.

This approach creates a controlled environment for testing the model's ability to learn from the graph structure and solve both a discrete (classification) and a continuous (regression) problem simultaneously.

V.2 Model Configuration

The GNN model used is a **LocalActivationGNN**, configured to have two distinct output branches for its two tasks. The configuration is **custom-made and defined** through a dictionary with the following parameters with assigned tasks as follows :=This involves a parameter that defines the number of features in each layer of the GNN. The architecture starts with 3 input features and progressively reduces the feature dimensions through layers of size 16, 8 and 4.Further, includes parameters with the responsibilities involving controlling the number of filter taps for each graph convolutional layer, which influences the receptive field of the filters,Specifying the number of hops (or network distance) for the activation function in each layer.Defining the architecture of the Multilayer Perceptron (MLP) branch used for the classification task. It consists of layers with 64 and 32 neurons. Similarly, the architecture for the MLP branch dedicated to the IPR regression task is also defined, with layers of 16 and 8 neurons.The parameter also involves a Graph Shift Operator, which is the adjacency matrix of the star network. This parameter provides the model with the underlying graph structure.The parameter for class sets the number

of output classes for the classification branch.

Pooling :=In Deep Neural Network Architectures there is an increase in the representation dimension with the increase in the stacking of the composing number of layers or with the increase in depth of the network or with the increase in the number of features.To keep the representation dimension under control many architectural solutions tend to resolve this issue by including pooling operation as an intermediate step in between the convolutional filter banks and the nonlinearity.Pooling operation refers to the summarizing of two step operation aimed at reducing dimensionality by first computing local summaries of the signal witin the graph equivalent and then subsampling it.The summarizing operation can be either linear or non-linear.The pooling operation can be majorly of two types - **average pooling** and the **max-pooling**.Average pooling has been considered in the GCNN architecture in this paper incorporated in the final layer of the GCNN.Pooling can also be used in feature extraction at multiple resolutions as the graph window it operates on can be rescaled.Since the study revolves around a range of network topologies the the permutation invariance is preserved during the sub sampling operation.Non linear activation function and pooling operation has been composed into one single localized activation function that preceeds sub sampling which passes onto next layer and the non linearity being the pointwise. [43]

V.3 Synthetic Dataset for Scale Free Network

The model's performance is tested on a synthetic dataset representing a scale-free network, a type of graph where a small number of nodes, known as "hubs," have a very large number of connections, while most other nodes have very few. The dataset is generated using a Barabási-Albert model, a common algorithm for creating such networks.

For each data sample, features are generated randomly for each node. Crucially, a distinct characteristic is introduced by significantly altering the features of the top two hub nodes, making their data values more prominent. The dataset is designed to challenge the model with two tasks:

- Classification: The binary label for each sample is determined by whether the absolute difference between the average feature values of the hub nodes and the non-hub nodes exceeds a specific threshold. This task requires the model to identify and differentiate the unique characteristics of the hub nodes within the graph structure.
- Regression: The model must predict a continuous Inverse Participation Ratio (IPR) value, which in this case is calculated based on the combined magnitude of features across all nodes. This tests the model's ability to learn and predict a global graph property from local node features.

V.4 Synthetic Dataset for Wheel Network

The dataset is built on the wheel network topology, a specific graph structure consisting of a central hub node connected to a ring of perimeter nodes. Each perimeter node is also connected to its two immediate neighbors on the ring, forming a continuous cycle. This unique structure allows for testing the model's ability to learn from both a central, highly-connected node and a more localized, circular subgraph.

The **WheelNetworkDataset** class generates a synthetic dataset with two specific learning objectives. For each sample, the node features are initialized randomly. A specific bias is then introduced by modifying the features of the central hub node and a handful of randomly selected perimeter nodes. This feature manipulation serves as the basis for the two distinct tasks:

- Classification: A binary classification label is assigned based on a comparison between the average feature values of the central hub and the modified perimeter nodes. The model must learn to classify each sample by identifying this key difference between these two sets of nodes, highlighting its capacity to detect localized feature patterns.
- Regression: The dataset also includes a continuous regression target, the Inverse Participation Ratio (IPR). This value is calculated from the feature magnitudes of all nodes, serving as a measure of feature distribution across the entire graph. The model must predict this global property by aggregating information from all nodes and their connections.

This multi-task dataset provides a controlled environment to evaluate the model's performance on a distinctive graph topology, assessing its capacity to perform both node-level feature analysis for classification and graph-level property prediction for regression.

VI ARMA Architecture

The descriptive power of the filter in equation (14) can be increased by increasing the order of K but this sets in complexity by increasing the parameters and the computational costs.It introduces numerical issues associated with higher orders of the matrix S^K.Hence setting in poor interpolatory and extrapolatory features.To counter the such effects we bring in rational rational spectral respsonse as they have better interpolatory and extrapolatory properties than polynomials.More complicated responses can be achieved in both numerator and denominator using rational functions with less learnable parameters.Auto Regressive Moving Average**(ARMA)** filters serve the purpose in such

cases.[44–46]

$$A(S) = \left(I + \Sigma_{p=1}^{P} a_p S^P\right)^{-1} (\Sigma_{q=1}^{Q} b_q S^q) = P^{-1}(S)Q(S) \tag{32}$$

Here $P(S) = \left(I + \Sigma_{p=1}^{P} a_p S^P\right)$ and $Q(S) = \Sigma_{q=1}^{Q} b_q S^Q$.The relationship for the input and output in the ARMA filter can be demonstrated as :-

$$\hat{u} = \left(I + \Sigma_{p=1}^{P} a_p \Lambda^p\right)^{-1} (\Sigma_{q=1}^{Q} a_q \Lambda^q)\tilde{x} \tag{33}$$

It also follows that ARMA filters are also pointwise in spectral domain charecterized by rational spectral respsonse function:-

$$a(\lambda) = \frac{\left(\Sigma_{q=0}^{Q} b_q \lambda_q\right)}{\left(1 + \Sigma_{p=1}^{P} a_p \lambda^p\right)} \tag{34}$$

The partial fraction decomposition of the rational function $a(\lambda)$ in equation (51) provides an equivalent representation of ARMA filters.Let $\gamma = [\gamma_1, \gamma_2, ..., \gamma_p]^T$ be the set of poles and $\beta = [\beta_1, \beta_2, ..., \beta_p]^T$ be the number of corresponding residues and $\alpha = [\alpha_1, \alpha_2, ..., \alpha_K]^T$ be a set of direct terms and equation (51) can be rewritten as :-

$$a(\lambda) = \Sigma_{p=1}^{P} \frac{\beta_p}{\lambda - \gamma_p} + \Sigma_{k=0}^{K} \alpha_k \lambda^k \tag{35}$$

here the α, β and γ are computed from a and b.A graph filter whose spectral response is as in equation (52) is the one in which spectral response λ is replaced by the shift operator variable S.It follows that if α, β and γ are chosen to make (52) and (51)equivalent, the filter in equation (49) is equivalent to:-

$$A(S) = \Sigma_{p=1}^{P} \beta_p (S - \gamma_p I)^{-1} + \Sigma_{k=0}^{K} \alpha_k S^k \tag{36}$$

The ARMA filter training can be trained using α, β and γ instead of training a and b.
For the Jacobi implementation of the single pole filters and circumventing the matrix inverses in equation (53), each single pole filter in (53) is taken into consideration separately and implemented out with the input output relationship:-

$$u_p = \beta_p (S - \gamma_p I)^{-1} x \tag{37}$$

The above expression is equivalent to the linear equation $(S - \gamma_p I)u_p = \beta_p x$, which we can solve iteratively through jacobi recursion.This requires separating the $(S - \gamma_p I)$ into diagonal and off diagonal elements.Beginning by defining the diagonal degree matrix $D = diag(S)$:-

$$S = D + (S - D) = diag(S) + (S - diag(S)) \tag{38}$$

hence we can write $(S - \gamma_p I_N) = (D - \gamma_p I_N) + (S - D)$, which is a decomposition on diagonal terms $(D - \gamma_p I_N)$ and the off diagonal terms $(S - D)$.The jacobi iteration k for (54) is given by the recursive expression:-

$$u_{pk} = -(D - \gamma_p I_N)^{-1}[\beta_p x - (S - D)u_{p(k-1)}] \tag{39}$$

with the initialization of $u_{p0} = x$.To execute this iteration operation we can write an explicit relationship between u_{pk} and x.Inorder to perform such a recursive operation we define the parameterized shift operator:-

$$R(\gamma_p) = -(D - \gamma_p I_N)^{-1}(S - D) \tag{40}$$

and using the above we can write the Kth iterate of the jacobi recursion as :-

$$u_{pk} = \beta_p \Sigma_{k=0}^{K-1} R^k(\gamma_p)x + R^K(\gamma_p)x \tag{41}$$

Signal u_{pk} in (58) for a convergent jacobi recursion relation converges to the output u_p of the single pole filter in (54).Truncating (58) at a finite K yields an approximation in which single-pole filters are written as polynomials on the shift operator $R(\gamma_p)$, single pole filter is approximated as a convolutional filter of order K in which shift operator of graph S is replaced by the shift operator $R(\gamma_p)$ defined in (57).This convolutional filter uses parameters β_p for $k = 0, ..., K-1$ and 1 for $k = K$.
For using Jacobi iterations to approximate all single pole filters in (53) and that we truncate all of these iterations at K, we can write ARMA filters as :-

$$A(S) = \Sigma_{p=1}^{P} H_K(R(\gamma_p)) + \Sigma_{k=0}^{K} \alpha_k S^k \tag{42}$$

where $H_K(R_{\gamma_p})$ is a K order jacobi approximation of the ARMA filter, which as per (58) is given by:-

$$H_K(R(\gamma_p)) = \beta_p \Sigma_{k=0}^{K-1} R^k(\gamma_p) + R^K(\gamma_p) \tag{43}$$

An ARMA GNN has $(2P + K + 1)F^2$ parameters per layer and a computational complexity of order $O(F^2P(MK + N))$.This decomposes as $O(PN)$ to invert the diagonal matrices $(D - \gamma_p I_N)$;$O(PM)$ to scale the non zeros of $(S - D)$ by the inverse diagonal;$O(PKM)$ to obtain the outputs of jacobi ARMA filters (59) of order K.

VII Training and Testing Strategy for ARMA GNN

The training and evaluation process for the ARMA GNN model is designed to handle a multi-task learning problem, simultaneously optimizing for both a node classification task and an Inverse Participation Ratio (IPR) regression task. The training function orchestrates the training loop, which is divided into distinct training and validation phases for each epoch. During the training phase, the model computes a forward pass to generate both a classification output and an IPR prediction. Two separate loss functions—Cross-Entropy Loss for classification and Mean Squared Error for regression—are calculated. These are then combined into a single, weighted total loss using configurable classification weight and ipr weight parameters, allowing for a flexible balance between the two objectives. This total loss is back propagated to update the model's parameters. The validation phase, which runs without backpropagation, is crucial for monitoring the model's generalization performance on unseen data. It tracks validation loss for both tasks and measures classification accuracy and Mean Absolute Error (MAE) for the IPR prediction, providing key indicators for potential overfitting.

Following the training, the evaluation function provides a final, unbiased assessment of the model's performance on a dedicated test set. The model is set to evaluation mode to ensure consistent predictions. The function processes the test data and collects all predictions and actual values for both tasks. It then computes a comprehensive set of performance metrics. For the classification task, it reports standard metrics like accuracy, F1 score, precision, and recall. For the IPR regression task, it provides a full suite of metrics including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R-squared (R 2) score, and Pearson correlation. By returning a single, detailed dictionary, this function facilitates a thorough analysis of the model's success in mastering both the categorical classification and the continuous regression challenges.

VIII Synthetic Dataset Preparation for ARMA

VIII.1 Star Network Dataset

The training and evaluation of the ARMA GNN model are conducted on a synthetic dataset representing a star network topology, a simple graph structure with a central node connected to all other peripheral nodes. This dataset is generated by the StarNetworkDataset class, which creates samples with two distinct learning objectives. For the classification task, a binary label is assigned based on a significant feature perturbation applied exclusively to the central node, challenging the model to recognize and distinguish the hub's unique characteristics. For the regression task, each sample is also associated with a continuous Inverse Participation Ratio (IPR) value, a measure of feature distribution across the graph, which tests the model's ability to predict a global graph property. The model itself is configured as an ARMA GNN, which leverages parameters such as dimNodeSignals, nDenominatorTaps, and nResidueTaps to control the depth and behavior of its graph-filtering layers. It receives the static adjacency matrix of the star network as its Graph Shift Operator (GSO). Finally, the model uses distinct Multi-Layer Perceptrons (MLP) in its final layers—configured by dimLayersMLP for the classification head and dimLayerIPR for the regression head—to process the learned graph representations and output both the categorical and continuous predictions simultaneously.

VIII.2 Wheel Network Dataset

The provided code defines a synthetic dataset for a **wheel network**, a specific graph topology characterized by a central hub node connected to all other nodes, which themselves form a circular rim. This dataset is meticulously prepared to train a model on a multi-task learning problem, encompassing both a classification and a regression task.

For each data sample, features are randomly generated for all nodes. To introduce a key distinction, the features of the central hub node are significantly perturbed by adding a value drawn from a normal distribution. Subsequently, a small, random subset of the perimeter nodes on the rim also has its features modified. The binary label for the classification task is then determined by comparing the average feature value of the hub with the average feature value of the selected rim nodes; if the absolute difference between these two averages exceeds a threshold of 2.0, the label is set to '1', otherwise it is '0'. This forces the model to learn to distinguish the unique properties of the hub from those of the rim. Simultaneously, a continuous **Inverse Participation Ratio (IPR)** value is calculated for each graph based on its feature magnitudes. The IPR, a measure of feature distribution, is

defined by the formula:

$$IPR = \frac{\sum_i |x_i|^4}{(\sum_i |x_i|^2)^2}$$

where x_i represents the features of node i. This IPR value serves as the target for the regression task, challenging the model's ability to predict a global graph property from local node features.

VIII.3 Scale Free Network for IPR Prediction for ARMA GNN

synthetic dataset for a scale-free network, a specific type of graph where the distribution of node degrees follows a power law, meaning a few nodes (hubs) have many connections while most have very few. The network topology itself is generated using the Barabási-Albert model, a standard algorithm for creating such structures. This dataset is meticulously prepared for a multi-task learning problem, encompassing both a classification and a regression task.

For each data sample, features are randomly generated for all nodes. To create a key challenge, the features of the top two most connected nodes (the hubs) are made intentionally distinct by adding a significant, normally distributed perturbation. The binary classification label for each sample is determined by a simple rule: a label of 1 is assigned if the absolute difference between the mean features of the hub nodes and the non-hub nodes exceeds a threshold; otherwise, the label is 0. This forces the model to learn to identify and leverage the unique characteristics of the hubs to make an accurate prediction. In parallel, the model is also tasked with a regression problem. It must predict a continuous Inverse Participation Ratio (IPR) value for each graph. The IPR is a metric calculated from the feature magnitudes of all nodes, quantifying how localized or distributed the features are across the network.

IX GCAT Architecture

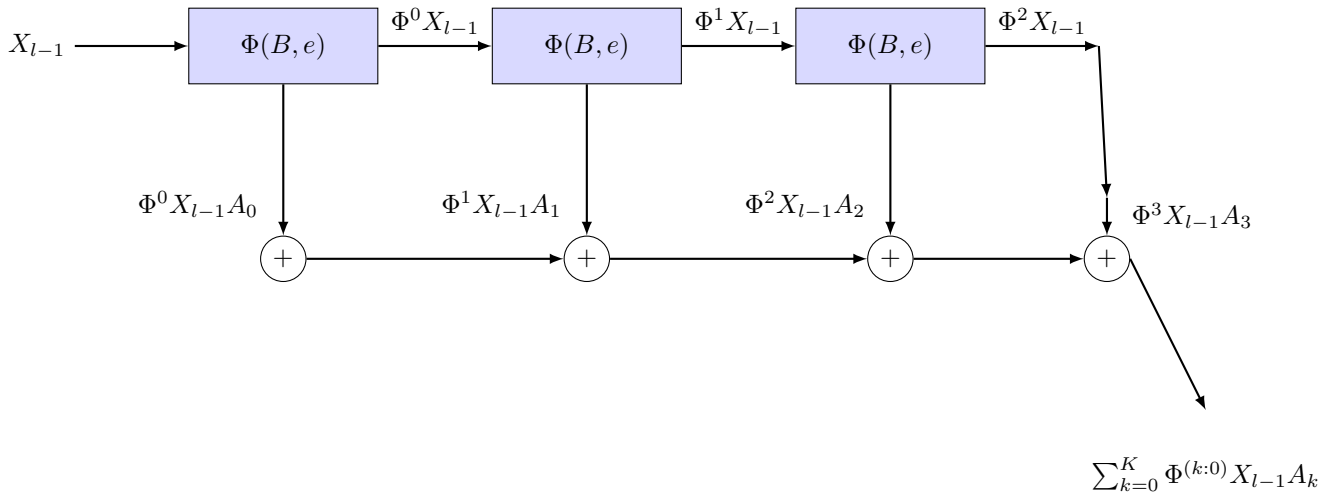

Figure 5: Higher order graph Attention filter :- Graph Convolutional Attention Filter.The input features X_{i-1} are shifted by the same edge varying shift operator $\phi(B,e)$ and weighted by the differential parameter matrices A_k.The edge varying parameters in all $\phi(B,e)$ are parameterized by the same matrix B and vector e following the attention mechanism. [30]

The **GCNN**(Graph Convolutional Neural Networks) parameterize the **EdgeNet** by using fized underlying shift operator S and only learning the filter parameters.This shift operator often estimated from data disjointly from the GNN task, or its sepecific weights may be unknown.And graph attention mechanism comes in handy in these cases where we can parameterize the **EdgeNet** in such a way that one can learn both shift operator weights and the convolutional filter parameters.Here enters the **GCAT**(Graph Convolutional Attention Network) which uses the filter in (21) but they are convolutional in a layer specific matrix $\phi_l = \phi$ which may be different from the shift operator S[24].

$$X_l = \sigma\left(\Sigma_{k=0}^{K} \phi^k X_{l-1} A_k\right), here \quad \phi = \phi_l \quad and A_k = A_{lk} \tag{44}$$

Both the parameters mentioned above are layer specific.Matrix ϕ shares the same sparsity pattern as of S.The above equation defines a GNN same as defined in (23).Matrix ϕ is learned from the features X_{l-1} passed from layer $l-1$ following the attention mechanism.A trainable matrix is considered $B \in \mathbf{R}^{\mathbf{F_{l-1}} \times \mathbf{F_l}}$ and vector $e \in \mathbf{R}^{\mathbf{2F_l}}$, and then

computing the edge scores as :-

$$\alpha_{ij} = \sigma\left(e^T \left[[X_{l-1}B]_i, [X_{l-1}B]_j\right]^T\right) \tag{45}$$

for all edges $i, j \in E$.In the above equation we begin with vector with features X_{l-1} and mix them as per the parameters in the training matrix B producing a collection of graph signals $X_{l-1}B$ in which each node i has F_i features that correspond to the ith row $[X_{l-1}B]_i$ of the product matrix $X_{l-1}B$.The features of node j are concatenated with those of the features of node i producing the resulting vector of $2F_l$ components which is multiplied with the vector e.Thhe product produces the scores α_{ij} after passing through the nonlinearity $\sigma(...)$.$B = B_l$ and $e = e_l$ are global parameters for all scores $\alpha_{ij} = \alpha_{lij}$ and are dependent on the layer index l.The score α_{ij} can be directly passed as an entry for the matrix ϕ but to promote the attention sparsity we pass the score α_{ij} through a local soft maximum operator as :-

$$\phi_{ij} = \frac{e^{\alpha_{ij}}}{\left(\Sigma_{j' \in N_i \cup i} e^{\alpha_{ij'}}\right)} \tag{46}$$

The soft maximum assigns the the edge weights ϕ_{ij} close to 1 to the largest of the edge scores α_{ij} and weights ϕ_{ij} close to 0 for the rest of the edge score α_{ij}.
The computation scores for the α_{ij} depends on the $F_{l-1} \times F_l$ parameters in B and $2F_l$ parameters in e.For the GAT in equation (61) the number of learnable parameters is at most $F^2 + 2F + F^2(K + 1)$, which independent of the number of edges.The graph sparsity is obeyed throughout by ϕ and the computational complexity of (61) and its parameterization is of the order of $O(F(NF + KM))$ lesser compared to ARMA GNN.

The training and evaluation framework for the GCAT GNN employs a sophisticated multi-task learning approach that simultaneously optimizes both classification accuracy and Inverse Participation Ratio (IPR) regression. During training, the model processes graph-structured data through a dual-objective loss function that combines cross-entropy loss for classification tasks with mean squared error loss for IPR prediction, allowing adjustable weighting between these complementary objectives. The validation phase meticulously tracks the interplay between classification performance and localization accuracy, providing real-time feedback on model convergence and generalization capabilities. The comprehensive evaluation module delivers a holistic assessment of model performance by calculating both standard classification metrics (accuracy, F1-score, precision, recall) and specialized regression metrics (MAE, RMSE, R^2) for IPR prediction, enabling researchers to analyze the intricate relationship between graph topological properties and signal localization behavior across diverse datasets and application domains.

X Synthetic Dataset Generation

X.1 Star Network Datset for GCAT

The StarNetworkDataset serves as a sophisticated synthetic benchmark for evaluating graph neural networks, featuring a star topology where a central node connects to all peripheral nodes with weighted edges while maintaining self-loop connections. This dataset generates multi-dimensional node features with intentionally amplified central node characteristics to create clearly separable classes, alongside calculated Inverse Participation Ratio (IPR) values that quantify signal localization patterns across the graph structure. The comprehensive execution block orchestrates a complete machine learning pipeline, initializing the Graph Convolutional Attentional Network with optimized architectural parameters, implementing a dual-objective training regimen that balances classification accuracy with IPR regression performance, and conducting thorough evaluation with detailed metric reporting and visualization. This integrated framework produces publication-ready results including loss curves, accuracy trends, IPR prediction scatter plots, error distributions, and confusion matrices, providing researchers with both quantitative performance measures and qualitative insights into the model's ability to capture complex graph topological properties and signal localization behavior.

X.2 Scale Free Network Dataset

The ScaleFreeDataset represents a sophisticated synthetic benchmark specifically designed to evaluate graph neural networks on complex real-world network topologies. Unlike regular or star networks, this dataset employs a scale-free network structure characterized by a power-law degree distribution, where a few highly connected hub nodes coexist with many sparsely connected peripheral nodes—mimicking the fundamental architecture of social networks, biological systems, and technological infrastructures. The dataset generation process carefully

amplifies feature values around identified hub nodes while maintaining random characteristics in peripheral regions, creating a natural classification task based on the statistical disparity between hub and non-hub node features. Each sample incorporates calculated Inverse Participation Ratio values that capture the localization behavior of signals across the heterogeneous network structure, providing a dual learning objective that tests models' abilities to simultaneously recognize topological patterns and quantify signal concentration. This dataset specifically challenges graph learning algorithms to distinguish between global network properties emerging from local connection patterns and localized signal behaviors that vary across the hierarchy of node connectivity, making it an ideal testbed for evaluating how well neural architectures can capture the intricate interplay between network topology and signal dynamics in complex systems.

X.3 Wheel Network Dataset

The WheelNetworkDataset implements a specialized synthetic benchmark that captures the distinctive topology of wheel graphs, where a central hub node connects to all peripheral nodes forming a complete rim while the peripheral nodes remain unconnected to each other. This dataset generation process carefully engineers node features to emphasize the structural dichotomy between the central hub and the surrounding rim: the hub node receives consistently amplified feature values, while a strategically selected subset of rim nodes receives either strongly positive or negative feature modifications, creating a clear spectral separation between center and periphery. The classification task is explicitly designed to test a model's ability to detect the feature disparity between the central hub and the modified rim nodes, with labels determined by whether the absolute difference between hub and rim feature means exceeds a defined threshold. Simultaneously, each sample incorporates calculated Inverse Participation Ratio values that quantify how signal energy distributes across the wheel's unique topology, particularly testing whether models can recognize the characteristically different localization patterns that emerge between the highly connected hub and the sparsely connected rim nodes. This dual-objective dataset specifically challenges graph neural networks to leverage both the explicit connectivity patterns of wheel graphs and the implicit feature relationships that arise from this distinctive topology, making it particularly valuable for evaluating architectural performance on systems with clear core-periphery structures, such as transportation networks, star-topology computer networks, and certain biological neural structures.

XI Mathematical Privilege for GCNN

The GCNN architecture training works in the following sequence **Forward Propogation** $\rightarrow$ **Loss Calculation** $\rightarrow$ **Backward Propogation** $\rightarrow$ **Weight Update**.The analysis is performed for a single GCNN layer but can be very well extended to multiple layers for analysis easily.The architecture considered for this analysis is a single GCNN layer , a MLP(Multi Layer Perceptron) and a regression layer of FCNN(Fully Connected Neural Network).The training process beneath is described via forward and backward propagation and predicting the ipr value:-

Forward Propagation :-
GCN Layer :- $H^{(1,i)} = \sigma(\hat{A}H^{(0,i)}W)$
GCN Final Layer :- $z^{(i)} = \frac{1}{n_i}\Sigma_{j=1}^{n_i} h_j^{(1,i)}$
Linear layer :- $\hat{y}^{(i)} = z^{(i)}W^{(lin)} + b$
Loss Function :- $L = \frac{1}{N}\Sigma_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})^2$
In the above $\hat{A}^{(i)}$ is the normalized adjacency matrix.$H^{(0,i)}$ is the initial feature matrix.Further $h_j^{(1,i)} = \sigma(\Sigma_{k=1}^{n_i}\hat{A}_{jk}^{(i)}h_k^{(0,i)}W)$ is the feature vector of node j in graph i and the j^{th} row of the updated feature matrix $H^{(1,i)}$.$W^{(lin)}$ and b are the learnable weights of the linear layer.

Backward Propagation :- Inorder to compute the gradients to update weight matrices the chain rule is applied to propagate the error from output layer back through network layers.The gradient of loss with repect to the output of the linear layer is calculated as :-

$$\frac{\partial L}{\partial \hat{y}^{(i)}} = \frac{2}{N}(\hat{y}^{(i)} - y^{(i)}) \tag{47}$$

The gradients of the linear layer are being calculated and the and it is known that each graph i contributes to the overall loss L.We accumulate gradient contributions from each graph when computing gradient of loss with respect to the weight matrix $W^{(lin)}$.To obtain the gradient of the loss with repsect to the weights $W^{(lin)}$, the chain rule is applied :-
Model Architeure Racapitulation

GCN layers:-

$$H^{(l)} = \sigma\left(\hat{A}H^{(l-1)}W^{(l-1)}\right) \quad , here \quad l = 1, 2, 3 \tag{48}$$

Average pooling (readout) after $H^{(3)}$:

$$z = \frac{1}{n}\Sigma_{j=1}^{n} h_j^{(3)} \tag{49}$$

First regression layer(linear transform):

$$a = zW_1 + b_1 \tag{50}$$

Second or final regression layer:

$$\hat{y} = aW_2 + b_2 \tag{51}$$

Loss(MSE):

$$L = \frac{1}{N}\Sigma_{i=1}^{N}\left(\hat{y}^{(i)} - y^{(i)}\right)^2 \tag{52}$$

Where $W_1 \in \mathbf{R}^{k_2 \times d_1}, b_1 \in \mathbf{R}^{d_1}$ and $W_2 \in \mathbf{R}^{d_1 \times 1}, b_2 \in \mathbf{R}$ and $a \in \mathbf{R}^{d_1}, \hat{y} \in \mathbf{R}$ **Forward Pass**

$$z \to a = zW_1 + b_1 \to \hat{y} = aW_2 + b_2 \tag{53}$$

Backward Pass : Gradients for the last regression layer
gradient with respect to $\hat{y}$

$$\frac{\partial L}{\partial \hat{y}^{(i)}} = \Sigma_{i=1}^{N}\frac{2}{N}\left(\hat{y}^{(i)} - y^{(i)}\right)$$

Now considering gradient with respect to W_2 we have the following:-

$$\frac{\partial L}{\partial W_2} = \Sigma_{i=1}^{N}\frac{\partial L}{\partial \hat{y}^{(i)}}.\frac{\partial \hat{y}^{(i)}}{\partial W_2}$$

Since $\hat{y} = aW_2 + b_2$ and W_2 is $d_1 \times 1$ so we have on differentiation :-

$$\frac{\partial \hat{y}^{(i)}}{\partial W_2} = a^{(i)T} \tag{54}$$

hence we have :-

$$\frac{\partial L}{\partial W_2} = \Sigma_{i=1}^{N}\frac{2}{N}\left(\hat{y}^{(i)} - y^{(i)}\right).a^{(i)T}, \quad with \quad shape = d_1 \times 1 \tag{55}$$

for taking gradient with respect to the second adjusting parameter b_2 we have :-

$$\frac{\partial \hat{y}^{(i)}}{\partial b_2} = 1 \quad and \quad \frac{\partial L}{\partial b_2} = \Sigma_{i=1}^{N}\frac{2}{N}\left(\hat{y}^{(i)} - y^{(i)}\right) \tag{56}$$

Next heading on to taking gradient with respect to First Regression Layer Parameters we need $\frac{\partial L}{\partial a_i}$ to proceed backward and we have :-

$$\frac{\partial L}{\partial a^{(i)}} = \frac{\partial L}{\partial \hat{y}^{(i)}}\frac{\partial \hat{y}^{(i)}}{\partial a^{(i)}} \quad and \quad \frac{\partial \hat{y}^{(i)}}{\partial a^{(i)}} = W_2^T \tag{57}$$

Since $\hat{y} = aW_2 + b_2$ so we have :-

$$\frac{\partial L}{\partial a^{(i)}} = \frac{2}{N}(\hat{y}^{(i)} - y^{(i)}).W_2^T \quad with \quad the \quad shape \quad 1 \times d_1 \tag{58}$$

Gradient with respect to W_1

$$\frac{\partial L}{\partial W_1} = \Sigma_{i=1}^{N}\frac{\partial L}{\partial a^{(i)}}.\frac{\partial a^{(i)}}{\partial W_1} \quad , \frac{\partial a^{(i)}}{\partial W_1} = z^{(i)T} \tag{59}$$

$$\frac{\partial L}{\partial a^{(i)}} = \delta_a^{(i)} \tag{60}$$

as $\delta_a^{(i)}$ is $1 \times d_1$, $z^{(i)}$ is $1 \times k_2$, so $z^{(i)T}$ is $k_2 \times 1$ so, $z^{(i)T}\delta_a^{(i)}$ is $k_2 \times d_1$, which matches W_1 shape, hence the final expression stands as :-

$$\frac{\partial L}{\partial W_1} = \Sigma_{i=1}^{N} z^{(i)T}\left[\frac{2}{N}(\hat{y}^{(i)}) - y^{(i)}W_2^T\right] \tag{61}$$

Now working on the gradient with respect to b_1:-

$$\frac{\partial a^{(i)}}{\partial b_1} = I_{d_1} \quad , \frac{\partial L}{\partial b_1} = \Sigma_{i=1}^{N} \frac{\partial L}{\partial a^{(i)}} = \Sigma_{i=1}^{N} \frac{2}{N} (\hat{y}^{(i)} - y^{(i)}) W_2^T \tag{62}$$

Now finally working out with respect to the final readout layer $z^{(i)}$ we have :-

$$\frac{\partial L}{\partial z^{(i)}} = \frac{\partial L}{\partial a^{(i)}} \frac{\partial a^{(i)}}{\partial z^{(i)}} \quad , \frac{\partial a^{(i)}}{\partial z^{(i)}} = W_1^T \tag{63}$$

$$\frac{\partial L}{\partial z^{(i)}} = \left[\frac{2}{N} (\hat{y}^{(i)} - y^{(i)}) W_2^T \right] W_1^T \tag{64}$$

followed by simplification and reordering :-

$$\frac{\partial L}{\partial z^{(i)}} = \left[\frac{2}{N} (\hat{y}^{(i)} - y^{(i)}) \right] (W_1 W_2)^T \tag{65}$$

Thus from the above calculation it is evident that the with the increase in the number of regression layers and applying pooling function before passing it through both of the read outlayers increase chances of the framework to deal with mixed topologies and also deal with complex ipr patterns more efficiently. Since the final layer of GCN is utilized for implementing the average function as the pooling function thus includin it in our calculations we have :-

$$\frac{\partial L}{\partial h_j^{(1,i)}} = \frac{\partial L}{\partial z^{(i)}} \frac{\partial z^{(i)}}{\partial h_j^{(1,i)}} = \Sigma_{i=1}^{N} \frac{2}{N} (\hat{y}^{(i)} - y^{(i)}) (W_1 W_2)^T \frac{1}{n_i} \tag{66}$$

The total gradient with respect to $\mathbf{W}$ accumulates the contribution from all nodes in all graphs:-

$$\frac{\partial L}{\partial W} = \Sigma_{i=1}^{N} \left(\frac{\partial L}{\partial h_j^{(1,i)}} \frac{\partial h_j^{(1,i)}}{\partial W} \right) \quad proceeding \quad with := \frac{\partial h_j^{(1,i)}}{\partial W} \tag{67}$$

The layer output for the i^{th} graph is represented as $H^{(1,i)} = \sigma(\hat{A} H^{(0,i)} W)$.Hence for a single node j in the graph i , its node representation after GCN layer is:-

$$h_j^{(1,i)} = \sigma \left(\Sigma_{k=1}^{n_i} \hat{A}_{jk}^{(i)} h_k^{(0,i)} W \right) = \sigma(q_j^{(i)}), \quad here \quad q_j^{(i)} = \Sigma_{k=1}^{n_i} \hat{A}_{jk}^{(i)} h_k^{(0,i)} W \tag{68}$$

In order to compute $\frac{\partial h_h^{(1,i)}}{\partial W}$, we apply the chain rule and the result is as follows :-

$$\frac{\partial h_j^{(1,i)}}{\partial W} = \frac{\partial h_j^{(1,i)}}{\partial q_j^{(i)}} \frac{\partial q_j^{(i)}}{\partial W} \tag{69}$$

Partial derivative can be calculated with respect to $q_j^{(i)}$ as :-

$$\frac{\partial h_j^{(1,i)}}{\partial q_j^{(i)}} = \sigma'(q_j^{(i)}) \tag{70}$$

$q_j^{(i)}$ is a linear combination of the rows $H^{(0,i)}$ weighted by $\hat{A}_j^{(i)}$.In the matrix notation we can write as :-

$$\frac{\partial q_j^{(i)}}{\partial W} = \hat{A}_j^{(i)} H^{(0,i)} \tag{71}$$

where $A_j^{(i)}$ is the j^{th} row of $\hat{A}^{(i)}$.Now, we can combine the results of the chain rule and get:-

$$\frac{\partial h_j^{(1,i)}}{\partial W} = \sigma'(q_j^{(i)}) \hat{A}_j^{(i)} H^{(0,i)} \tag{72}$$

Now combining equations (19) and (26) into equation (21) we have as follows :-

$$\frac{\partial L}{\partial W} = \left[\frac{2}{N} (\hat{y}^{(i)} - y^{(i)}) \right] (W_1 W_2)^T \frac{1}{n_i} . \sigma'(q_j^{(i)}) \hat{A}_j^{(i)} H^{(0,i)} \tag{73}$$

Finally it is beign observed that the learnable weight matrices are updated using gradient descent as follows :-

$$W \leftarrow W - \eta \frac{\partial L}{\partial W}, \quad W_{lin}^{(1)} \leftarrow W_{lin}^{(1)} - \eta \frac{\partial L}{W_{lin}^{(1)}}, \quad W_{lin}^{(2)} \leftarrow W_{lin}^{(2)} - \eta \frac{\partial L}{W_{lin}^{(2)}}, \tag{74}$$

$$b_1 \leftarrow b_1 - \eta \frac{\partial L}{\partial b_1}, \quad b_2 \leftarrow b_2 - \eta \frac{\partial L}{\partial b_2} \tag{75}$$

The above process includes the parameter η which is the learning parameter.The above process is repeated iteratively as mentioned earlier which includes a forward propagation followd by loss calculation followed by backward propagation and finally a weight update.The above derivation clearly portrays the mathematical upperhand over other GCNN architectures interms of an increased number of learnable weight matrices that helps acclimatize the architecture to complex network topologies and the gradient descent used in updating the parameters suggest a better convergence.

XII Mathematical Superiority of ARMA GNN Vs GCNN

To begin with proving the superiority of ARMA over GCNN revolves around the fact that rational functions have better expressive capacity and shows rapid convergence compared to polynomials[47].To prove the aforementioned we will be heavily relying on weierstrass approximation theorem to show that continuous real valued functions on a compact interval can be unformly approximated by polynomials.In short words polynomials are uniformly dense in $C([a,b]\mathbf{R})$ with respect to sup-norm(the absolute value of the function in the provided domain). The proof of Stone - Weierstrass theorem requires few preliminaries to be established which go on as follows:-

Definition(1) (Unital Sub-Algebra , Separating Points):- Let K be a compact metric space .Considering the banach algebra beneath:-

$$C(K,\mathbf{R}) := [f : K \to \mathbf{R}| \quad f \quad is \quad continuous] \tag{76}$$

equpped with the sup norm ,

$$||f||_\infty := sup_{t\in K}|f(t)| \tag{77}$$

Then the following two cases arise ,

- $A \subset C(K,\mathbf{R})$ is a unital sub-algebra if $1 \in A$ and if $f,g \in A, \alpha,\beta \in \mathbf{R}$ implies that $\alpha f + \beta g \in A$ and $fg \in A$.
- $A \subset C(K,\mathbf{R})$ separates points of K if for all $s,t \in K$ with $s \neq t$, there exists $f \in A$ such that $f(s) \neq f(t)$.

Generalization can follow for the K which can be redefined from compact metric space to being a compact topological space.
Before moving on to the stone-weistrass theorem a proof has to devised for lattice with significance of it being used when defining classes of function in compact metric space.

Definition(2) (Lattice):-A subset $S \in C(K,\mathbf{R})$ is a lattice if, for all $f,g \in S$, , $f\mathrm{V}g \in S$ and $f \wedge g \in S$, where $(f\mathrm{V}g)(\mathrm{x}) := max[f(x),g(x)]$ and $(f \wedge g)(x) := min(f(x),g(x))$, to work on the above lemma the taylor series of the function $\sqrt{(1-g(t))}$, where $0 \leq g(t) \leq 1$, is investigated.It is useful to first study the taylor series of $\sqrt{1-t}$. Formally,

$$\sqrt{1-t} = 1 - \Sigma_{n=1}^{\infty} a_n t^n \tag{78}$$

where, for $n \in N$,

$$a_n = (-1)^{n-1}\binom{\frac{1}{2}}{C_n} = \frac{(-1)^{n-1}}{n!}\Pi_{k=0}^{n-1}\left(\frac{1}{2}-k\right) = 2^{1-2n}\frac{(2n-2)!}{n!(n-1)!}, \quad n \in N, a_n \geq 0 \tag{79}$$

We observe for the ratio test for convergence that :-

$$lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = lim_{n\to\infty}\frac{2n-1}{2(n+1)} = 1 \tag{80}$$

Thus the convergence test proves that it happens in pointwise fashion within the interval $t \in (-1,1)$.Now defining $\psi(t) = 1 - \Sigma_{n=1}^{\infty} a_n t^n$.so $\psi(t)$ converges for $t \in (-1,1)$,Evaluating we obtain :-

$$\psi(t) = -2(1-t)\frac{d\psi}{dt} \implies -\frac{1}{2}\int \frac{dt}{1-t} = \int \frac{d\psi}{\psi} \implies \psi(t) = c\sqrt{1-t}, \quad for \quad some \quad c \in \mathbf{R} \tag{81}$$

Hence evaluating both sides at $t = 0$ gives $c = 1$.Thus, we have $\psi(t) = \sqrt{1-t}$ pointwise for $t \in [0, 1)$.

Now we must show that $\psi(1) = 0$.Hence Stirling's inequality states that :-

$$e^{\frac{7}{8}-n} n^{\frac{1}{2}} < n! < e^{1-n} n^{n+\frac{1}{2}}.$$

Hence, for $n \geq 2$,

$$a_n < 2^{1-2n} \frac{1}{e^{\frac{7}{8}} n^{\frac{1}{2}}} \frac{(2n-2)^{2n-\frac{3}{2}}}{(n-1)^{n-\frac{1}{2}}} < \frac{1}{\sqrt{2} e^{\frac{7}{8}}} \frac{(n-1)^{2n-\frac{3}{2}}}{(n-1)^{2n}} = \frac{1}{\sqrt{2} e^{\frac{7}{8}}} \frac{1}{(n-1)^{\frac{3}{2}}}.$$

So,

$$\sum_{n=1}^{\infty} a_n < \frac{1}{2} + \frac{1}{\sqrt{2} e^{\frac{7}{8}}} \sum_{n=2}^{\infty} \frac{1}{(n-1)^{\frac{3}{2}}} < \infty$$

where the comparison test was used in the first inequality and p-series test in the second. Since $\left(\sum_{n=1}^{k} a_n\right)_{k \geq 1}$ is monotonically increasing and bounded above, $\psi(1)$ exists. Then, by **Abel's Theorem**[1],

$$\psi(1) = 1 - \sum_{n=1}^{\infty} a_n = 1 - \lim_{t \to 1^-} \sum_{n=1}^{\infty} a_n t^n = \lim_{t \to 1^-} \sqrt{1-t} = 0.$$

Hence, again using Abel's Theorem, $\psi(t) = \sqrt{1-t}$ uniformly for $t \in [0, 1]$.

Lemma Let $A \subset C(\mathbf{K}, \mathbf{R})$ be a closed unital sub-algebra. Then

i) if $f \in A$ and $f \geq 0$, then $\sqrt{f} \in A$;

ii) if $f \in A$, then $|f| \in A$;

iii) A is a lattice.

Proof. To see i), consider without restriction

$$0 \leq f \leq 1.$$

Then, we can write

$$f = 1 - g$$

with

$$0 \leq g \leq 1.$$

Using a Taylor series expansion, we can write, formally,

$$\sqrt{f(t)} = \sqrt{1 - g(t)} = 1 - \sum_{n=1}^{\infty} a_n g^n(t)$$

where the coefficients are as in the above remark. The Taylor series approximates $\sqrt{f}$ uniformly in $\| \cdot \|_{\infty}$. Indeed,

$$\left\| \sqrt{f} - \left(1 - \sum_{n=1}^{N} a_n g^n\right) \right\|_{\infty} \leq \sup_{t \in K} \left| \sqrt{f(t)} - \left(1 - \sum_{n=1}^{N} a_n g^n(t)\right) \right|$$

$$= \sup_{z \in [0,1]} \left| \sqrt{z} - \left(1 - \sum_{n=1}^{N} a_n z^n\right) \right|.$$

So, given $\epsilon > 0$, by the uniform convergence of the Taylor series of $\sqrt{1-t}$ on $[0, 1]$, there exists $N \in \mathbf{N}$ such that

$$\left\| \sqrt{f} - \left(1 - \sum_{n=1}^{N} a_n g^n\right) \right\|_{\infty} < \epsilon \quad \forall N \geq \bar{N}.$$

[1]Theorem (Abel). *Let $f(x) := \sum_{n=0}^{\infty} c_n (x - x_0)^n$, and assume that the series converges at $x = x_0 + R$, for some $R \in (0, \infty)$. Then the series is uniformly convergent on $[x_0, x_0 + R]$ and*

$$\lim_{x \to (x_0+R)^-} f(x) = f(x_0 + R) = \sum_{n=0}^{\infty} c_n R^n.$$

That is,

$$\lim_{N\to\infty}\left\|\sqrt{f}-\left(1-\sum_{n=1}^{N}a_n g^n\right)\right\|_\infty=0.$$

Since for all $n\in\mathbf{N}\cup\{0\}$ $g^n\in A$, and A is a sub-algebra, $1-\sum_{n=1}^{N}a_n g^n\in A$. And, since A is closed by hypothesis, $\sqrt{f}\in A$. This completes the proof of i). To prove ii), note that

$$|f|=\sqrt{f^2}.$$

So for $f\in A$, $f.f=f^2\in A$, since A is an algebra.Applying i) to f^2 implies $|f|\in A$.To prove iii) , it is to be noted that,

$$f\wedge g=\frac{1}{2}(f+g-|f-g|)\quad and\quad f\vee g=\frac{1}{2}(f+g+|f-g|)\tag{82}$$

Applying ii) to the above identities gives the desired result.
Its high time now I must disclose the proof of the Stone weistrass theorem which justifies the effeciency of the polynomial nature of GCNN and followed by the superiority of ARMA as a rational function to function better than the GCNN

Stone-Weierstrass Theorem (1937).:-
Let K be a compact metric space and $A\subset C(K,\mathbf{R})$ a unital sub-algebra which separates points of K .Then A is dense in $C(K,\mathbf{R})$.
An equivalent statement is that if A is a closed unital subalgebra which separates points of a compact set K, and $A\subset C(K,\mathbf{R})$,then $A=C(K,\mathbf{R})$.We will proceed using this formulation.
Proof. Let $A\subset C(K,\mathbf{R})$ be a closed unital sub-algebra that separates points of K. Let $\epsilon>0$ be given. For any $f\in C(K,\mathbf{R})$ we will show that there exists $g\in A$ such that

$$\|f-g\|_\infty<\epsilon.$$

Consider points $s,t\in K$. Since A separates points, there exists $h\in A$ such that $h(s)\neq h(t)$. For some $\lambda,\mu\in\mathbf{R}$, define $\tilde{h}:K\to\mathbf{R}$ by

$$\tilde{h}(v):=\mu+(\lambda-\mu)\frac{h(v)-h(t)}{h(s)-h(t)}\quad\forall v\in K.$$

Note that $\tilde{h}\in A$ and $\tilde{h}(s)=\lambda$, $\tilde{h}(t)=\mu$. Thus, for $s\neq t$, there exists $f_{s,t}\in A$ such that

$$f_{s,t}(s)=f(s)$$

and $f_{s,t}(t)=f(t)$.

So, $f_{s,t}$ approximates f in neighbourhoods around s and t. Now, fix s, and let t vary. Put

$$U_t:=\{v\in K|f_{s,t}(v)<f(v)+\epsilon\}.$$

U_t is open because it is the pre-image of an open set. Also, $t\in U_t$. So, $\bigcup_{t\in K}U_t$ is clearly an open cover of K. By the compactness of K, there exists finitely many $t_1,\ldots,t_n\in K$ such that

$$K\subset\bigcup_{i=1}^{n}U_{t_i}.$$

Put

$$h_s:=\min_{1\le i\le n}f_{s,t_i}.$$

Then,

$$\begin{aligned}h_s&\in A\\ h_s(s)&=f(s)\\ h_s&<f+\epsilon.\end{aligned}$$

Now, define

$$V_s:=\{v\in K|h_s(v)>f(v)-\epsilon\}.$$

Note that V_s is open and $K\subset\bigcup_{s\in K}V_s$. By compactness, there exists finitely many $s_1,\ldots,s_m\in K$ such that

$$K\subset\bigcup_{j=1}^{m}V_{s_j}.$$

Put $g = \max_{1 \le j \le m} h_{s_j}$. Then, $g \in A$ and

$$f - \epsilon < g < f + \epsilon.$$

That is,

$$\|f - g\|_\infty < \epsilon.$$

So, A is dense in $C(K, \mathbf{R})$. And, since A is closed, $A = C(K, \mathbf{R})$. **Corollary:-** Let X be a compact subset of $\mathbf{R}^n$ for some $n \in \mathbf{N}$. Then the algebra of all polynomials $P(X, \mathbf{R})$ in the coordinates $x_1, \ldots, x_n$ is dense in $C(K, \mathbf{R})$.**Remarks** The case in which $n = 1$ in the above corollary is the Weierstrass Approximation Theorem.

The Stone weierstrass theorem states that for rational function they can approximate continous functions more effeciently than polynomials, especially functions with → Sharp Transitions, Poles or Singularities and Rapid decay.A formulation for the same can be assumed to be such as :-
For the same number of parameters (P+Q = K), rational functions can achieve lower approximation error:-

$$inf_{a \in P_K} \left|\left| f - a \right|\right|_\infty \geq inf_{r \in R_{P,Q}} \left|\left| f - r \right|\right|_\infty \quad where \quad P + Q = K \tag{83}$$

Here P_K is the set of polynomials of degree K, and $R_{P,Q}$ is the set of rational functions of order (P,Q).The inequality needs to be proved about a fixed degree K.The need of the hour is to establish a faster convergence rate as $K \to \infty$[48].

XII.1 Asymptotic Convergence Rate

Hence what needs to be checked now is the convergene rates for either of the **Polynomials** comapred to the **Rational Functions**
The best uniform approximation erros are defined as :-

- **Polynomial Error:**$E_K(f) = inf_{a \in P_K} \left|\left| f - a \right|\right|_\infty$
- **Rational Function Error:**$R_N(f) = inf_{r \in R_{N,N}} \left|\left| f - r \right|\right|_\infty$ (Here we set P=Q=N so the total degree sums up as K=2N)

The comaprison is for functions f that converge on the interval of approximation that is obey the above theorems.
The proof following is primarily in the field of Approximation Theory and associated with the work of **J.L.Walsh,S.N.Mergelya** and later continued by **A.A.Gonchar,E.B.Saff** and **V.K.Totik**Working initially on the polynomial approximation rate we have:-

XII.1.1 The Polynomial Approximation Rate(Geometric Convergence)

For a function f that is analytic in a closed interval [a,b] the convergence rate of the best polynomial approximation error $E_K(f)$ is geometric.Let D_ρ be the largest ellipse in the complex plane with foci a and b inside which f is analytic , and let ρ be the sum of its semi-axes(the conformal radius of ellipse).The **Geometric Rate for Polynomials** goes as follows for error $E_K(f)$:-

$$E_K(f) \sim \frac{C_f}{\rho^{K+1}} \quad or \quad more \quad formally \quad \to lim_{K \to \infty} sup[E_K(f)]^{1/K} = \frac{1}{\rho} \tag{84}$$

Where C_f is a constant related to f.This is a geometric convergence rate, determined solely by the distance ρ to the nearest singularity of f in the complex plane. The further the nearest singularity is from the interval, the faster the error decreases.

XII.1.2 The Rational Function Approximation(Super Geometric Convergence)

For the same analytic function f, the best rational function approximation error $R_N(f)$ with $P = Q = N$, so the total degree $K = 2N$ can converge much faster.In many cases the best rational approximation error exhibits a double geometric rate of convergence.

$$R_N(f) \sim C'_f . \frac{1}{\rho^{2N+1}} \to lim_{N \to \infty} sup[R_N(f)]^{1/N} = \frac{1}{\rho^2} \tag{85}$$

ρ being the same parameter used for the polynomial converegence.The most advantage arises when the function f has a pole or a strong singularity near the interval.

- **Polynomials P_K:**To approximate a function with a singularity z_0, polynomials must use all their coefficients to fight the singularity's influence on the real axis, leading to the geometric rate limited to $\frac{1}{\rho}$.
- **Rational Functions $R_{N,N}$:**The denominator polynomial $q(x)$ can be specifically designed to cancel or absorb the effect of the pole at z_0.The optimal rational function $r(x) = \frac{p(x)}{q(x)}$ will have a pole of its own very close to z_0.This "pole cancelling " mechanism frees up to $2N$ parameters to approximate the remaining smooth part of the funciton allowing the overall error to converge much faster.

To demonstrate the much faster convergence , we compare the error for the sam etotal number of parameters , $K = 2N$ and for $\rho > 1$(necessary for convergence), the rational error $R_N(f)$ is determined by the term $(\rho^2)^{-N}$, while the polynomial error $E_K(f)$ is determined by the slower term ρ^{-2N}.

$$\frac{E_{2N}(f)}{R_N(f)} \approx \frac{\frac{C_f}{\rho^{2N}}}{\frac{C_{f'}}{\rho^{2N}}} . \frac{1}{(other \quad term)} \tag{86}$$

The core result, proven by Donald Newman for the function $f(x) = |x|$ and generalized by others, is that for many functions, the rational error converges at the rate of the square of the polynomial rate:

$$lim_{N\to\infty} \frac{E_{2N}(f)}{R_N(f)} = \infty \tag{87}$$

This demonstrates that the rational function error decays at a super-geometric or double-geometric rate compared to the polynomial error, confirming the phrase "converges much faster".

XIII Mathematical Privilege of GCAT Vs GCNN

From the equation (63) we have the score α_{ij} to be used directly as an entry for the matrix ϕ but to encourage attention sparsity the α_{ij} is passed through **local soft maximum operator** :-

$$\phi_{ij} = \frac{e^{\alpha_{ij}}}{\left(\Sigma_{j'\in N_i \cup i} e^{\alpha_{ij'}}\right)} \tag{88}$$

Since the soft maximum operator assigns a rational function form for the matrix ϕ which is a layer specific matrix and also convolutional in nature for the GCAT and since being different from the shift operator S the operation takes place as follows:-

$$X_l = \sigma\left(\Sigma_{k=0}^{K} \phi^k X_{l-1} A_k\right), here \quad \phi = \phi_l \quad and A_k = A_{lk} \tag{89}$$

Now coming back to the Newman's proof[49] for best uniform polynomial and rational approximation errors for $f(x) = |x|$ on the interval $[-1, 1]$ is
Polynomial Approximation $= E_N(|x|) \sim \frac{C_1}{n}$, where C_1 is a constant.
And the **Rational Approximation** rate being $= R_{n,n}(|x|) \sim C_2 e^{-C_3\sqrt{n}}$, where C_2, C_3 being constants.

A. The Non-Analytic Case ($f(x) = |x|$)

For $|x|$:

- Polynomial Rate: $O(1/n)$ (algebraic)
- Rational Rate: $O(e^{-\pi\sqrt{n}})$ (root-exponential)

The ratio of the errors shows the enormous difference:

$$\lim_{n\to\infty} \frac{E_n(|x|)}{R_{n,n}(|x|)} = \lim_{n\to\infty} \frac{C_1/n}{C_2 e^{-\pi\sqrt{n}}} = \infty$$

The rational error vanishes incredibly fast, while the polynomial error only decays as $1/n$. The polynomial's failure is due to the **singularity at the origin** ($|x|$ is not differentiable at $x = 0$). A rational function uses the poles of its denominator, clustered exponentially near the singularity, to absorb the non-smoothness.

B. The Analytic Case (The $\sim 1/\rho^{2N}$ Analogy)

The "rate-squared" analogy is **literally true for analytic functions**, which is the more generalized result:

- Polynomial Rate: $E_K(f) \sim \frac{C_E}{\rho^K}$
- Rational Rate: $R_{N,N}(f) \sim \frac{C_R}{\rho^{2N}}$

If we compare them using the same total parameter count, $K \approx 2N$:

$$R_{N,N}(f) \approx \frac{C_R}{\rho^{2N}} = \frac{C_R}{(\rho^N)^2}$$

If the polynomial error for degree N is $E_N(f) \sim 1/\rho^N$, then the rational error $R_{N,N}(f)$ is proportional to the **square of the polynomial error of half its degree**. This is where the analogy originates and why rational approximation is considered "double-geometric."Hence the entries for the matrix ϕ being in rational format depict a more rapid convergence compared to the GCNN hence is superior to it.

XIV Results

XIV.1 GCNN Results

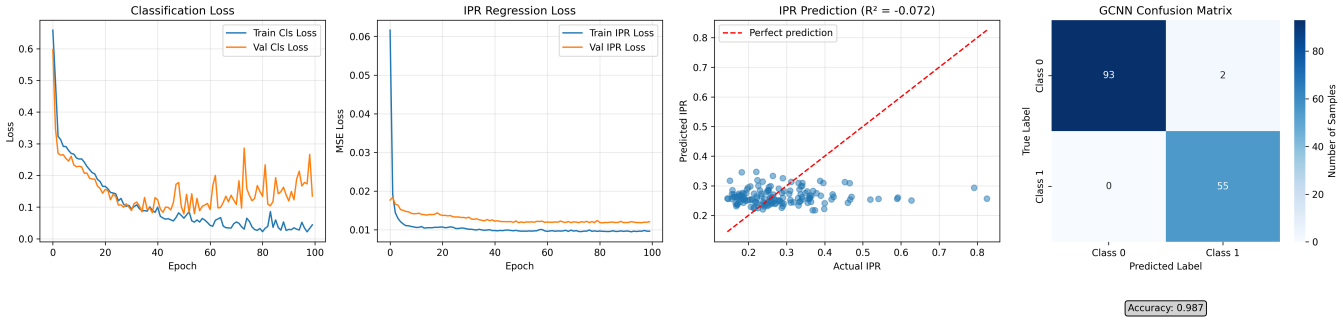

Figure 6: IPR Prediction for Star Network using GCNN.The GCNN model is trained with a strongly localized states.The input comes in as a the star network with a synthetic dataset preparation and with the associated targeted IPR values.The observation in the output relies on the true and the predicted ipr values.The observation primarily relies on the Confusion Matrix and the IPR distrbution and about how far the predicted ipr values distribute across the perfect prediction.The Accuracy score from the confusion matrix is := .987 ,the precision score := 1,the Recall := 0.979 and the F1-Score(provides a better sense about the model's overall performance particularly for imbalanced datasets, helpful when both false postives and false negatives are important) := 0.989

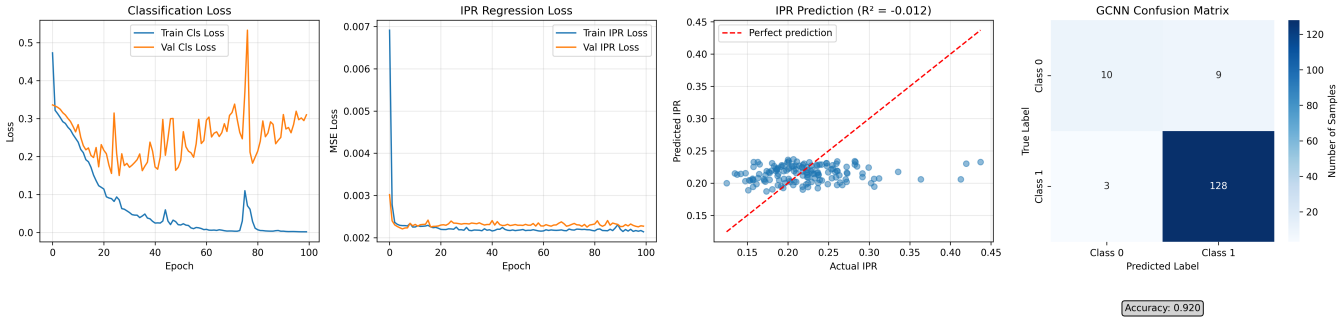

Figure 7: IPR Prediction for Scale Free Network using GCNN.The GCNN model is trained with a weakly localized states.The input comes in as a the scale free network with a synthetic dataset preparation and with the associated targeted IPR values.The observation in the output relies on the true and the predicted ipr values.The observation primarily relies on the Confusion Matrix and the IPR distrbution and about how far the predicted ipr values distribute across the perfect prediction.The Accuracy score from the confusion matrix is :=0.920 ,the precision score := 0.769,the Recall := 0.526 and the F1-Score(provides a better sense about the model's overall performance particularly for imbalanced datasets, helpful when both false postives and false negatives are important) := 0.625

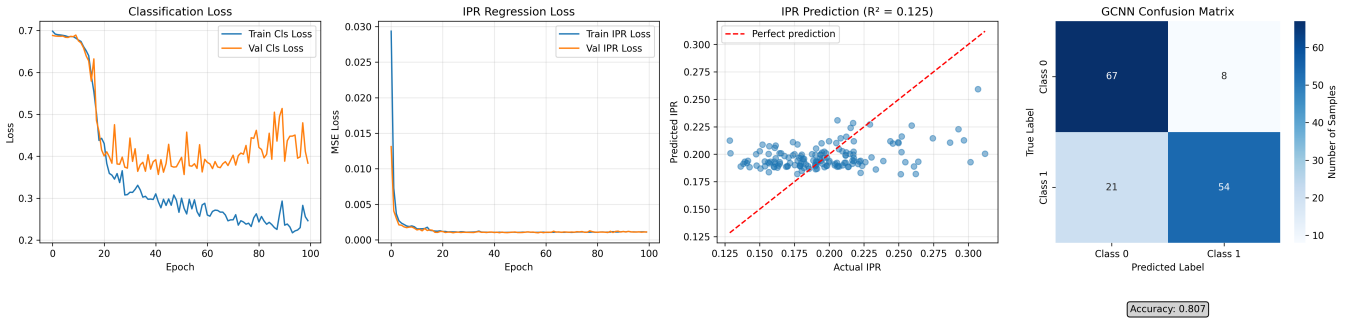

Figure 8: IPR Prediction for Wheel Network using GCNN.The GCNN model is trained with a strongly localized states.The input comes in as a the Wheel network with a synthetic dataset preparation and with the associated targeted IPR values.The observation in the output relies on the true and the predicted ipr values.The observation primarily relies on the Confusion Matrix and the IPR distrbution and about how far the predicted ipr values distribute across the perfect prediction.The Accuracy score from the confusion matrix is := 0.807 ,the precision score := 0.788,the Recall := 0.893 and the F1-Score(provides a better sense about the model's overall performance particularly for imbalanced datasets, helpful when both false postives and false negatives are important) := 0.837

XIV.2 ARMA Results

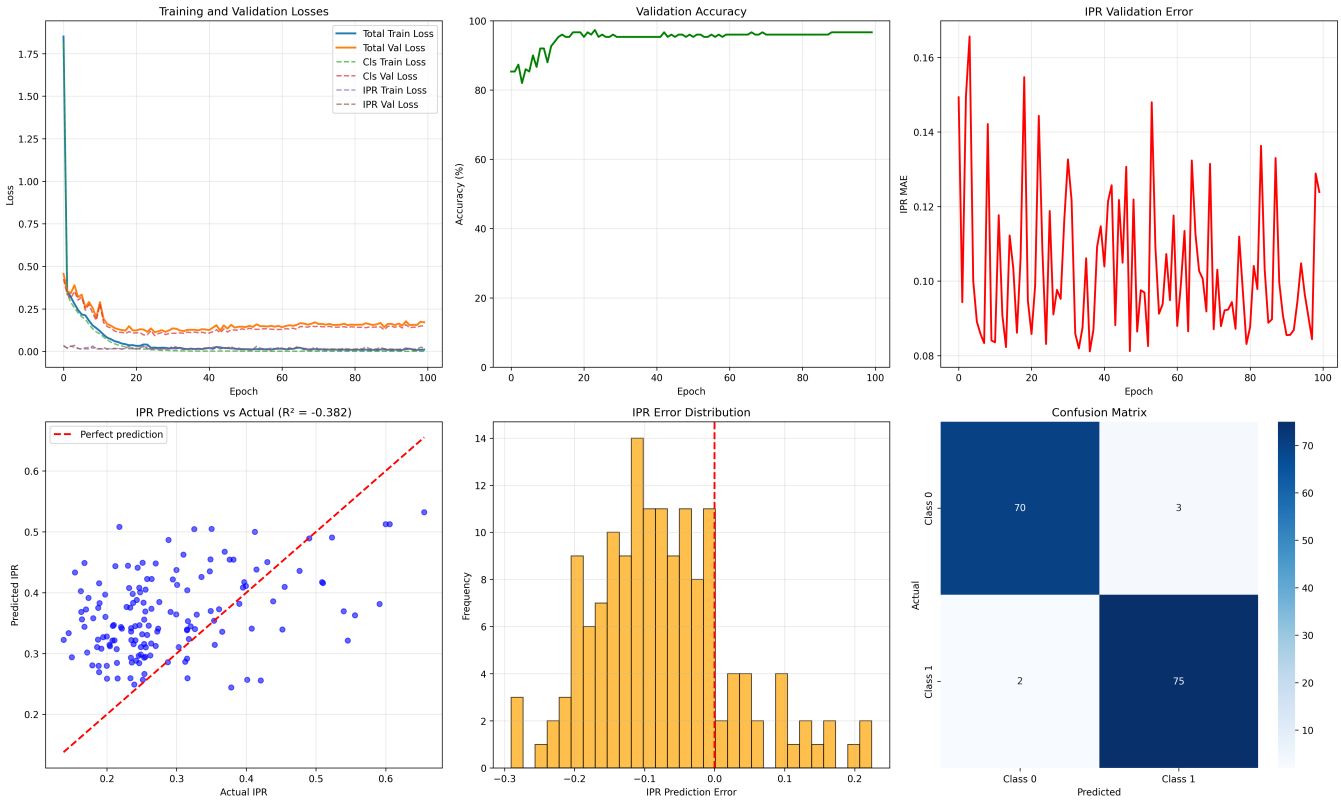

Figure 9: IPR Prediction for Star Network using ARMA.The ARMA model is trained with a strongly localized states.The input comes in as a the star network with a synthetic dataset preparation and with the associated targeted IPR values.The observation in the output relies on the true and the predicted ipr values.The observation primarily relies on the Confusion Matrix and the IPR distrbution and about how far the predicted ipr values distribute across the perfect prediction.Compared to the GCNN the ipr predicted values show a better nature as they tend to align more interms of their accuracy along the perfect IPR prediction.The Accuracy score from the confusion matrix is := .967 ,the precision score := 0.972,the Recall := 0.959 and the F1-Score(provides a better sense about the model's overall performance particularly for imbalanced datasets, helpful when both false postives and false negatives are important) := 0.966

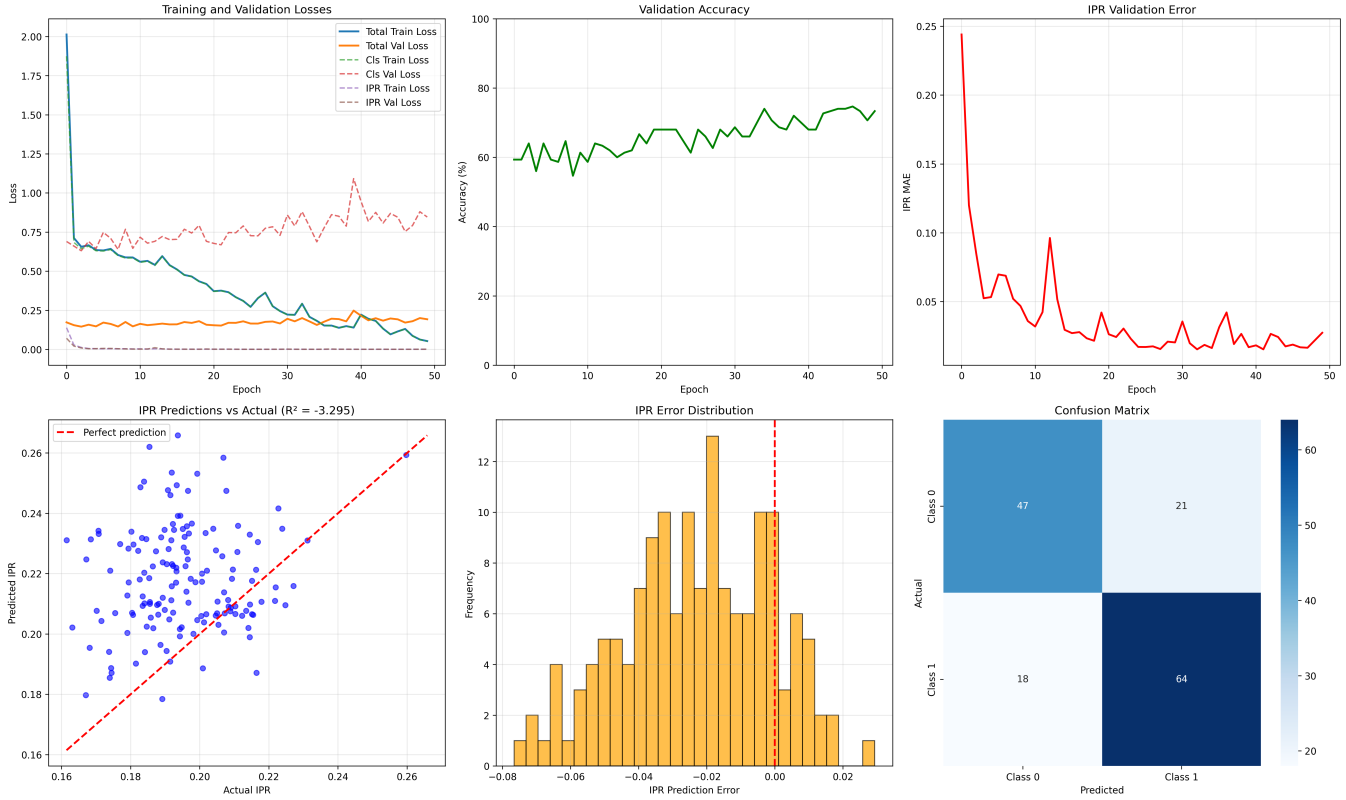

Figure 10: IPR Prediction for Wheel Network using ARMA.The ARMA model is trained with a strongly localized states.The input comes in as a the wheel network with a synthetic dataset preparation and with the associated targeted IPR values.The observation in the output relies on the true and the predicted ipr values.The observation primarily relies on the Confusion Matrix and the IPR distrbution and about how far the predicted ipr values distribute across the perfect prediction.Compared to the GCNN the ipr predicted values show a better nature as they tend to align more interms of their accuracy along the perfect IPR prediction.The Accuracy score from the confusion matrix is := .740 ,the precision score := 0.723,the Recall := 0.691 and the F1-Score(provides a better sense about the model's overall performance particularly for imbalanced datasets, helpful when both false postives and false negatives are important) := 0.707

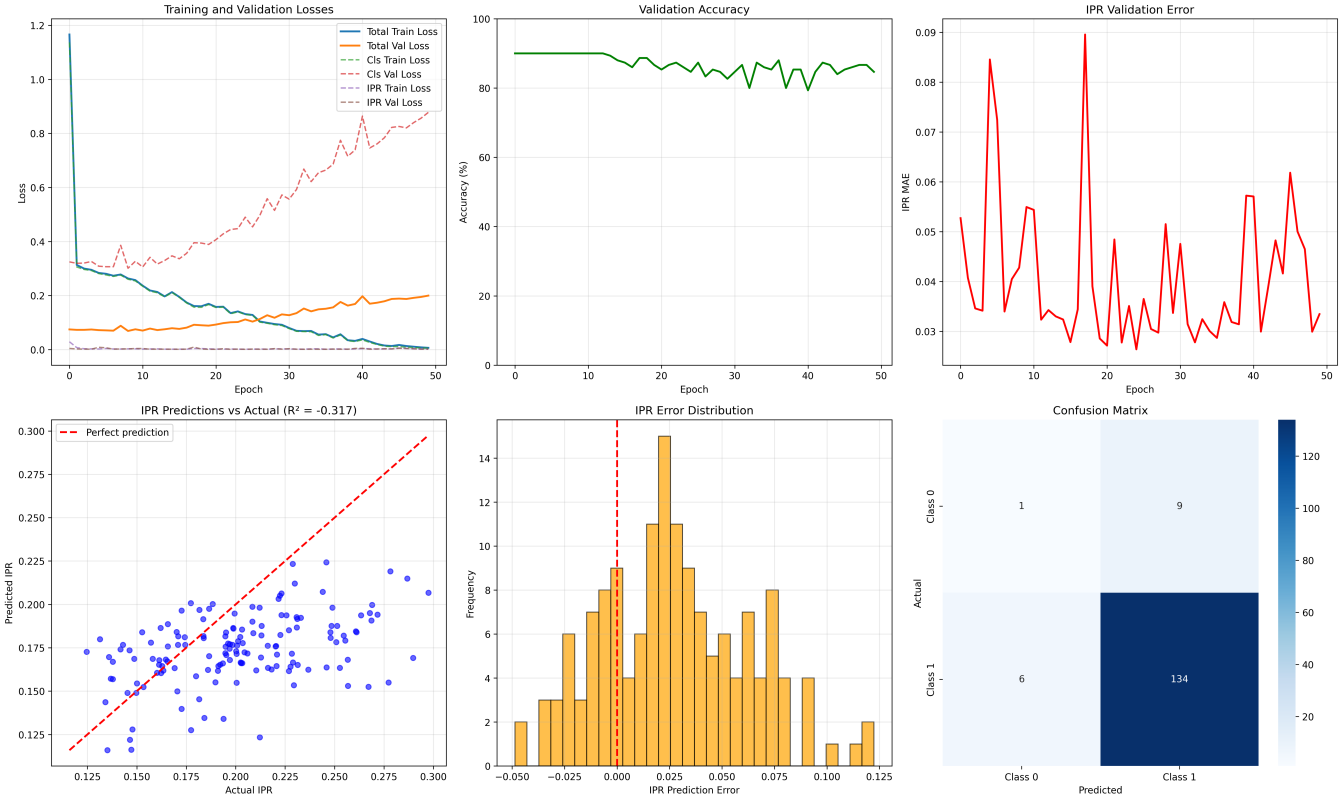

Figure 11: IPR Prediction for Sclae Free Network using ARMA.The ARMA model is trained with a weaklylocalized states.The input comes in as a the scale free network with a synthetic dataset preparation and with the associated targeted IPR values.The observation in the output relies on the true and the predicted ipr values.The observation primarily relies on the Confusion Matrix and the IPR distrbution and about how far the predicted ipr values distribute across the perfect prediction.Compared to the GCNN the ipr predicted values show a better nature as they tend to align more interms of their accuracy along the perfect IPR prediction.The Accuracy score from the confusion matrix is := .900 ,the precision score := 0.143,the Recall := 0.100 and the F1-Score(provides a better sense about the model's overall performance particularly for imbalanced datasets, helpful when both false postives and false negatives are important) := 0.118

XIV.3 GCAT Results

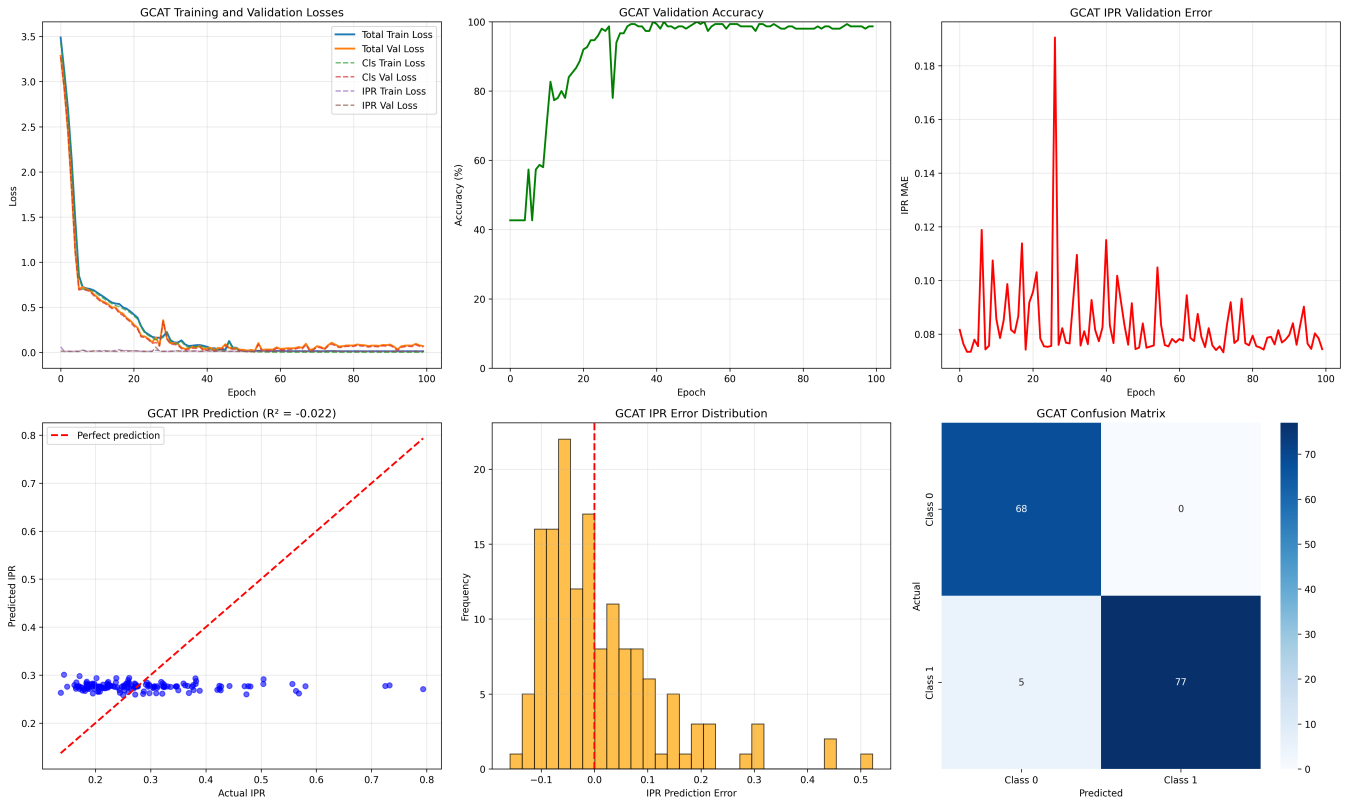

Figure 12: IPR Prediction for Star Network using GCAT.The GCAT model is trained with a strongly localized states.The input comes in as a the star network with a synthetic dataset preparation and with the associated targeted IPR values.The observation in the output relies on the true and the predicted ipr values.The observation primarily relies on the Confusion Matrix and the IPR distrbution and about how far the predicted ipr values distribute across the perfect prediction.Compared to the GCNN the ipr predicted values show a better nature as they tend to perform better in the confusion matrix compared to the GCNN.The Accuracy score from the confusion matrix is := .0.977 ,the precision score := 0.932,the Recall := 1 and the F1-Score(provides a better sense about the model's overall performance particularly for imbalanced datasets, helpful when both false postives and false negatives are important) := 0.965

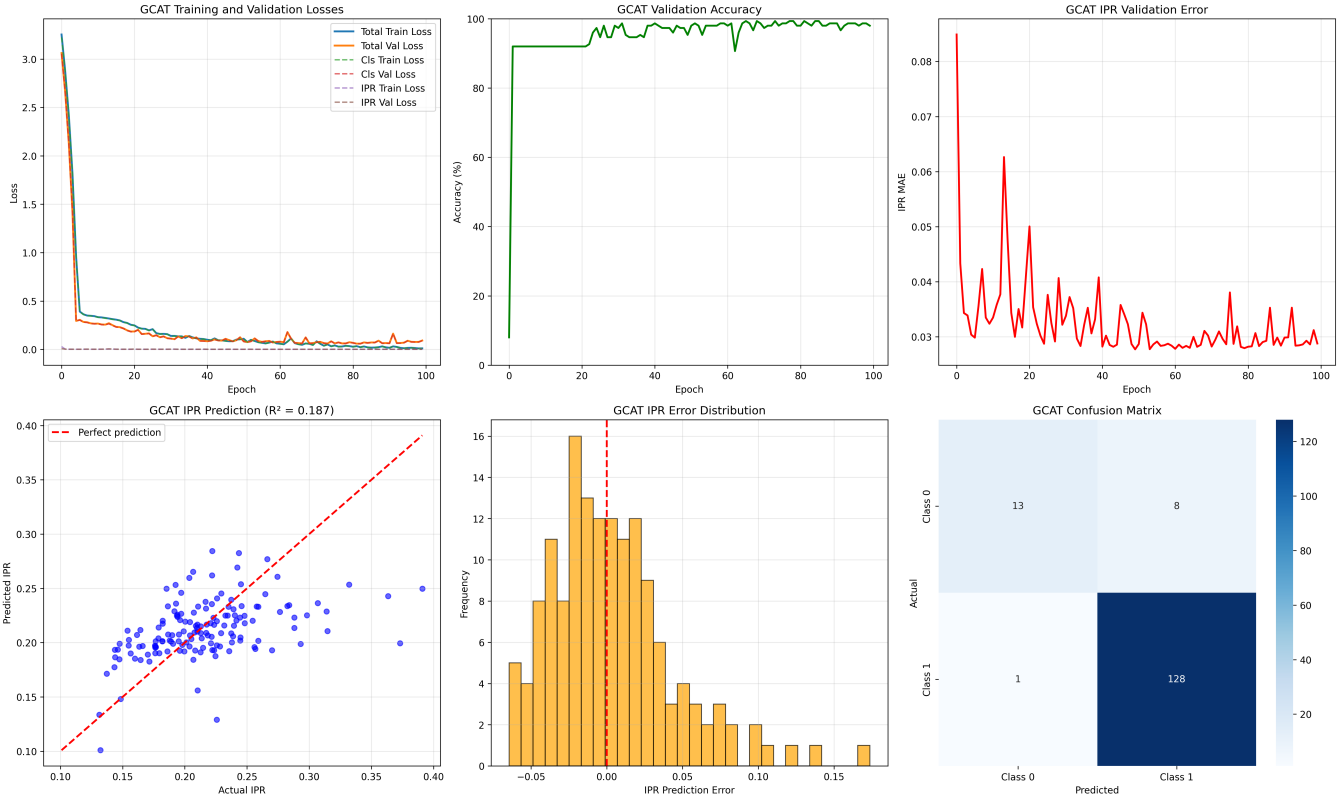

Figure 13: IPR Prediction for Scale Free Network using GCAT.The GCAT model is trained with a weakly localized states.The input comes in as a the scale free network with a synthetic dataset preparation and with the associated targeted IPR values.The observation in the output relies on the true and the predicted ipr values.The observation primarily relies on the Confusion Matrix and the IPR distrbution and about how far the predicted ipr values distribute across the perfect prediction.Compared to the GCNN the ipr predicted values show a better nature as they tend to perform better in the confusion matrix compared to the GCNN.The Accuracy score from the confusion matrix is := .0.940 ,the precision score := 0.929,the Recall := 0.619 and the F1-Score(provides a better sense about the model's overall performance particularly for imbalanced datasets, helpful when both false postives and false negatives are important) := 0.743

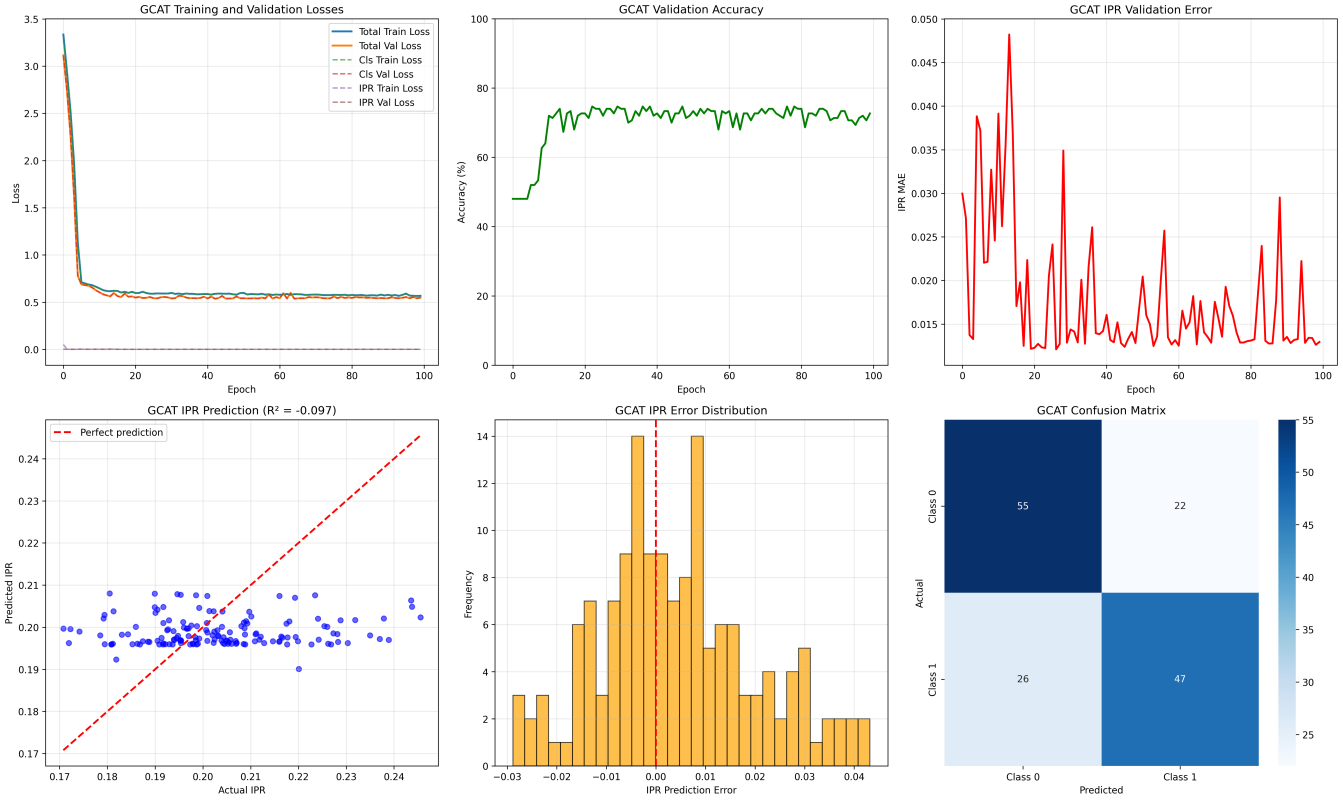

Figure 14: IPR Prediction for Wheel Network using GCAT.The GCAT model is trained with a strongly localized states.The input comes in as a the wheel network with a synthetic dataset preparation and with the associated targeted IPR values.The observation in the output relies on the true and the predicted ipr values.The observation primarily relies on the Confusion Matrix and the IPR distrbution and about how far the predicted ipr values distribute across the perfect prediction.Compared to the GCNN the ipr predicted values show a better nature as they tend to perform better in the validation accuracy compared to the GCNN.The Accuracy score from the confusion matrix is := .0.680 ,the precision score := 0.679,the Recall := 0.714 and the F1-Score(provides a better sense about the model's overall performance particularly for imbalanced datasets, helpful when both false postives and false negatives are important) := 0.696

XV Conclusion

Future implementations for a better framework can be achieved through the Local Activation Function which can be utilized in the training procedure.Contrary to the pointwise nonlinearities used in the frameworks so far can be replaced with the localized activation function that can act on multiple nodal components at a time as a result for which both the forward and the backward calculations are carried out at the same time by the Neural Networks thus the both the forward and the backward passes data need to be updated.The fact that local activation functions are equally applicable to training via backpropagation like the conventional GNNs and hence return closed form expressions for the trainable parameter's gradient updates is well established in this reference[43].The learnable parameters finds a surge in addition to the two regression layer and the local activations includes another.
Thus with the increase in the number of learnable parameters the then formed framework with local activation function adopts better to complex network topologies performing better with a rapid convergence as the weight matrices are updated using gradient descent .Thus frameworks made with local activation function can perform better than pointwise non linearities.Besides the above approaches can be taken towards other distinct forms of frameworks including edge varying and the node varying.Physics informed graph neural network or graph tensor networks can form better alternatives.

XVI References

References

[1] Steven H Strogatz. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. Chapman and Hall/CRC, 2024.

[2] Steven H Strogatz. "Exploring complex networks". In: *nature* 410.6825 (2001), pp. 268–276.

[3] Danielle S Bassett et al. "Dynamic reconfiguration of human brain networks during learning". In: *Proceedings of the National Academy of Sciences* 108.18 (2011), pp. 7641–7646.

[4] H Jeff Kimble. "The quantum internet". In: *Nature* 453.7198 (2008), pp. 1023–1030.

[5] Gautam Mahapatra et al. "Multilevel Digital Contact Tracing". In: *arXiv preprint arXiv:2007.05637* (2020).

[6] Zhenzhen Yang et al. "Estimating the influence of disruption on highway networks using GPS data". In: *Expert Systems with Applications* 187 (2022), p. 115994.

[7] Stephen Eubank et al. "Modelling disease outbreaks in realistic urban social networks". In: *Nature* 429.6988 (2004), pp. 180–184.

[8] Ira M Longini Jr et al. "Containing pandemic influenza at the source". In: *Science* 309.5737 (2005), pp. 1083–1087.

[9] Huaiyu Tian et al. "Urbanization prolongs hantavirus epidemics in cities". In: *Proceedings of the National Academy of Sciences* 115.18 (2018), pp. 4707–4712.

[10] Benjamin D Dalziel et al. "Urbanization and humidity shape the intensity of influenza epidemics in US cities". In: *Science* 362.6410 (2018), pp. 75–79.

[11] Sarika Jalan and Priodyuti Pradhan. "Wheel graph strategy for PEV localization of networks". In: *Europhysics Letters* 129.4 (2020), p. 46002.

[12] Franco Scarselli et al. "Graph neural networks for ranking web pages". In: *The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05)*. IEEE. 2005, pp. 666–672.

[13] Franco Scarselli et al. "The graph neural network model". In: *IEEE transactions on neural networks* 20.1 (2008), pp. 61–80.

[14] Fernando Gama et al. "Graphs, convolutions, and neural networks: From graph filters to graph neural networks". In: *IEEE Signal Processing Magazine* 37.6 (2020), pp. 128–138.

[15] Joan Bruna et al. "Spectral networks and locally connected networks on graphs". In: *arXiv preprint arXiv:1312.6203* (2013).

[16] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. "Convolutional neural networks on graphs with fast localized spectral filtering". In: *Advances in neural information processing systems* 29 (2016).

[17] Jian Du et al. "On graph convolution for graph CNNs". In: *2018 IEEE Data Science Workshop (DSW)*. IEEE. 2018, pp. 1–5.

[18] Thomas N Kipf and Max Welling. *Semi-supervised learning with graph convolutional networks*. 2017.

[19] Keyulu Xu et al. "How powerful are graph neural networks?" In: *arXiv preprint arXiv:1810.00826* (2018).

[20] Ron Levie, Elvin Isufi, and Gitta Kutyniok. "On the transferability of spectral graph filters". In: *2019 13th International conference on sampling theory and applications (SampTA)*. IEEE. 2019, pp. 1–5.

[21] Martin Simonovsky and Nikos Komodakis. "Dynamic edge-conditioned filters in convolutional neural networks on graphs". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 3693–3702.

[22] Federico Monti et al. "Geometric deep learning on graphs and manifolds using mixture model cnns". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5115–5124.

[23] James Atwood and Don Towsley. "Diffusion-convolutional neural networks". In: *Advances in neural information processing systems* 29 (2016).

[24] Petar Velickovic, Guillem Cucurull, and Arantxa Casanova. "Adriana Romero Pietro Lio and Yoshua Bengio. Graph attention networks". In: *ICLR*. 2018.

[25] Zonghan Wu et al. "A comprehensive survey on graph neural networks". In: *IEEE transactions on neural networks and learning systems* 32.1 (2020), pp. 4–24.

[26] Jie Zhou et al. "Graph neural networks: A review of methods and applications". In: *AI open* 1 (2020), pp. 57–81.

[27] Ziwei Zhang, Peng Cui, and Wenwu Zhu. "Deep Learning on Graphs: A Survey". In: *IEEE Transactions on Knowledge and Data Engineering* PP (Mar. 2020), pp. 1–1. DOI: 10.1109/TKDE.2020.2981333.

[28] John Lee et al. "Attention Models in Graphs: A Survey". In: *ACM Transactions on Knowledge Discovery from Data* 13 (Nov. 2019), pp. 1–25. DOI: 10.1145/3363574.

[29] Priodyuti Pradhan and Amit Reza. "Predicting Steady-State Behavior in Complex Networks with Graph Neural Networks". In: *arXiv preprint arXiv:2502.01693* (2025).

[30] Elvin Isufi, Fernando Gama, and Alejandro Ribeiro. "EdgeNets: Edge Varying Graph Neural Networks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.11 (2022), pp. 7457–7473. DOI: 10.1109/TPAMI.2021.3111054.

[31] Anqi Mao, Mehryar Mohri, and Yutao Zhong. *Cross-Entropy Loss Functions: Theoretical Analysis and Applications*. 2023. arXiv: 2304.07288 [cs.LG]. URL: https://arxiv.org/abs/2304.07288.

[32] "Mean Squared Error". In: *The Concise Encyclopedia of Statistics*. New York, NY: Springer New York, 2008, pp. 337–339. ISBN: 978-0-387-32833-1. DOI: 10.1007/978-0-387-32833-1_251. URL: https://doi.org/10.1007/978-0-387-32833-1_251.

[33] Fernando Gama et al. "Convolutional neural network architectures for signals supported on graphs". In: *IEEE Transactions on Signal Processing* 67.4 (2018), pp. 1034–1049.

[34] Gabriel Taubin. "Geometric Signal Processing on Polygonal Meshes". In: *Eurographics State of the Art Reports* 4 (June 2001).

[35] David Shuman, Pierre Vandergheynst, and Pascal Frossard. "Distributed Signal Processing via Chebyshev Polynomial Approximation". In: *IEEE Transactions on Signal and Information Processing over Networks* 4 (Nov. 2011). DOI: 10.1109/TSIPN.2018.2824239.

[36] Aliaksei Sandryhaila and Jose Moura. "Discrete signal processing on graphs: Graph filters". In: Oct. 2013, pp. 6163–6166. DOI: 10.1109/ICASSP.2013.6638849.

[37] Sunil K. Narang and Antonio Ortega. "Compact Support Biorthogonal Wavelet Filterbanks for Arbitrary Undirected Graphs". In: *IEEE Transactions on Signal Processing* 61 (Oct. 2012). DOI: 10.1109/TSP.2013.2273197.

[38] Oguzhan Teke and Palghat Vaidyanathan. "Extending Classical Multirate Signal Processing Theory to Graphs - Part I: Fundamentals". In: *IEEE Transactions on Signal Processing* 65 (Jan. 2017), pp. 409–422. DOI: 10.1109/TSP.2016.2617833.

[39] Santiago Segarra, Antonio Marques, and Alejandro Ribeiro. "Optimal Graph-Filter Design and Applications to Distributed Linear Network Operators". In: *IEEE Transactions on Signal Processing* PP (May 2017), pp. 1–1. DOI: 10.1109/TSP.2017.2703660.

[40] Elvin Isufi et al. "Autoregressive Moving Average Graph Filtering". In: *IEEE Transactions on Signal Processing* PP (Oct. 2016), pp. 1–1. DOI: 10.1109/TSP.2016.2614793.

[41] Haggai Maron et al. "Invariant and equivariant graph networks". In: *arXiv preprint arXiv:1812.09902* (2018).

[42] Fernando Gama, Joan Bruna, and Alejandro Ribeiro. "Stability properties of graph neural networks". In: *IEEE Transactions on Signal Processing* 68 (2020), pp. 5680–5695.

[43] Luana Ruiz et al. “Invariance-preserving localized activation functions for graph neural networks”. In: *IEEE Transactions on Signal Processing* 68 (2019), pp. 127–141.

[44] N Alan Heckert and James J Filliben. “Nist/sematech e-handbook of statistical methods; chapter 1: Exploratory data analysis”. In: (2003).

[45] Lloyd N Trefethen. *Approximation theory and approximation practice, extended edition*. SIAM, 2019.

[46] Jiani Liu, Elvin Isufi, and Geert Leus. “Filter design for autoregressive moving average graph filters”. In: *IEEE Transactions on Signal and Information Processing over Networks* 5.1 (2018), pp. 47–60.

[47] Matt Young. “The stone-weierstrass theorem”. In: *MATH 328 Notes*. Queen’s University at Kingston, 2006.

[48] Penco Petrov Petrushev and Vasil Atanasov Popov. *Rational approximation of real functions*. 28. Cambridge University Press, 2011.

[49] Donald J Newman. *Approximation with rational functions*. 41. American Mathematical Soc., 1979.