Indonesian Journal of Electrical Engineering and Computer Science
Vol. 42, No. 2, May 2026, pp. 403~413
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v42.i2.pp403-413

Fairness dynamics in graph neural networks: a comparative study of graph-structured neural models with and without gradient-based training

Ananda Chatterjee[1], K. A. Venkatesh[2]
[1]Department of Applied Mathematics, Alliance University, Bangalore, India
[2]School of Advance Computing, Alliance University, Bangalore, India

Article Info

Article history:

Received Dec 6, 2025
Revised Mar 4, 2026
Accepted May 26, 2026

Keywords:

Backpropagation
Fairness
Graph attention network
Graph convolutional network
Graph sample and aggregate

ABSTRACT

Graph neural networks (GNNs) are gaining more and more popularity in high stakes domain due to their ability to learn both from features and relationships. Nevertheless, there are concerns regarding how this accuracy centric optimization used by these models will impact fairness when deployed in socially sensitive areas. This work explores the interplay between predictive accuracy and fairness in GNNs when applied in judicial risk assessment system. A comparative study was performed among three canonical architectures such as graph convolutional networks (GCN), graph sample and aggregate (GraphSAGE) and graph attention networks (GAT) under trained and untrained settings on judicial risk assessment dataset. Fairness was evaluated through metrices like demographic parity (DP), equalized opportunity (Eopp), and equalized odds (Eodds) along with predictive performance metrices. Sensitivity analysis was conducted to investigate the effect of graph construction choices and neighborhood sizes in determing fairness and predictive accuracy. Experimental evidences proved that backpropagation improved predictive performance but in tandem fairness degradation happened. Untrained models exhibited lower fairness gap but that is superficial as weak predictive outcome of those models made group differences suppressed. Among the three trained models GAT was able to strike a good balance between accuracy and fairness while increase in neighborhood size caused little bit improvement in fairness via graph smoothing. The novelty of this work lies with its empericial characterization of GNNs under realistic settings. This study emphasizes the fact that how learning methodology, architectural designs, graph formation influence fairness outcomes. This work enlightens how graph-based models can be applied to decision making scenario and encourages embedding of fairness aware training strategies to it.

Corresponding Author:

Ananda Chatterjee
Department of Applied Mathematics, Alliance University
Bangalore, Karnataka, 562106, India
Email: canandaPHD724@sam.alliance.edu.in

1. INTRODUCTION

The rise in the applications of artificial intelligence in socially sensitive areas has revived the discussions on fairness, accountability, and transparency. Finance, healthcare, and criminal justice are also the common uses of automated decision systems, the fact is that their forecasts directly influence the life of people. The most talked about example is probably the correctional offender management profiling for

alternative sanctions (COMPAS) recidivism prediction tool, which is applied in the United States to forecast the risk of defendants to reoffend within two years. Research showed that the system allowed the higher risk score to be given to African-American defendants in a systematic manner than it was given to Caucasian defendants with the same criminal history [1]. This finding brought to the fore an even greater problem, which is that machine learning models, despite their supposed neutrality, can reproduce or even amplify inequities in training data, often leading to discrimination across demographic groups [2], [3]. Simultaneously, graph neural networks (GNNs) have also become a potent type of models of learning on graph-structured data by leveraging node attributes and realational information. GNNs identify dependencies that are missed by traditional vector-based models by accumulating information of a node by the neighbors in an iterative manner [4], [5]. Other forms of graphs have become the backbone in areas of social network analysis and molecular property prediction by attaining strong predictive performance, such as graph convolutional networks (GCNs) [6], [7], graph sample and aggregate (GraphSAGE) [8] and graph attention networks (GATs) [9] would be used. Despite the fact that these frameworks have been demonstrated as predictive to a great extent, not much is known about their result in maintaining fairness, especially in scenarios which involve high stake alike criminal justice.

Majority of the existing reserach on algorithmic fairness have revolved around defining and diminishing group inequities in conventional supervised learning environments. Broadly accepted criteria are demographic parity (DP) [10], equalized opportunity (EOpp) and equalized odds (EOdds) [11] along mitigation methodologies covering pre-processing such as resampling and reweighting of training data [12], in-processing via adverserial debiasing [13] and post processing methods [14]. In recent times, fairness concerns span over graph based learning. FairGNN integrates adverserial objectives to attenuate bias in GNNs [15], [16]. EDITs offers a standardized framework to evaluate fairness in graph learning [17] while FS-GNN ameliorates fairness through sparsification of graph structure [18], [19]. Although these methods suggest explicit debiasing strategies, they mainly assess fairness on fully trained models and implicitly assume fairness as a static architectural property.

This approach overlooks a gap, that is the idea of fairness not only depends on model architecture but it is also impacted by the optimization algorithm which is induced by training. Untrained models whose parameters are randomly initialized capture only architectural inductive biases while trained models embed both architecture and learned correlations through backpropagation [20], [21]. Differentiating between two settings is crucial to comprehend whether the observed fairness emerges due to architectural design or it is amplified because of accuracy driven optimization. To our knowledge, no prior study has systematically compared the fairness dynamics of GCN, GraphSAGE, and GAT across both trained and untrained regimes on a socially sensitive dataset such as COMPAS.

The present study addresses this gap by analyzing how fairness varies with and without backpropagation in three canonical GNN architectures. Using the COMPAS dataset as a benchmark, we evaluate predictive performance through Accuracy and ROC-AUC [22], [23], alonside multiple fairness metricesss including DP, EOpp, and EOdds and calibration based measures [24], [25]. In this study sensitivity analyis was conducted over graph construction parameters, specifically using neighborhood size to assess how bias diffuses through message passing. Our results demonstrate that backpropagation consistently enhances predictive accuracy but simultaneously increases disparities across sensitive groups, with the severity of fairness erosion varying by architecture This comparative analysis will bring up a new concept of fairness in graph learning: fairness as not a fixed property of an architecture, but an outcome by the interconnection between architecture and optimization [21]. The innovation has both the methodological and practical value as it provides future directions to create the fairness-conscious training of GNNs on high-stakes applications and as indicated in the case of recent sparsification-based fairness optimization work in GNNs [19] and in another study of incorporating fairness with explainability in automatic decision-making systems [26]-[28].

2. LITERATURE REVIEW

The concept of fairness in prediction has been well-researched throughout the larger machine learning community and a range of formal definitions and methods have been developed. One of the most popular ones is DP that mandates the ratio between positive predictions to be the same among protected groups irrespective of real results [10]. The EOpp goes one step further and demands equality of true positive rates whereas EOdds extends it further to true positive and false positive rates [11]. Many research in algorithmic fairness has been based on these definitions. Related formulations look into different measurement and ingrained trade-offs which naturally occur due to the conjunctions between fairness constraints and predictive performance [13], [14]. Earlier methods attempted to equalize datasets in the phase of training by re-weighting or re-sampling to balance group distributions [12]. In-processing approaches alters

the learning goal directly with the help of, adversarial debiasing where a predictor and an adversary are trained together to make group-invariant representations [13]. Post-processing methods, such as the EOdds framework, modify group-specific thresholds on predicted probabilities to decrease group disparities [11].

While these developments laid important groundwork in classical supervised learning, fairness concerns became more complex with the rise of deep learning. Deep architectures, though celebrated for their ability to capture hierarchical and nonlinear patterns, are also vulnerable to encoding spurious correlations from training data [2], [14]. Studies in computer vision and natural language processing have shown that convolutional and recurrent networks can reflect and even amplify biases in race and gender, motivating fairness-aware deep learning approaches such as adversarial debiasing, gradient reversal, and fairness-regularized loss functions [14]. This concern has been taken care of in case of language models, where bias in pre trained embeddings has been minimized with a fairness-constrained representation-learning framework [14].

The challenge becomes even more pronounced when learning extends to geometric domains, including graphs, manifolds, and other non-Euclidean structures. GCNs introduced spectral graph convolutions with degree normalization [7]. GraphSAGE advanced this approach by sampling neighbors for inductive scalability [8], and GATs introduced attention mechanisms that assign variable importance to neighbors [9]. Surveys such as Wu *et al.* [4] provide comprehensive overviews of this field [4], [5], while Bronstein *et al.* [6] proposed the broader framework of geometric deep learning, unifying neural methods across graphs and other structured domains [6].

Fairness research in graph learning remains relatively nascent. FairGNN integrated adversarial debiasing with label propagation to reduce group disparities. Agarwal *et al.* [16] developed EDITs, a benchmark for standardized fairness evaluation in graph learning [17]. Kose and Shen [18] empirically showed that attention mechanisms in GATs may exacerbate unfairness if left unregularized. A recent work introduced a framework FS-GNN which enhanced the fairness by taking account the graphical input and architecture of the model [19]. Broader surveys highlight the fragmented nature of fairness research in GNNs, calling for systematic benchmarks and standardized evaluation protocols [14].

Another limitation that is common to present literature is that it is characterized by a focus on trained models. In nearly every research, parameters are optimized by the use of backpropagation and the fairness performance is measured. However, optimization is not neutral: it does not act on the statistical regularities of the data, but tends to amplify existing biases [15]. This leads to a basic question, long ignored, which is the extent to which the unfairness of any model is due to its inductive bias on architecture, and to what extent it is created or exacerbated by training. To overcome this difference, it is necessary to compare the regimes that are trained and untrained systematically to disentangle these effects. In all the cases, nothing has been previously studied comparing this question on GCNs, GraphSAGE, and GATs on tasks that are sensitive to fairness. It has been noted in a recent study that explainability should be accompanied by fairness in an automatic decision system in an effort of attaining technical validity and legal responsibility [26]. The present work therefore contributes new value by explicitly disentangling architectural bias from optimization-driven bias, offering a deeper understanding of fairness dynamics in geometric deep learning [29].

3. METHOD

The methodological framework comprised dataset preparation, graph construction, model design, training regimes, and evaluation metrics. The dataset used was the COMPAS two-year recidivism dataset [1], which contains demographic and criminal history information for more than 7,000 defendants. The prediction target was whether the individual reoffended within the period of two years. The sensitive attribute was race, with the analysis restricted to African-American and Caucasian defendants to form a binary sensitive variable. Records with missing or inconsistent fields were removed. Numerical features, including age, priors count, and juvenile counts, were standardized using z-scores, while categorical features such as sex and charge degree were encoded using one-hot encoding.

Since the dataset is tabular, a graph structure was induced via k-nearest neighbors (k=10) using Euclidean distance in feature space. All individuals was represented by nodes, while the similar defendant were connected through edges. The adjacency was symmetrized to ensure undirected connectivity. The resulting graph reflects similarity in demographic and criminal history profiles. The data was divided into training (70%), validation (15%), and testing (15%) with the help of stratified sampling to maintain class proportions. Boolean masks ensured consistent splits across models.

Three architectures were evaluated. The GCN uses normalized spectral convolutions:

$$H^{(l+1)} = \sigma\left(\left(\widetilde{D}\right)^{-1/2} \tilde{A} \left(\widetilde{D}\right)^{-1/2} H^{(l)} W^{(l)}\right)[7] \quad (1)$$

where $\tilde{A} = A + I$, $\widetilde{D}$ is the degree matrix, and $W^{(l)}$ are the trainable weights and σ is the activation function.

GraphSAGE performs inductive neighbor sampling:

$$h_v^{(l+1)} = \sigma\left(W^{(l)} \cdot \text{concat!}\left(h_v^{(l)}, \text{Mean}\{h_u^{(l)} : u \in \mathcal{N}(v)\}\right)\right)[8] \quad (2)$$

GAT introduces attention-based aggregation:

$$h_v^{(l+1)} = \sigma\left(\sum_{u\in\mathcal{N}(v)} \alpha_{vu}\, W^{(l)} h_u^{(l)}\right)[9] \quad (3)$$

where α_{vu} is the normalized attention coefficient nodes v and u.

All models were implemented with two layers, hidden dimension 64 and drop out 0.2. Two training regimes were applied. In the backpropagation regime, weights were optimized using Adam with learning rate 0.01, weight decay 5×10^{-4} and cross-entropy loss. Early stopping was triggered by validation ROC-AUC with patience of 30 epochs and a maximum of 200 epochs [20], [22]. In the no-backpropagation regime, models were initialized randomly but not updated. The evaluation covered both predictive performance and fairness. Predictive measures included accuracy and ROC-AUC [22], [23]. Fairness was assessed through DP, EOpp, and EOdds, computed as in [10], [11].

4. RESULTS AND DISCUSSION

An experiment was conducted across three different architectures to evaluate the interaction between the perfomance and fairness of those architectures under two training regimes. The perfomance was assessed different graph construction and fairness diagnostics. Table 1 (in Appendix) summarizes the aggregated results to compare the perfomance (ROC-AUC) and fairness through of three different architectures. The sensivity of the results to graph construction was assessed through varying neighbourhood sizes K {5,10,20} and edge weighting strategies. The robustness of the results was gauged through confidence intervals.

The results in Table 1 (in Appendix) shows that training with back propagation improves accuracy as reflected by ROC-AUC compared to the untrained counterparts, but it is accompanied by increased fairness disparities too as reflected by DP, Eodds, Eopps. To analyze the sensivity to graph construction neighboorhood size K was varied through three different values (5,10,20) along with weighting schemes and it was observed that with the increase of K the accuracy gets improved due to more information propagatio along with that the increase of k lead to slight decrease in fairness gap as big neighborhood size causes over smoothening which resuces the fairness disparity.

Edge weighting further modulates the behaviour. The similarity based weighting schemes like (cosine, rbf) amplifies the influence the structurally similar nodes causing node level bias to get injected when demographic properties correlate with feature similarity. Inspite of the variations the relative ranking of the models and the tradeoff between accuracy and fairness remained constant which showed the weighting schemes had no effect on the results.

For all the three models in absense of back propagation their predictions became weak, as they produced uniform prediction scores resulting in similar prediction and error rates across all demograhic groups. Consequently the group fairness metrics appear small though predcitions were weak because they rose from uniformative predictions which failed to discriminate between positive and negative outcomes thus masking underlying disparities. On the other side with backpropagation the cross entropy loss gets optimized resulting in the seperation between the postive and negative samples become sharpened which causes ROC-AUC to increase. However due to training the correlations between predictive feature and sensitivity attributes get amplified resulting in decision rates and error rates get diverted which leads to increased fairness gap.

Across all the graph architectures GraphSAGE exhibited best predictive perfomance but its shows increased fairness disparities. The inductive bias and aggregation (mean/LSTM/pool) across neighbors optimize the ranking perfomance resulting in providing good accuracy, but as aggregation remains uniform across neighbors which lacks mechanism to suppress biased signals causing fairness disparities to occur. GAT achives good accuracy while keeping low DP and equalized odd gaps, indicating fairness roubustness among the three. The attention mechanism of the GAT causes down weight bias propagating neighbors leads to comparatively lower fairness disparities over other two architectures. The bias modulation due to attention mechanism make GAT favourable for good accuracy fairness trade off. Whereas GCN due to uniform aggregation amplifies the majority group information which improves the overall discrimination but makes the fairness disparity worse. As a result GCN is neither good for highest accuracy nor good as fairness robust model. Overall GAT offers most balances trade-off between fairness and accuracy, while GraphSAGE offers highest accuracy and GCN provides limited improvement scope both for fairness and accuracy.

The Table 2 summarizes additional fairness metrics that capture the aspects not covered by DP/EOpps/EOdds based gaps such as calibration error using expected calibration error (ECE), predictive parity

using ΔPPV and subgroup fairness using SubgroupGap_DP. These metrics help to assess the consistency of model confidence and precision across different demographic groups under both trained and untrained regimes.

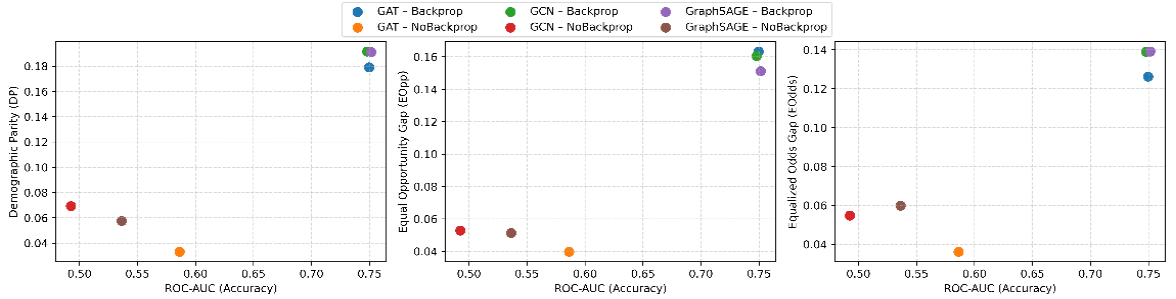
Table 2. Calibration, predictive parity, and subgroup fairness evaluation of GCN, GraphSAGE, and GAT on the COMPAS dataset under two training regimes

Model	Training	Weighting	K	ECE	ΔPPV	SubgroupGap_DP
GAT	Backprop	cosine	5	0.232023	0.099937	0.104429
GAT	Backprop	cosine	10	0.228423	0.102523	0.103904
GAT	Backprop	cosine	20	0.225415	0.106820	0.097748
GAT	Backprop	rbf	5	0.232023	0.099937	0.104429
GAT	Backprop	rbf	10	0.228423	0.102523	0.103904
GAT	Backprop	rbf	20	0.225415	0.106820	0.097748
GAT	Backprop	unweighted	5	0.232023	0.099937	0.104429
GAT	Backprop	unweighted	10	0.228423	0.102523	0.103904
GAT	Backprop	unweighted	20	0.225415	0.106820	0.097748
GAT	NoBackprop	cosine	5	0.165123	0.164984	0.019294
GAT	NoBackprop	cosine	10	0.165513	0.180237	0.017192
GAT	NoBackprop	cosine	20	0.160874	0.167541	0.015240
GAT	NoBackprop	rbf	5	0.165123	0.164984	0.019294
GAT	NoBackprop	rbf	10	0.165513	0.180237	0.017192
GAT	NoBackprop	rbf	20	0.160874	0.167541	0.015240
GAT	NoBackprop	unweighted	5	0.165123	0.164984	0.019294
GAT	NoBackprop	unweighted	10	0.165513	0.180237	0.017192
GAT	NoBackprop	unweighted	20	0.160874	0.167541	0.015240
GCN	Backprop	cosine	5	0.223134	0.097296	0.110435
GCN	Backprop	cosine	10	0.220542	0.093943	0.106381
GCN	Backprop	cosine	20	0.218240	0.103303	0.101276
GCN	Backprop	rbf	5	0.231031	0.079354	0.109985
GCN	Backprop	rbf	10	0.228987	0.077877	0.109309
GCN	Backprop	rbf	20	0.223540	0.078925	0.108634
GCN	Backprop	unweighted	5	0.224328	0.089718	0.111712
GCN	Backprop	unweighted	10	0.220568	0.101718	0.105556
GCN	Backprop	unweighted	20	0.216731	0.100986	0.103378
GCN	NoBackprop	cosine	5	0.143680	0.169679	0.040390
GCN	NoBackprop	cosine	10	0.143041	0.170488	0.040240
GCN	NoBackprop	cosine	20	0.147634	0.170902	0.037538
GCN	NoBackprop	rbf	5	0.145323	0.166961	0.039715
GCN	NoBackprop	rbf	10	0.145270	0.165139	0.040090
GCN	NoBackprop	rbf	20	0.146688	0.163207	0.040541
GCN	NoBackprop	unweighted	5	0.143688	0.169679	0.040390
GCN	NoBackprop	unweighted	10	0.142834	0.166726	0.039565
GCN	NoBackprop	unweighted	20	0.146802	0.173090	0.037538
GraphSAGE	Backprop	cosine	5	0.229155	0.081991	0.111486
GraphSAGE	Backprop	cosine	10	0.225454	0.090314	0.106306
GraphSAGE	Backprop	cosine	20	0.227659	0.100281	0.102102
GraphSAGE	Backprop	rbf	5	0.229155	0.081991	0.111486
GraphSAGE	Backprop	rbf	10	0.225454	0.090314	0.106306
GraphSAGE	Backprop	rbf	20	0.227659	0.100281	0.102102
GraphSAGE	Backprop	unweighted	5	0.229155	0.081991	0.111486
GraphSAGE	Backprop	unweighted	10	0.225454	0.090314	0.106306
GraphSAGE	Backprop	unweighted	20	0.227659	0.100281	0.102102
GraphSAGE	NoBackprop	cosine	5	0.112164	0.165592	0.033559
GraphSAGE	NoBackprop	cosine	10	0.111722	0.178630	0.032883
GraphSAGE	NoBackprop	cosine	20	0.110962	0.171490	0.031456
GraphSAGE	NoBackprop	rbf	5	0.112164	0.165592	0.033559
GraphSAGE	NoBackprop	rbf	10	0.111722	0.178630	0.032883
GraphSAGE	NoBackprop	rbf	20	0.110962	0.171490	0.031456
GraphSAGE	NoBackprop	unweighted	5	0.112164	0.165592	0.033559
GraphSAGE	NoBackprop	unweighted	10	0.111722	0.178630	0.032883
GraphSAGE	NoBackprop	unweighted	20	0.110962	0.171490	0.031456

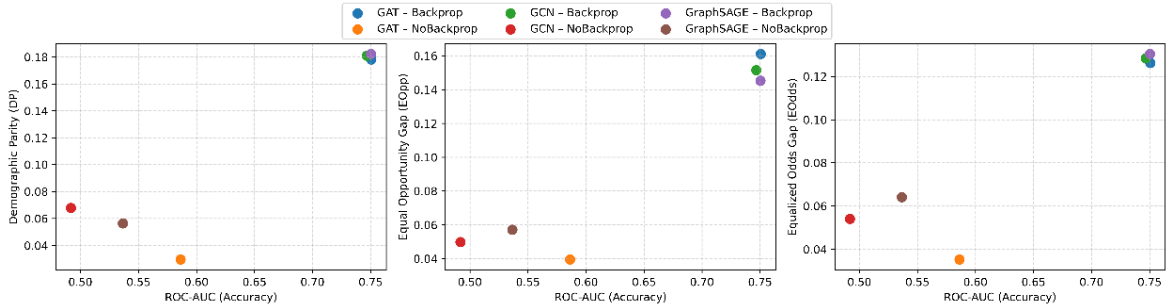
Across all the models when untrained produced prediction around 0.5 leading to less calibration [24] gap. While after training due to optimization of cross entropy loss the predictions become confident; when the confidence rises faster than emperical correctness the calibration error increases [25]. From the result it is observed that predictive parity [28] gaps decreases after training. When models remained untrained positive predictions become weak leading to large predictive parity gaps. After training the predictions become discriminative, the predcitive postives correspond to true high risk users in both groups. This reduces condition precision resulting into lower ΔPPV. In case of Subgroupgap_DP [10] for untrained models the gap

is samll as while weights are randomly initialized predictions become uniform across all groups. After training the correlation between predictive feature and the sensitive attributes gets amplified leading to the predcition rates and error rates across the groups get diverged causing subgroupgap disparity [15].

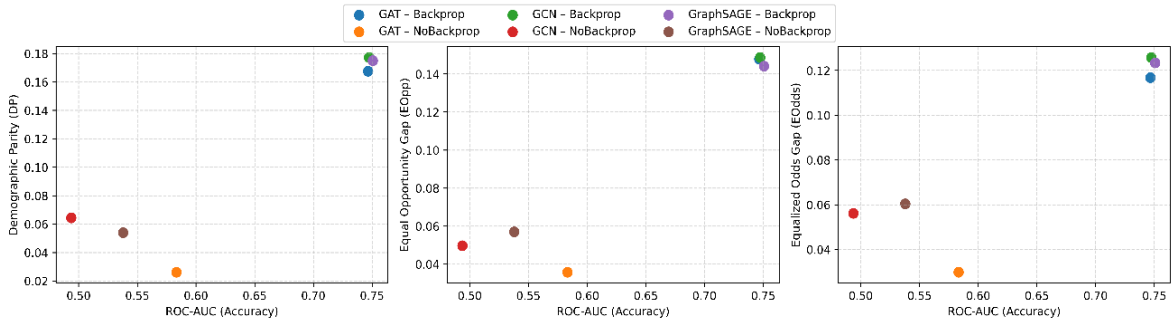
Figure 1 represents the accuracy-fairness trade off across all the models for differnt neighborhood sizes where $k \in \{5, 10, 20\}$ under trained and trained regimes, outlining the relationship between ROC-AUC and fairness metrices which includes DP, EOpp, and Eodds as shown in Figures 1(a)-(c). The figure shows that optimization through backpropagation improves predcitive performance but in tandem increases the group disparities confirming that fairness is not only dependent on architecrual differences but also varies because of dynamic training. Moreover it is also revelaed from the figure that increase in neighborhood size k slightly stabilizes the fairness gaps without effecting the accuracy indicating that graph construction influnces bias diffusion while message passing [30]. Backpropagation shifts the models towards higher ROC-AUC, indicating improved ranking performance [22], but in tandem the fairness gap gets increased indicating larger gaps for DP, EOpp, EOdds [11] while the untrained models cause lower fairness gaps because of weak, random prediction. Among the trained models GAT achived comparable accuracy with GraphSAGE while producing comparable lesser gaps among three models [18]. The attention mechanism helps in mitigating the bias propagation resulting in lesser fairness disparity in case of GAT [9]. For every model it is observed from the Figure 1. That with the increase of k ROC-AUC increases and simultaneously the fairness gap reduces [30], [31]. Therefore larger neighborhood size acts as somewhat debiasing that can be presumed.



(a)



(b)



(c)

Figure 1. Accuracy-fairness trade-off across different neighborhood sizes: (a) k=5, (b) k=10, and (c) k=20. Each subplot illustrates the relationship between ROC-AUC and group fairness metrices which are DP, EOpp, EOdds across trained and untrained settings

The findings from this study has practical implicatons specially in judicial risk assessment contexts such as COMPAS [1], [29]. Elevated fairness gaps implies unqual false postive and false negative rates across all demographic groups leading to disproportionate detention or release [3]. Even smaller fainess gap can scale into considerable societal harm when deployed at population level [28]. Results from this investigation exemplifies accucary centric optimization if left unchecked can strengthens existing structural inequities ingrained in social and criminal systems [14], [21], [30]. Although the emperical results suggest that both GAT and larger neighborhood size mitigate bias implicitly but those are not explicit mitigation mechanisms. Several debiasing strategies have been worked upon by the researchers. One of them is Adverserial debiasing which mitigates the information regarding sensitive attributes from the learned representations [13]. Fairness aware loss functions aims to penalizes the disparities across sensitve groups by adding a regularization term [13], [15], [21]. Another methodology Fair GNN focused on building unbiased GNN models using to techniques like adverserial debiasing, fairness constraint mainitaining improved classification accuracy utilizing graph architecture and allowing limited sensitive information [16].

The work documented here highlights a tradeoff between accuracy and fairness by comparing three canonical architectures GCN, GraphSAGE, and GAT under trained and untrained regimes on a dataset of judicial risk assessment [29]. The results indicated backpropagation improved perfomence for all three models but it was accompanied with increased DP, Eopp, and EOdds gaps [11]. This corroborates that enhancement in accuracy amplify the disparities embedded in the structures of GNN [15].

The takeaways from study was lower fairness gap in untrained models don't denote equal decision making rather it comes from weak, random predictions [23]. The outcomes of such models yield poor ranking perfomance which artificially repress group conditioned disparities [21]. This shows the significance of measuring predictive perfomance along with fairness in high stake areas like judicial assessment where shallowly fair but inaccurate decision can cause harmful effects [3].

The comparative study also revealed significant architectural dissimilarity. Although GCN and GraphSAGE are efficient in predictive accuracy after training but at the same time they cause fairness to increase due to their uniform or inductive aggregation mechanism leads to passing of group correlated signals across the graph [7], [8]. In compare to that attention mechanism of GAT causes to supress bias propapagating signal and at the same time allow informative signal to pass which makes GAT to maintain a balanced predictive perfomance and fair decision making [9], [18]. Even though it is not explicitly debiasing method but it does show that structural differences play a pivotal role in fair dynamics.

The sensitivity analysis reveals that the fairness gets better without affecting accuracy with increase in the number of neighborhood sizes [31]. This fact of graph smoothening shows that graph construction parameters affect diffusion of biases during message passing which supports the need to investigate graph design as an aspect of fairness evaluation [6]. Nevertehless the continued existance of disparities evn in bigger neighborhood sizes show that such implicit effect can't stand alone [15].

Practically the findings from this study have direct involvement in judicial decision support systems [26]. Unequal error rates or acceptance probabilities can be tranfomed into unequal dtention or release decision across different demographic groups which can strengthen the existing inequity in the society [3]. It can be deduced from the results that accuracy centric optimization shouldn't be deployed without fairness protection, and highlights that it is imperative to measure fairness in realistic scenerio [14], [21].

The novelty of this investigation lies in its comprehensive heuristical characterization of fairness dynamics under trained and untrained graph neural architectures, coupled with sensitivity analysis on graph construction options and various complimentory fairness metrices, including calibration related measures [24], [25]. Instead of suggesting a new debiasing methodology this research sheds light on how learning dynamics, GNN architecture and neighborhood size influnce fairness outcome. These findings pave the way for further reserach on fairness aware learning, and encourages incorporation of adverserial debiasing and fairness controlled objectives in message passing framework [16], [27].

5. CONCLUSION

The work demonstrates an examination which highlights an interplay between predictive accuracy and fairness in GNN applied to a judicial risk assessment dataset. Performance of GCN, GraphSAGE, and GAT were compared under trained and untrained settings and the outcomes exhibited that perfomance for each of these three models were improved after backpropogation but that was accompanied with increased fairness gaps in three fairness metrices. This result indicates that the fairness degradation is not inicidental but it arises due to structural learning dynamics of graph based models. The takeway from this study is lower fairness gap observed in untrained models should not be interpreted as an equitable decision making rather these models show limited discriminative power and weak ranking perfomance which repress the group difference. This shows the importance of evaluation of accuracy along with fairness otheriwse inaccurate but apparently fair decision can cause social harm.

A comparative analysis among three canonical architectures was perfomed which showed how fairness outcomes vary based on structural differences. GCN and GraphSAGE showed strong predictive performance after training but they exhibit fairness degradation due to uniform or inductive aggregation mechanism which cause group correlated signals pass across the graphs. In contrast GAT was able to attain a good balance between accuracy and fairness. The attention mechanism of GAT made the informative message to pass across the graph while down weighting the bias propagating signals. Sensitivity analysis further shows that increase in neighborhood size leads to decrease in fairness gaps without affecting the accuracy, although these implicit effects remain insufficient in abolishing the disparities. From practical scenerio these findings from this study carry significant implications in judicial decision making where unequal error rates might lead to unequal detention or release decisions. The outcomes wary against accuracy centric optimization without fairness safeguards, and highlights the necessity of fairness determination under realistic conditions. The novelty of this investigation lies in its empirical characterization of fairness analysis across different demographic groups coupled with sensitivity analysis over graph construction options and various set of fairness metrices, including calibration-based measures. Instead of suggesting a new debiasing algorithm this work proposes how learning dynamics, structural choices, and neighborhood size help in shaping the fairness behaviour. These findings underscore the requirement of further research on fairness aware learning and incorporation of adversarial debiasing with fairness regulated tasks in message passing system.

ACKNOWLEDGMENTS

The authors acknowledge Alliance University for offering appropriate academic environment which facilitated this research.

FUNDING INFORMATION

It is hereby declared by the authors that no funding has been received to conduct this research work.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ananda Chatterjee		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓			
K. A. Venkatesh	✓	✓		✓	✓	✓				✓	✓	✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration

CONFLICT OF INTEREST STATEMENT

The authors are hereby reconfirming that no conflict of interest is linked with this publication and no financial sponsorship has been provided to influence the research outcome. It is testified that the manuscript has been read and endorsed by all the listed authors.

INFORMED CONSENT

The datasets used in this study is publicly available, containing anonymized records and have no personally identifiable information. No interaction or intervention with subjects was done. Data analysis was performed to conduct academic research. Since the datasets are publicly available to be used by scientists, with the prior agreement of the providers of the data, no separate individual informed consent was necessary.

ETHICAL APPROVAL

This study uses publicly available COMPAS dataset released by ProPublica investigations “Machine Bias” [1] which does not consist of any personal identifiable information. No human or animal subjects were recruited, and no private or institutional data was collected to conduct this investigation, therefore formal approval by a review board was not necessary according to our institutional policy.

DATA AVAILABILITY

The data corroborates the outcomes of this investigation are openly available as part of the ProPublica investigation "Machine Bias" [1]. The COMPAS recidivism dataset used in this work can be accessed at: https://raw.githubusercontent.com/propublica/compas-analysis/master/compas-scores-two-years.csv. No new data were generated in the course of this research.

REFERENCES

[1] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," in *Ethics of Data and Analytics*, Auerbach Publications, 2022, pp. 254–264.

[2] A. Chouldechova, "Fair prediction with disparate impact: a study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017, doi: 10.1089/big.2016.0047.

[3] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2017, pp. 797–806, doi: 10.1145/3097983.3098095.

[4] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4–24, Jan. 2021, doi: 10.1109/tnnls.2020.2978386.

[5] J. Zhou *et al.*, "Graph neural networks: A review of methods and applications," *AI Open*, vol. 1, pp. 57–81, 2020, doi: 10.1016/j.aiopen.2021.01.001.

[6] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: grids, groups, graphs, geodesics, and gauges," *arXiv preprint arXiv:2104.13478*, 2021.

[7] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[8] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[9] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio, "Graph attention networks," *arXiv preprint arXiv:1710.10903*.

[10] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, Jan. 2012, pp. 214–226, doi: 10.1145/2090236.2090255.

[11] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[12] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, Dec. 2011, doi: 10.1007/s10115-011-0463-8.

[13] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness beyond disparate treatment and disparate impact: learning classification without disparate mistreatment," in *Proceedings of the 26th International Conference on World Wide Web*, Apr. 2017, pp. 1171–1180, doi: 10.1145/3038912.3052660.

[14] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021, doi: 10.1145/3457607.

[15] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.

[16] A. Chen *et al.*, "Fairness-aware graph neural networks: a survey," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 6, pp. 1–23, Apr. 2024, doi: 10.1145/3649142.

[17] C. Agarwal, H. Lakkaraju, and M. Zitnik, "Towards a unified framework for fair and stable graph representation learning," *In Uncertainty in artificial intelligence*, pp. 2114–2124, 2021.

[18] O. D. Kose and Y. Shen, "FairGAT: fairness-aware graph attention networks," *ACM Transactions on Knowledge Discovery from Data*, vol. 18, no. 7, pp. 1–20, 2024, doi: 10.1145/3645096.

[19] J. Zhao *et al.*, "FS-GNN: improving fairness in graph neural networks via joint sparsification," *Neurocomputing*, vol. 648, p. 130641, Oct. 2025, doi: 10.1016/j.neucom.2025.130641.

[20] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[21] I. Spinelli, R. Bianchini, and S. Scardapane, "Drop edges and adapt: A fairness enforcing fine-tuning for graph neural networks," *Neural Networks*, vol. 167, pp. 159–167, Oct. 2023, doi: 10.1016/j.neunet.2023.08.002.

[22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, Jun. 2006, doi: 10.1016/j.patrec.2005.10.010.

[23] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982, doi: 10.1148/radiology.143.1.7063747.

[24] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *In International conference on machine learning*, pp. 1321–1330, 2017.

[25] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Advances in neural information processing systems*, vol. 30, 2017.

[26] T. Kirat, O. Tambou, V. Do, and A. Tsoukiàs, "Fairness and explainability in automatic decision-making systems. A challenge for computer science and law," *EURO Journal on Decision Processes*, vol. 11, p. 100036, 2023, doi: 10.1016/j.ejdp.2023.100036.

[27] E. Dai and S. Wang, "Learning fair graph neural networks with limited and private sensitive attribute information," *IEEE Transactions on Knowledge and Data Engineering*, pp. 1–14, 2022, doi: 10.1109/tkde.2022.3197554.

[28] Z. Cong, B. Shi, S. Li, J. Yang, Q. He, and J. Pei, "FairSample: training fair and accurate graph convolutional neural networks efficiently," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 4, pp. 1537–1551, Apr. 2024, doi: 10.1109/tkde.2023.3306378.

[29] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, no. 1, Jan. 2018, doi: 10.1126/sciadv.aao5580.

[30] W. Yu, X. Lin, J. Liu, J. Ge, W. Ou, and Z. Qin, "Self-propagation graph neural network for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 12, pp. 5993–6002, Dec. 2022, doi: 10.1109/tkde.2021.3076772.

[31] M. Fey and J. E. Lenssen, "Fast graph representation learning with PyTorch Geometric," *arXiv preprint arXiv:1903.02428*, 2019.

APPENDIX

Table 1. The performance and fairness of GCN, GraphSAGE and GAT on COMPAS dataset under two training regimes

Model	Training	Weighting	K	ROC_AUC (95% CI)	DP (95% CI)	EOpps (95% CI)	EOdds (95% CI)
GAT	Backprop	cosine	5	0.75[0.718, 0.782]	0.179[0.123, 0.243]	0.163[0.084, 0.278]	0.126[0.079, 0.201]
GAT	Backprop	cosine	10	0.751[0.719, 0.782]	0.178[0.098, 0.218]	0.161[0.067, 0.263]	0.126[0.074, 0.193]
GAT	Backprop	cosine	20	0.747[0.713, 0.777]	0.167[0.131, 0.254]	0.147[0.049, 0.239]	0.116[0.049, 0.170]
GAT	Backprop	rbf	5	0.75[0.718, 0.782]	0.179[0.123, 0.243]	0.163[0.084, 0.278]	0.126[0.074, 0.193]
GAT	Backprop	rbf	10	0.751[0.719, 0.782]	0.178[0.098, 0.218]	0.161[0.067, 0.263]	0.126 [0.049, 0.170]
GAT	Backprop	rbf	20	0.747[0.713, 0.777]	0.167[0.131, 0.254]	0.147[0.049, 0.239]	0.116[0.049, 0.170]
GAT	Backprop	unweighted	5	0.75[0.718, 0.782]	0.179[0.123, 0.243]	0.163[0.084, 0.278]	0.126[0.074, 0.193]
GAT	Backprop	unweighted	10	0.751[0.719, 0.782]	0.178[0.098, 0.218]	0.161[0.084, 0.278]	0.126[0.074, 0.193]
GAT	Backprop	unweighted	20	0.747[0.713, 0.777]	0.167[0.001, 0.076]	0.161[0.067, 0.263]	0.116[0.049,0.17]
GAT	NoBackprop	cosine	5	0.586[0.498, 0.574]	0.033[0.001, 0.067]	0.147[0.049, 0.239]	0.036[0.018, 0.103]
GAT	NoBackprop	cosine	10	0.586[0.498, 0.574]	0.029[0.002, 0.069]	0.039[0.008, 0.152]	0.035[0.018, 0.097]
GAT	NoBackprop	cosine	20	0.583[0.493, 0.569]	0.026[0.001, 0.076]	0.039[0.006, 0.139]	0.029[0.015, 0.096]
GAT	NoBackprop	rbf	5	0.586[0.498, 0.574]	0.033[0.001, 0.067]	0.035[0.005, 0.139]	0.036[0.018, 0.103]
GAT	NoBackprop	rbf	10	0.586[0.498, 0.574]	0.029[0.002, 0.069]	0.039[0.008, 0.152]	0.035[0.018, 0.097]
GAT	NoBackprop	rbf	20	0.583[0.493, 0.569]	0.026[0.001, 0.076]	0.039[0.006, 0.139]	0.029[0.015, 0.096]
GAT	NoBackprop	unweighted	5	0.586[0.498, 0.574]	0.033[0.001, 0.067]	0.035[0.005, 0.139]	0.036[0.018, 0.103]
GAT	NoBackprop	unweighted	10	0.586[0.498, 0.574]	0.029[0.002, 0.069]	0.039[0.006, 0.139]	0.035[0.018, 0.097]
GAT	NoBackprop	unweighted	20	0.583[0.493, 0.569]	0.026[0.117, 0.247]	0.035[0.005, 0.139]	0.029[0.015, 0.096]
GCN	Backprop	cosine	5	0.748[0.713, 0.779]	0.189[0.104, 0.242]	0.163[0.064, 0.261]	0.137[0.070, 0.199]
GCN	Backprop	cosine	10	0.747[0.714, 0.779]	0.182[0.116, 0.249]	0.145[0.042, 0.232]	0.129[0.053, 0.188]
GCN	Backprop	cosine	20	0.747[0.716, 0.781]	0.173[0.116, 0.240]	0.146[0.049, 0.236]	0.122[0.074, 0.197]
GCN	Backprop	rbf	5	0.745[0.712, 0.776]	0.188[0.117, 0.249]	0.149[0.040, 0.227]	0.134[0.063, 0.188]
GCN	Backprop	rbf	10	0.743[0.711, 0.775]	0.187[0.116, 0.242]	0.143[0.049, 0.234]	0.133[0.065, 0.197]
GCN	Backprop	rbf	20	0.745[0.713, 0.778]	0.186[0.120, 0.249]	0.143[0.056, 0.247]	0.133[0.068, 0.193]
GCN	Backprop	unweighted	5	0.748[0.713, 0.779]	0.191[0.106, 0.240]	0.160[0.062, 0.257]	0.138[0.073, 0.201]
GCN	Backprop	unweighted	10	0.747[0.714, 0.779]	0.181[0.117, 0.248]	0.152[0.042, 0.232]	0.128[0.055, 0.188]
GCN	Backprop	unweighted	20	0.747[0.716, 0.781]	0.177[0.010, 0.110]	0.149[0.055, 0.240]	0.125[0.074, 0.196]
GCN	NoBackprop	cosine	5	0.493[0.606, 0.675]	0.069[0.006, 0.102]	0.053[0.011, 0.170]	0.054[0.016, 0.104]
GCN	NoBackprop	cosine	10	0.492[0.607, 0.676]	0.068[0.003, 0.098]	0.052[0.008, 0.159]	0.054[0.014, 0.099]
GCN	NoBackprop	cosine	20	0.494[0.607, 0.677]	0.064[0.007, 0.104]	0.051[0.006, 0.145]	0.056[0.009, 0.093]
GCN	NoBackprop	rbf	5	0.489[0.601, 0.669]	0.068[0.007, 0.106]	0.048[0.006, 0.156]	0.053[0.012, 0.097]
GCN	NoBackprop	rbf	10	0.489[0.602, 0.670]	0.069[0.011, 0.111]	0.049[0.007, 0.161]	0.053[0.013, 0.100]
GCN	NoBackprop	rbf	20	0.491[0.601, 0.669]	0.069[0.010, 0.110]	0.052[0.011, 0.170]	0.054[0.015, 0.103]
GCN	NoBackprop	unweighted	5	0.493[0.606, 0.675]	0.068[0.006, 0.103]	0.049[0.007, 0.155]	0.054[0.016, 0.104]
GCN	NoBackprop	unweighted	10	0.492[0.607, 0.675]	0.064[0.003, 0.098]	0.049[0.006, 0.145]	0.054[0.012, 0.098]

Table 1. The performance and fairness of GCN, GraphSAGE and GAT on COMPAS dataset under two training regimes (*Continued*)

Model	Training	Weighting	K	ROC_AUC (95% CI)	DP (95% CI)	EOpps (95% CI)	EOdds (95% CI)
GCN	NoBackprop	unweighted	20	0.494[0.607, 0.677]	0.191[0.130, 0.257]	0.151[0.044, 0.238]	0.056[0.009, 0.093]
GraphSAGE	Backprop	cosine	5	0.751[0.720, 0.785]	0.182[0.123, 0.249]	0.145[0.050, 0.244]	0.139[0.075, 0.204]
GraphSAGE	Backprop	cosine	10	0.75[0.718, 0.782]	0.175[0.115, 0.241]	0.144[0.054, 0.241]	0.130[0.074, 0.202]
GraphSAGE	Backprop	cosine	20	0.751[0.718, 0.781]	0.191[0.130, 0.257]	0.151[0.044, 0.238]	0.123[0.069, 0.190]
GraphSAGE	Backprop	rbf	5	0.751[0.720, 0.785]	0.182[0.123, 0.249]	0.145[0.050, 0.244]	0.139[0.075, 0.204]
GraphSAGE	Backprop	rbf	10	0.75[0.718, 0.782]	0.175[0.115, 0.241]	0.144[0.054, 0.241]	0.130[0.074, 0.202]
Model	Training	Weighting	K	ROC_AUC (95% CI)	DP (95% CI)	EOpps (95% CI)	EOdds (95% CI)
GraphSAGE	Backprop	rbf	20	0.751[0.718, 0.781]	0.191[0.130, 0.257]	0.151[0.044, 0.238]	0.123[0.069, 0.190]
GraphSAGE	Backprop	unweighted	5	0.751[0.720, 0.785]	0.182[0.123, 0.249]	0.145 [0.050, 0.244]	0.139[0.075, 0.204]
GraphSAGE	Backprop	unweighted	10	0.75[0.718, 0.782]	0.175[0.115, 0.241]	0.144[0.054, 0.241]	0.130[0.074, 0.202]
GraphSAGE	Backprop	unweighted	20	0.751[0.718, 0.781]	0.057[0.007, 0.092]	0.051[0.004, 0.113]	0.123[0.069, 0.190]
GraphSAGE	NoBackprop	cosine	5	0.536[0.511, 0.584]	0.056[0.013, 0.094]	0.057[0.004, 0.127]	0.059[0.011, 0.089]
GraphSAGE	NoBackprop	cosine	10	0.537[0.510, 0.584]	0.054[0.013, 0.093]	0.057[0.004, 0.123]	0.064[0.017, 0.095]
GraphSAGE	NoBackprop	cosine	20	0.538[0.510, 0.585]	0.057[0.007, 0.092]	0.051[0.004, 0.113]	0.060[0.016, 0.093]
GraphSAGE	NoBackprop	rbf	5	0.536[0.511, 0.584]	0.056[0.013, 0.094]	0.057[0.004, 0.127]	0.059[0.011, 0.089]
GraphSAGE	NoBackprop	rbf	10	0.537[0.510, 0.584]	0.054[0.013, 0.093]	0.057[0.004, 0.123]	0.064[0.017, 0.095]
GraphSAGE	NoBackprop	rbf	20	0.538[0.510, 0.585]	0.057[0.007, 0.092]	0.051[0.004, 0.123]	0.060[0.016, 0.093]
GraphSAGE	NoBackprop	unweighted	5	0.536[0.511, 0.584]	0.056[0.013, 0.094]	0.057[0.004, 0.113]	0.059[0.011, 0.089]
GraphSAGE	NoBackprop	unweighted	10	0.537[0.510, 0.584]	0.054[0.013, 0.093]	0.056[0.004, 0.127]	0.064[0.017, 0.095]
GraphSAGE	NoBackprop	unweighted	20	0.538[0.510, 0.585]	0.054[0.013, 0.093]	0.05[0.004, 0.123]	0.060[0.016, 0.093]

BIOGRAPHIES OF AUTHORS

Ananda Chatterjee is a Ph.D. scholar in applied mathematics at Alliance University, Bengaluru, India specializing in fairness-aware machine learning and geometric deep learning. He has a total of nine years of industry experience, with the last five and a half years devoted to working as a data scientist in the financial domain, including banking and fintech organizations. His core expertise spans machine learning, deep learning, time series modeling, and computer vision, with hands-on experience in designing AI-driven decision systems for financial analytics. His academic qualifications include a B.Tech. in Electronics and Instrumentation, an M.Tech. in instrumentation and control, and a post graduate program in data science. He is the author of six research papers presented at national and international conferences and one book chapter, with one paper currently under publication. He can be contacted at email: canandaPHD724@sam.alliance.edu.in.

Dr. K. A. Venkatesh is a professor of Mathematics and Computer Science, Alliance University, Bengaluru, India with over 32 years of teaching experience at both undergraduate and postgraduate levels. He is a versatile researcher, having contributed significantly to various disciplines. Driven by a passion for education, he possesses a distinguished record of publications, including over 20 book chapters and 74 journal articles. His academic qualifications include a Ph.D. in Combinatorial Optimization, an MPhil degree, M.Sc. degree, and M.Tech. (Data Science and Engineering) from Bits Pilani. Dr. Venkatesh is an active member of several prestigious organizations, including the Ramanujan Mathematical Society, System Society of India and the Academy of Discrete Mathematics and Applications. He is also affiliated with IEEE and serves as a member in association for constraint programming and outreach member in SIAM, USA. He can be contacted at email: ka.venkatesh@alliance.edu.in.