Model-Agnostic Fairness Regularization for GNNs with Incomplete Sensitive Information

M. Tavassoli Kejani[1], F. Dornaika[*2, 3], and J. M. Loubes[1,4]

[1]*Institut de Mathématiques de Toulouse*, [2]*University of the Basque Country*, [3]*IKERBASQUE*, [4]*INRIA, Regalia Team*,
mahdi.tavassoli-kejani@univ-toulouse.fr, fadi.dornaika@ehu.eus, loubes@math.univ-toulouse.fr

Abstract

Graph Neural Networks (GNNs) have demonstrated exceptional efficacy in relational learning tasks, including node classification and link prediction. However, their application raises significant fairness concerns, as GNNs can perpetuate and even amplify societal biases against protected groups defined by sensitive attributes such as race or gender. These biases are often inherent in the node features, structural topology, and message-passing mechanisms of the graph itself. A critical limitation of existing fairness-aware GNN methods is their reliance on the strong assumption that sensitive attributes are fully available for all nodes during training—a condition that poses a practical impediment due to privacy concerns and data collection constraints. To address this gap, we propose a novel, model-agnostic fairness regularization framework designed for the realistic scenario where sensitive attributes are only partially available. Our approach formalizes a fairness-aware objective function that integrates both*equal opportunity* and *statistical parity* as differentiable regularization terms. Through a comprehensive empirical evaluation across five real-world benchmark datasets, we demonstrate that the proposed method significantly mitigates bias across key fairness metrics while maintaining competitive node classification performance. Results show that our framework consistently outperforms baseline models in achieving a favorable fairness-accuracy trade-off, with minimal degradation in predictive accuracy. The datasets and source code will be publicly released at https://github.com/mtavassoli/GNN-FC.

***Keywords*—** Graph Neural Networks (GNNs), Fairness, Semi-Supervised Classification, Equal Opportunity, Statistical Parity.

1 Introduction

In recent years, bias mitigation in AI-based algorithms has emerged as a critical concern, as surveyed in [1, 2]. Beyond ethical considerations, the European Union has introduced mandatory

*Corresponding author

regulations governing algorithms used in high-risk systems, which are defined as those with the potential to cause harm to health, safety, or fundamental rights [1]. These regulations emphasize the need to ensure that such systems do not infringe upon fundamental rights, particularly with respect to preventing discrimination and avoiding unfair biases prohibited by Union or national law.

Addressing bias in the development and evaluation of high-risk AI systems presents significant challenges. In some cases, it may be necessary to process special categories of personal data to enable bias detection and correction, provided this is done in strict accordance with legal conditions and includes appropriate safeguards to protect fundamental rights and privacy. Nevertheless, bias mitigation efforts are often hindered by the unavailability of key sensitive attributes, such as ethnic origin and gender. While the law may permit the collection of such data under certain circumstances to enhance fairness and comply with data protection regulations, these sensitive attributes are frequently scarce and difficult to obtain. For instance, only 14% of teenage users make their full profiles public on Facebook [3], limiting the availability of necessary information for bias detection.

One class of AI-based models where concerns regarding bias and fairness are increasingly pertinent is the Graph Neural Networks (GNNs) [4]. GNNs are a subset of deep learning models designed specifically to operate on graph-structured data, enabling the capture of intricate relational and structural information. Due to their capacity for modeling complex dependencies, GNNs have become useful in various high-stakes applications, including social networks [5] and recommendation systems [6]. Consequently, addressing issues of bias and fairness in GNNs is of paramount importance. Machine learning algorithms often inherit and can amplify biases present in training data. These biases have been extensively studied in recent years, as discussed in [7, 8, 9] and references therein. For graph neural networks, biases may arise from several sources, including node attributes (referred to as attribute bias in [10]), the connections between nodes (structural bias in [10]), and the message-passing mechanisms employed in GNNs (potential bias in [10]), all of which can result in unfair or biased predictions. For example, in book recommendation systems leveraging GNNs, the model might exhibit a bias towards recommending books by male authors [11]. Therefore, ensuring the fairness, safety, and reliability of GNN models and their predictions is critical, making this a vital area of ongoing research.

Many methods have been proposed to reduce the effects of bias in traditional machine learning algorithms. In the context of GNNs, existing approaches for bias mitigation remain limited. Several fairness-aware GNN methods have been proposed [12, 13, 14, 15, 16, 17, 18, 19], which can be broadly categorized into four groups [20]: (1) **pre-processing methods**, which remove bias before applying GNNs, such as [21]; (2) **in-training methods**, which address bias during GNN training, like [22]; (3) **post-processing methods**, which remove bias from the resulting GNN embeddings, as in [23]; and (4) **hybrid methods**, which combine two or more of the aforementioned techniques, like [24].

While existing methods contribute to improving fairness in AI models, they often come at the cost of reduced accuracy and may even exacerbate it. Furthermore, many of these approaches require access to all sensitive attribute labels during the training phase [6], which is not feasible in practical real-world scenarios.

To address these challenges, we propose a novel in-training, model-agnostic approach, **E**qual **O**pportunity **S**tatistical **P**arity (EOSP), which operates with limited sensitive attribute labels (available only for labeled nodes) during training to enable fair learning in GNNs. To the best of our knowledge, this is the first work to integrate statistical parity and equal opportunity directly as regularization terms for bias mitigation in GNNs. The EOSP method incorporates these two key fairness constraints directly into the training process: a statistical parity term, which en-

[1] https://artificialintelligenceact.eu/

sures similar model behavior across demographic groups, and an equal opportunity term, which guarantees that errors are distributed fairly rather than enforcing identical decisions for everyone.

This approach contrasts with existing adversarial frameworks, such as LAFTR [25]. While LAFTR reduces the influence of sensitive attributes by training an adversary to predict them, our method directly incorporates fairness penalties. Furthermore, we specifically design and evaluate our approach for the graph learning domain, where the relational structure presents unique challenges for fairness that LAFTR was not designed to address.

The main contributions of the paper are listed below.

- **Model-Agnostic and Composable Framework:** We design a flexible, model-agnostic framework that can be seamlessly integrated into any standard GNN architecture and is composable with other fairness-aware techniques.
- **Effectiveness with Partial Sensitive Attributes:** Our method effectively mitigates bias when sensitive attribute labels are available only for labeled nodes, relaxing the strong and often unrealistic assumption of complete attribute knowledge required by prior methods.
- **Extensive Empirical Validation:** We demonstrate the effectiveness of our approach through comprehensive experiments on five real-world datasets, with particular emphasis on balanced accuracy, a fair performance metric often overlooked in prior fairness literature, alongside standard metrics such as F1-score and AUC.

The rest of the paper is structured as follows. We begin by reviewing related work in Section 2 and introducing necessary preliminaries in Section 3. Our proposed methodology is then detailed in Section 4, followed by experimental setup in section 5 and experimental results in Section 6. Finally, we conclude the paper in Section 7.

2 Related Work

The literature on fairness in machine learning offers various methods for bias mitigation, generally categorized into pre-processing of training data, in-processing with fairness constraints during model optimization, and post-processing of model outputs. Representative approaches include fairness-constrained optimization, latent space disentanglement, optimal transport, and counterfactual reasoning, as discussed in [26, 27, 28, 29, 30].

For graph-based learning, several studies [31, 32, 33, 34, 35] explore bias detection and mitigation in Graph Neural Networks (GNNs). According to [17], the main fairness notions considered are group fairness (balanced outcomes across sensitive groups), individual fairness (similar results for similar nodes), and counterfactual fairness (invariance of outcomes under counterfactual scenarios). This work focuses on two main directions of fair GNN: group fairness and counterfactual fairness.

Group Fairness in GNNs Group fairness ensures that demographic groups defined by sensitive attributes, such as gender, receive equitable treatment in learning tasks. For example, FairGNN [12] employs adversarial learning, where the GNN predicts target labels while an adversarial network simultaneously attempts to infer sensitive attributes from node embeddings. The GNN is trained to minimize task loss while reducing the adversary's predictive capability, thereby improving fairness. FairVGNN [10] enhances fairness by feature channels correlated with sensitive attributes and clamping weights to reduce the influence of sensitive information, supported by adversarial discriminators. FairDrop [21] proposes an edge-dropout scheme to mitigate homophily and improve fairness in graph representation learning. Fairwalk [14] introduce fair random walk strategies to learn equitable network embeddings, while BIND [36] proposes

the Probabilistic Distribution Disparity (PDD) metric to measure and eliminate bias-contributing nodes. FairSAD [37] separates sensitive attribute information into an independent component with a channel masking mechanism that adaptively identifies and isolates sensitive factors.

More recently, DAB-GNN [38] introduces a GNN framework that disentangles attribute, structural, and potential biases into separate node embeddings. A bias contrast optimizer ensures that these embeddings remain independent, while a fairness harmonizer aligns subgroup distributions within each bias-specific embedding to reduce the influence of sensitive attributes. FairSR [39] integrates structural rebalancing with adversarial learning to ensure equitable outcomes for nodes with varying degrees but similar functionalities in prediction tasks.

Counterfactual Fairness in GNNs Research on counterfactual fairness in GNNs aims to maintain consistent predictions between factual and counterfactual scenarios. For example, NIFTY [15] and GEAR [16] employ perturbation strategies and GraphVAE-based models to generate realistic counterfactuals and enforce fairness constraints. Despite progress, generating realistic counterfactuals remains challenging. CAF [17] addresses this by developing a structural causal model that improves fairness. FairGB [40] tackles unfairness through group balancing using real counterfactual node mixup and a contribution alignment loss. FCLCA [41] proposes a framework to learn counterfactual fairness by addressing graph structure bias. It generates two counterfactual graphs via structural augmentation and employs contrastive learning to maximize consistency between node representations across these augmented graphs. Additionally, FCLCA incorporates adversarial debiasing to further reduce the influence of sensitive attributes, yielding fairer node representations

In this paper, we propose a model-agnostic framework that enforces fairness by incorporating equal opportunity and statistical parity as regularization terms, which can be integrated into standard GNN architectures and combined with other fairness-aware techniques.

3 Background

3.1 Notations

Notations and Problem Setting. We consider a dataset of n individuals. Each individual i is associated with a feature vector $\mathbf{x}_i \in \mathbb{R}^d$ and a binary sensitive attribute $\mathrm{s}_i \in \{0, 1\}$ (e.g., gender or race). The goal is to predict a binary target label $\mathrm{y}_i \in \{0, 1\}$. We operate in a semi-supervised setting, where the index set $\{1, \ldots, n\}$ is partitioned into a labeled set $\mathcal{L} = \{1, \ldots, \ell\}$ and an unlabeled set $\mathcal{U} = \{\ell + 1, \ldots, n\}$, with $n = \ell + u$. Both target labels y_i and sensitive attributes s_i are assumed to be observed only for individuals in the labeled set $\mathcal{L}$.

Graph Data We model the data as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where the node set $\mathcal{V}$ (with $|\mathcal{V}| = n$) represents the individuals, and the edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ represents relationships between them. The graph structure is encoded by an adjacency matrix $\mathbf{A} \in \{0, 1\}^{n \times n}$, where $\mathbf{A}_{ij} = 1$ if an edge exists between nodes i and j, and 0 otherwise. For standard graph-theoretic definitions, we refer the reader to [42]. Although node features $\mathbf{x_i}$ may be used to construct the graph, in this work we assume a given graph structure and a binary sensitive attribute for clarity. The proposed framework can, however, be extended to weighted graphs and non-binary sensitive attributes.

Objective Our objective is to learn a fair classifier that accurately predicts target labels Y while minimizing dependence on sensitive attributes S. We employ an encoder-classifier framework where:

- A GNN encoder $f_\theta : \mathbb{R}^{n\times d} \times \{0,1\}^{n\times n} \to \mathbb{R}^{n\times d'}$ maps the input features and graph structure to a set of node representations $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_n\} \in \mathbb{R}^{n\times d'}$, where $\mathbf{h}_i$ is the node representation for individual i.
- A classifier $g_\phi : \mathbb{R}^{n\times d'} \to (0,1)^{n\times 1}$ maps the node representations to a vector of predicted probabilities $\mathrm{P} = (\mathrm{p}_1, \ldots, \mathrm{p}_n)$, where $\mathrm{p}_i \in (0,1)$ is the predicted probability for individual i.
- Final predictions are obtained through thresholding: $\hat{\mathrm{Y}} = \mathbb{I}[\mathrm{P} > 0.5]$, where $\hat{\mathrm{y}}_i$ is the predicted label for individual i.

We seek representations that are predictive of target labels while invariant to sensitive attributes.

3.2 Graph Neural Networks

Graph Neural Networks (GNNs) learn node-level representations through message passing, where each node aggregates feature information from its neighbors. GNN architectures including GCN [43], GIN [44], GraphSAGE [45], and GAT [46] implement this approach. For semi-supervised label propagation, a two-layer GCN model is formulated as:

$$\mathbf{H} = \sigma\left(\hat{\mathbf{A}}\,\sigma\left(\hat{\mathbf{A}}\mathbf{X}\mathbf{W}^{(0)}\right)\mathbf{W}^{(1)}\right) \tag{1}$$

where $\mathbf{X}$ is a feature matrix, $\hat{\mathbf{A}} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$ is the normalized adjacency matrix, $\mathrm{D}_{ii} = \sum_j \mathrm{A}_{ij}$ is the degree matrix, $\sigma(\cdot)$ is the ReLU activation function, $\mathbf{W}^{(0)} \in \mathbb{R}^{d\times h}$ is the input-to-hidden weight matrix, and $\mathbf{W}^{(1)} \in \mathbb{R}^{h\times d'}$ is the hidden-to-hidden weight matrix.

For binary classification, the model outputs probabilities as:

$$\mathrm{P} = g_\phi\left(\mathbf{H}\right) \tag{2}$$

The base model is is trained by minimizing the cross-entropy loss over labeled nodes:

$$\mathcal{L}_{\text{pred}} = -\frac{1}{|\mathcal{L}|}\sum_{i\in\mathcal{L}}\left[\mathrm{y}_i \log(\mathrm{p}_i) + (1-\mathrm{y}_i)\log(1-\mathrm{p}_i)\right] \tag{3}$$

where $\mathrm{p}_i \in (0,1)$ is the predicted probability and $\mathrm{y}_i \in \{0,1\}$ is the corresponding class label.

This objective alone may lead to biased predictions.

3.3 Fairness in Graph Neural Networks

Evaluating fairness in Graph Neural Networks (GNNs) is a critical research concern that has been explored from multiple fundamental perspectives. Chen et al. [47] categorize fairness evaluation metrics for GNNs into four levels: *prediction-level*, *graph-level*, *neighborhood-level*, and *embedding-level* metrics. In this work, we focus on *prediction-level fairness*, which assesses how model outputs are influenced by sensitive attributes. Within this context, two key fairness notions are commonly considered: *statistical parity* and *equal opportunity*, defined as follows:

Statistical Parity (SP) requires predictions to be statistically independent of sensitive attributes, ensuring equal selection rates across groups:

$$\Delta_{\text{SP}} = |P(\hat{\mathrm{y}} = 1 \mid \mathrm{s} = 1) - P(\hat{\mathrm{y}} = 1 \mid \mathrm{s} = 0)| \tag{4}$$

Equal Opportunity (EO) requires equal true positive rates across sensitive groups, ensuring qualified individuals have equal chances of positive outcomes:

$$\Delta_{\mathrm{EO}} = |P(\hat{\mathrm{y}} = 1 \mid \mathrm{y} = 1, \mathrm{s} = 1) - P(\hat{\mathrm{y}} = 1 \mid \mathrm{y} = 1, \mathrm{s} = 0)| \tag{5}$$

Lower values of both Δ_{SP} and Δ_{EO} indicate better fairness. In the literature, the two metrics mentioned above are commonly used to assess the fairness of trained GNNs.

4 Fairness-Aware Regularization Framework

Our proposed method, Equal Opportunity Statistical Parity (EOSP), integrates fairness directly into GNN training through differentiable regularization terms. While fairness is typically evaluated on test data, we employ the labeled training set as a practical approximation to guide the model toward fair representations. In other words, our approach explicitly integrates fairness metrics into the end-to-end training of a given GNN, considering them as differentiable losses.

4.1 Statistical Parity Regularizer

Statistical Parity requires that predictions are independent of sensitive attributes. While the evaluation metric operates on discrete predictions:

$$\Delta_{SP} = |P(\hat{\mathrm{y}} = 1 \mid \mathrm{s} = 0) - P(\hat{\mathrm{y}} = 1 \mid \mathrm{s} = 1)| \tag{6}$$

we propose a differentiable approximation for training and used it as a loss function:

$$\mathcal{L}_{\mathrm{SP}} = \left| \frac{1}{|\mathcal{D}_1|} \sum_{i \in \mathcal{D}_1} \mathrm{p}_i - \frac{1}{|\mathcal{D}_0|} \sum_{j \in \mathcal{D}_0} \mathrm{p}_j \right| \tag{7}$$

where $\mathcal{D}_1 = \{i \in \mathcal{L} : \mathrm{s}_i = 1\}$ and $\mathcal{D}_0 = \{j \in \mathcal{L} : \mathrm{s}_j = 0\}$ represent the sensitive groups in the labeled training set.

This loss is differentiable because it uses continuous probabilities p_x instead of discrete predictions $\hat{\mathrm{y}}_x$, enabling gradient-based optimization.

4.2 Equal Opportunity Regularizer

Equal Opportunity requires equal true positive rates across sensitive groups. While the evaluation metric uses discrete predictions:

$$\Delta_{EO} = |P(\hat{\mathrm{y}} = 1 \mid \mathrm{y} = 1, \mathrm{s} = 0) - P(\hat{\mathrm{y}} = 1 \mid \mathrm{y} = 1, \mathrm{s} = 1)| \tag{8}$$

we define a differentiable approximation for training and used it as a loss function:

$$\mathcal{L}_{\mathrm{EO}} = \left| \frac{1}{|\mathcal{P}_1|} \sum_{i \in \mathcal{P}_1} \mathrm{p}_i - \frac{1}{|\mathcal{P}_0|} \sum_{j \in \mathcal{P}_0} \mathrm{p}_j \right| \tag{9}$$

where $\mathcal{P}_1 = \{i \in \mathcal{L} : \mathrm{y}_i = 1, \mathrm{s}_i = 1\}$ and $\mathcal{P}_0 = \{j \in \mathcal{L} : \mathrm{y}_j = 1, \mathrm{s}_j = 0\}$.

This formulation is differentiable as it operates on probability scores p_x rather than binary decisions, allowing seamless integration with gradient-based learning.

4.3 Final Objective Function

The Graph Neural Network (GNN) model introduced in the previous section optimizes predictive accuracy but does not account for its potential dependency on sensitive attributes s. To learn a model that is both accurate and fair, we propose regularizing the standard training objective with a composite fairness loss. Consequently, the model is trained to minimize the following global loss function:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{fairness}} \tag{10}$$

Here, $\mathcal{L}_{\text{pred}}$ is the standard cross-entropy loss (Eq. (3)) for the node classification task. The term $\mathcal{L}_{\text{fairness}}$ is our proposed fairness regularizer, defined as a weighted sum of two distinct fairness criteria:

$$\mathcal{L}_{\text{fairness}} = \alpha\, \mathcal{L}_{\text{eo}}(g_\phi(f_\theta(\mathbf{X}, \mathbf{A})), \mathrm{S}, \mathrm{Y}) + \beta\, \mathcal{L}_{\text{sp}}(g_\phi(f_\theta(\mathbf{X}, \mathbf{A})), \mathrm{S}) \tag{11}$$

In this formulation, $\mathcal{L}_{\text{EO}}$ enforces equal opportunity and $\mathcal{L}_{\text{SP}}$ enforces statistical parity. The hyperparameters α and β provide explicit control over the trade-off between these two fairness notions. Although these objectives are based on different statistical criteria, we reproduce in the Appendix the proof from [48] showing that they can be combined.

We employ Bayesian hyperparameter optimization [49] to tune α and β. A key advantage of this framework is its model-agnostic nature. For any generic GNN model f_θ, regardless of whether fairness is incorporated into its design, with a native loss function $\mathcal{L}_{\text{GNN}}$, it can be trained by minimizing:

$$\mathcal{L}_{\text{fair}} = \mathcal{L}_{\text{GNN}} + \mathcal{L}_{\text{fairness}} \tag{12}$$

This approach provides a flexible and direct method for integrating fairness constraints into existing GNN architectures and other fairness-aware techniques. The following section empirically demonstrates this integration in several standard GNN and fairness-aware models. To ensure reproducibility and facilitate further research, we provide our implementation at: https://github.com/mtavassoli/GNN-FC.

5 Experimental Setup

This section presents the comprehensive methodology used to evaluate the model-agnostic EOSP fairness constraint. We detail the datasets employed, the compared methods, training settings, evaluation metrics, and provide in-depth implementation details.

5.1 Datasets

We evaluated the proposed method on five publicly available real-world tabular datasets, commonly used in graph-based fairness research. A summary of each dataset is provided below:

1. **German [50]:** This dataset consists of 1000 individual clients from a German bank. The main objective is to evaluate the credit risk level for each individual, with a focus on the sensitive attribute "gender". In this dataset, the nodes represent individuals, and the links are based on significant similarities in their bank account details.
2. **Bail [51]:** This dataset contains records of 18,876 individuals granted bail by US state courts between 1990 and 2009. The primary goal is to classify whether a defendant is granted bail or not, focusing on the sensitive attribute "race". In this dataset, the nodes represent individuals, and the links are based on similarities in their demographics and prior criminal histories.

3. **Credit Defaulter [52]:** This dataset includes 30,000 individual credit card holders. The main objective is to predict whether an individual will default on credit card payments, with an emphasis on the sensitive attribute "age". In this dataset, the nodes represent individuals, and the links are based on high similarity in payment information.
4. **NBA [12]:** The dataset includes 400 NBA players with statistics from the 2016–2017 season, along with information such as nationality, age and salary. The players act as nodes in a graph, linked based on their Twitter connections. The main task is to predict whether a player's salary is above or below the league median, with nationality (U.S. or non-U.S.) used as a sensitive attribute.
5. **Pokec-n [53]:** This dataset is based on a popular online social network in Slovakia, similar to Facebook, and includes 66,569 individual users. The main objective is to predict a user's working field, with an emphasis on the sensitive attribute "region". In this dataset, the nodes represent users, and the links are based on friendship relations between them.

5.2 Compared Methods

We evaluate our model-agnostic EOSP method by integrating it with both standard GNN architectures (GCN [43], GraphSAGE [45]) and state-of-the-art fairness-aware methods (FairGNN [12], FairVGNN [10], CAF [17], FairGB [40], DAB-GCN [38]). For each model, we compare performance with and without EOSP to assess its impact on both accuracy and fairness.

5.3 Training Settings

For all methods, we tune the learning rate ($lr \in \{10^{-2}, 10^{-3}\}$), hidden size ($h \in \{16, 32\}$), dropout rate ($dr = 0.2$), and model depth (layers $\in \{1, 2, 3\}$). Hyperparameters for baseline methods follow their original publications. Train/validation/test splits follow [54], with labeled samples set to 100 (German), 100 (Bail), 6000 (Credit), 100 (NBA), and 500 (Pokec-n). Validation and test sets each comprise 25% of the data, allocated independently. All methods during training use labeled sensitive attributes equal to the training set size, except for CAF, which uses all available sensitive attributes.

Each experiment is run five times with different data splits and trained for 100 epochs. For the EOSP method, the hyperparameters α and β are tuned via Bayesian optimization using Optuna [49]. The search space for both α and β includes the values $\{0.01, 0.02, \ldots, 0.1, 0.2, 0.3, \ldots, 1, 2, 5\}$. Optimization is performed with 15 trials, and the best combination is chosen based on validation set performance. Hyperparameters of baseline methods are kept fixed to ensure a fair comparison.

All experiments run on nodes with dual Intel Xeon Gold 6136 processors (24 cores total, 3 GHz) and 192 GB RAM, using Python 3.9.18.

5.4 Evaluation Metrics

We evaluate both classification performance and prediction-level fairness. For classification, we use balanced accuracy (BACC), area under the ROC curve (AUC), and F1 score. For fairness, we employ statistical parity difference Δ_{SP} (Eq. (6)) and equal opportunity difference Δ_{EO} (Eq. (8)) metrics, where all metrics are expressed as percentages.

To select the best epoch during training, we compute the following hybrid score evaluated on the validation data:

$$\text{Score} = \text{BACC} + \frac{1}{2}\left[(100 - \Delta_{EO}) + (100 - \Delta_{SP})\right] \tag{13}$$

5.5 Implementation of the Model-Agnostic EOSP Fairness Constraint Using GNN-FC

The proposed EOSP fairness constraint is implemented as a modular component in the open-source `GNN-FC` library[2], which provides compatibility with standard GNN architectures (e.g., GCN) and fairness-aware frameworks (e.g., FairGNN). As outlined in Algorithm 1, the EOSP module integrates as an auxiliary objective during model training, ensuring full compatibility with PyTorch's automatic differentiation framework.

Algorithm 1: Integration of the EOSP Fairness Constraint

Require: class labels $Y_{\mathcal{L}}$, sensitive attribute labels $S_{\mathcal{L}}$, Hyperparameters (α,β)
Require:
1: **for** iteration $= 1$ to T **do**
2: Compute model loss $\mathcal{L}_{\text{GNN}}$
3: Compute fairness loss (EOSP) $\mathcal{L}_{\text{fairness}}$ (Eq. (11))
4: Compute total fair loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GNN}} + \mathcal{L}_{\text{fairness}}$ (Eq. (12))
5: Update the model by back-propagation
6: **end for**

6 Experimental Results

In this section, we aim to thoroughly evaluate our proposed model-agnostic EOSP framework. To this end, we address the following research questions:

1. **RQ1**: How do our proposed fairness losses affect the fairness and performance of existing GNNs?
2. **RQ2**: How sensitive is our method to its hyperparameters (α and β)?
3. **RQ3**: How does limited labeled sensitive attribute availability affect fairness-aware graph learning?
4. **RQ4**: What is the computational overhead of our fairness method?

6.1 RQ1: Performance Comparison

To evaluate the effectiveness of the proposed model-agnostic EOSP framework, we conducted a comparative analysis of classification performance and fairness outcomes across a range of models, both with and without the EOSP fairness constraint. Experiments were performed on the German, Bail, Credit Defaulter, NBA, and Pokec-n datasets. The results, including the means and standard deviations over five runs with different random splits of train/validation/test, are reported as percentages in Tables 1, 2, 3, 4, and 5.

On the German dataset (Table 1), applying EOSP to GCN, FairGNN, CAF, FairGB, and DAB-GCN yields consistent fairness improvements while maintaining comparable classification accuracy. This confirms that EOSP enhances fairness without compromising predictive utility. GCN-EOSP achieves the strongest fairness performance among all models, with $\Delta_{SP} = 1.98$

[2] https://github.com/mtavassoli/GNN-FC

Table 1: Evaluation of Node Classification and Fairness Performance on the German Dataset. The best result between methods with and without EOSP is bold.

Method	BACC(↑)	AUC(↑)	F1(↑)	$\Delta_{SP}(\downarrow)$	$\Delta_{EO}(\downarrow)$
GCN [43]	59.70(3.79)	**63.29(5.00)**	**69.25 (5.43)**	5.58(7.64)	5.36(6.10)
GCN-EOSP (Ours)	**60.02(5.25)**	63.21(5.32)	63.89(8.75)	**1.98(1.81)**	**1.86(1.04)**
SAGE [45]	60.34(4.71)	65.37(5.23)	53.82(23.20)	**6.98(6.58)**	11.23(6.52)
SAGE-EOSP (Ours)	**62.90(1.66)**	**67.68(3.12)**	**68.78(3.47)**	11.63(3.50)	**10.37(3.67)**
FairGNN [12]	61.01(4.56)	64.63(3.75)	67.18(3.05)	10.04(5.72)	8.52(6.51)
FairGNN-EOSP (Ours)	**61.35(4.03)**	**65.18(3.52)**	**68.79(3.95)**	**6.90(5.27)**	**4.94(2.31)**
FairVGNN [10]	**61.50(2.59)**	**65.51(2.83)**	**71.37(4.43)**	5.28(4.06)	**4.16(2.02)**
FairVGNN-EOSP (Ours)	60.19(4.79)	65.36(3.09)	68.00(7.85)	**4.04(1.29)**	4.46(3.46)
CAF [17]	61.68(4.02)	66.08(4.19)	**69.27(6.58)**	6.25(5.66)	9.42(5.83)
CAF-EOSP (Ours)	**62.13(5.55)**	**66.11(4.84)**	67.22(6.39)	**4.13(1.46)**	**4.19(2.33)**
Fair-GB [40]	**61.62(3.51)**	66.16(5.44)	65.77(7.95)	7.78(6.83)	8.57(7.31)
FairGB-EOSP (Ours)	61.10(3.52)	**66.44(4.69)**	**68.35(8.49)**	**7.24(4.73)**	**5.47(3.23)**
DAB-GCN[38]	**60.72(3.19)**	65.02(2.87)	67.01(3.69)	5.93(3.77)	5.77(4.31)
DAB-GCN-EOSP (Ours)	60.59(1.56)	**65.90(2.44)**	**67.01(2.40)**	**5.13(1.54)**	**5.62(2.73)**
Average Improvement	**+0.24**	**+0.55**	**+1.34**	**+0.83**	**+2.16**

and $\Delta_{EO} = 1.86$, representing substantial improvements over both vanilla GNNs and existing fairness-aware baselines. For classification performance, SAGE-EOSP attains the highest predictive scores, achieving a BACC of 62.90, AUC of 67.68, and F1-score of 68.78.

Table 2: Evaluation of Node Classification and Fairness Performance on the Bail Dataset. The best result between methods with and without EOSP is bold.

Method	BACC(↑)	AUC(↑)	F1(↑)	$\Delta_{SP}(\downarrow)$	$\Delta_{EO}(\downarrow)$
GCN [43]	91.07 (1.07)	93.94 (1.41)	89.21 (1.26)	7.18 (1.10)	1.57 (0.37)
GCN-EOSP (Ours)	**91.90 (0.98)**	**94.67 (1.30)**	**90.12 (1.20)**	**6.89 (0.82)**	**1.47 (0.80)**
SAGE [45]	**94.87 (0.93)**	**97.72 (0.50)**	**93.82 (0.94)**	7.49 (1.24)	1.38 (0.93)
SAGE-EOSP (Ours)	94.84 (1.00)	97.63 (0.51)	93.71 (1.39)	**6.91 (1.04)**	**1.33 (0.98)**
FairGNN [12]	90.37 (0.97)	**91.94 (1.48)**	86.15 (1.48)	6.89 (1.18)	2.82 (1.20)
FairGNN-EOSP (Ours)	**90.77 (0.54)**	91.32 (1.96)	**86.26 (2.11)**	**6.86 (0.74)**	**2.24 (1.18)**
FairVGNN [10]	**91.85 (1.04)**	**95.68 (1.11)**	89.65 (1.26)	7.10 (1.12)	**0.59 (0.20)**
FairVGNN-EOSP (Ours)	91.67 (0.60)	95.04 (0.96)	**89.76 (0.83)**	**6.16 (1.27)**	1.24 (0.81)
CAF [17]	94.43 (1.44)	97.16 (0.85)	93.08 (1.87)	6.88 (1.40)	1.20 (0.80)
CAF-EOSP (Ours)	**95.10 (0.99)**	**97.67 (0.63)**	**93.62 (1.30)**	**5.93 (1.40)**	**0.53 (0.27)**
FairGB [40]	93.06 (0.86)	96.35 (0.82)	91.75 (0.85)	6.68 (1.49)	1.35 (1.05)
FairGB-EOSP (Ours)	**94.50 (0.52)**	**97.46 (0.48)**	**93.18 (0.75)**	**5.82 (1.51)**	**1.05 (1.97)**
DAB-GCN [38]	71.55 (6.19)	**79.89 (5.04)**	63.18 (9.92)	11.67 (10.86)	10.40 (7.86)
DAB-GCN-EOSP (Ours)	**72.85 (4.99)**	79.85 (4.79)	**66.25 (5.76)**	**4.28 (4.38)**	**4.67 (4.92)**
Average Improvement	**+0.64**	**+0.14**	**+0.84**	**+1.56**	**+0.96**

On the Bail dataset (Table 2), EOSP consistently improves both fairness and predictive performance across GCN, SAGE, FairGNN, CAF, FairGB, and DAB-GCN. These results demonstrate the effectiveness of EOSP for both vanilla GNNs and existing fairness-aware baselines.

Among the models, CAF-EOSP achieves the strongest fairness gains, with $\Delta_{SP} = 5.93$ and $\Delta_{EO} = 0.53$, substantially reducing disparities compared to all other models. It also attains the best classification performance, achieving the highest balanced accuracy (95.10), AUC (97.67), and F1-score (93.62). These findings highlight EOSP's ability to achieve a superior balance between fairness and predictive accuracy.

Table 3: Evaluation of Node Classification and Fairness Performance on the Credit Dataset. The best result between methods with and without EOSP is bold.

Method	BACC(↑)	AUC(↑)	F1(↑)	Δ_{SP}(↓)	Δ_{EO}(↓)
GCN [43]	64.15 (0.37)	**69.16 (0.58)**	**81.34 (0.36)**	9.61 (2.07)	8.39 (2.26)
GCN-EOSP (Ours)	**64.29 (0.33)**	69.03 (0.40)	79.85 (0.70)	**1.54 (0.71)**	**1.48 (1.32)**
SAGE [45]	69.62 (0.29)	**75.57 (0.44)**	82.91 (2.99)	7.44 (3.04)	6.09 (3.00)
SAGE-EOSP (Ours)	**69.95 (0.56)**	75.42 (0.60)	**83.62 (1.02)**	**1.61 (1.06)**	**1.47 (0.78)**
FairGNN [12]	63.73 (0.48)	**70.27 (0.20)**	78.16 (0.54)	11.13 (3.58)	10.12 (3.99)
FairGNN-EOSP (Ours)	**64.37 (0.23)**	70.00 (0.29)	**79.46 (1.61)**	**1.20 (0.42)**	**0.71 (0.83)**
FairVGNN [10]	64.97 (1.12)	69.56 (0.92)	**85.53 (2.03)**	6.46 (2.93)	5.58 (1.83)
FairVGNN-EOSP (Ours)	**65.79 (0.28)**	**69.87 (0.38)**	79.05 (2.13)	**2.07 (0.95)**	**2.34 (1.37)**
CAF [17]	69.43 (1.14)	**75.37 (0.39)**	80.61 (4.74)	5.48 (3.22)	4.66 (2.22)
CAF-EOSP (Ours)	**69.57 (0.48)**	75.02 (0.34)	**80.91 (1.28)**	**2.31 (1.45)**	**1.93 (0.66)**
FairGB [40]	69.63 (0.44)	74.54 (0.13)	82.14 (2.95)	1.55 (1.05)	1.63 (0.85)
FairGB-EOSP (Ours)	**69.93 (0.25)**	**74.74 (0.40)**	**84.59 (0.62)**	**0.93 (0.49)**	**1.12 (0.75)**
DAB-GCN [38]	69.11 (1.08)	72.25 (0.62)	80.90 (2.57)	6.57 (3.17)	5.04 (3.27)
DAB-GCN-EOSP (Ours)	**69.58 (0.33)**	**73.80 (0.34)**	**81.95 (0.95)**	**1.76 (1.42)**	**2.03 (0.46)**
Average Improvement	**+0.40**	**+0.31**	-0.31	**+5.28**	**4.36**

On the Credit dataset (Table 3), applying EOSP improves fairness across all vanilla GNNs and existing fairness-aware baselines, reducing Δ_{SP} and Δ_{EO} by over 80% in some cases while maintaining comparable classification accuracy. This confirms that EOSP enables fairness improvements without compromising predictive utility. FairGB-EOSP achieves the strongest fairness performance among all models, with $\Delta_{SP} = 0.93$ and $\Delta_{EO} = 1.12$, representing substantial improvements over both vanilla GNNs and existing fairness-aware baselines. For classification performance, SAGE-EOSP and FairGB-EOSP attain the highest predictive scores, achieving BACC of 69.95 and 69.93 respectively with superior F1 scores.

On the NBA dataset (Table 4), EOSP enhances both fairness and predictive performance across GCN, SAGE, and DAB-GCN models. CAF-EOSP delivers the strongest fairness improvements, achieving $\Delta_{SP} = 4.30$ and $\Delta_{EO} = 7.47$, substantially reducing disparities compared to baseline methods. In terms of classification performance, FairGNN-EOSP achieves the highest AUC (81.79), while DAB-GCN-EOSP attains the best balanced accuracy (74.34) and F1-score (76.52). These results demonstrate EOSP's effectiveness in balancing fairness and predictive utility.

On the Pokec-n dataset (Table 5), applying EOSP to GCN, FairGNN, and FairVGNN yields consistent fairness improvements while maintaining or enhancing predictive accuracy. CAF-EOSP achieves the strongest statistical parity performance with $\Delta_{SP} = 1.19$, while CAF attains the best equal opportunity value of $\Delta_{EO} = 1.78$. For classification performance, FairGNN-EOSP reaches the highest balanced accuracy (68.45) and AUC (73.19). These results demonstrate that EOSP generally provides balanced enhancements across both fairness and utility objectives on this dataset.

The key comparisons evaluate each baseline model against its EOSP-enhanced version across

Table 4: Evaluation of Node Classification and Fairness Performance on the NBA Dataset. The best result between methods with and without EOSP is bold.

Method	BACC(↑)	AUC(↑)	F1(↑)	$\Delta_{SP}(\downarrow)$	$\Delta_{EO}(\downarrow)$
GCN [43]	65.51 (2.64)	**74.36 (2.31)**	66.83 (3.38)	**7.22 (4.67)**	15.27 (9.22)
GCN-EOSP (Ours)	**67.85 (3.83)**	74.19 (2.75)	**67.72 (2.25)**	8.47 (4.64)	**14.89 (3.21)**
SAGE [45]	62.03 (3.27)	69.90 (5.92)	59.27 (14.74)	9.62 (6.90)	17.46 (16.61)
SAGE-EOSP (Ours)	**69.11 (2.20)**	**77.41 (2.05)**	**69.34 (3.31)**	**8.16 (3.92)**	**9.98 (9.20)**
FairGNN [12]	68.25 (9.03)	79.62 (4.28)	66.29 (6.46)	7.74 (7.44)	16.40 (14.71)
FairGNN-EOSP (Ours)	**70.79 (3.97)**	**81.79 (2.48)**	**66.87 (10.61)**	**6.97 (4.04)**	**10.61 (10.13)**
FairVGNN [10]	65.81 (2.48)	73.08 (2.41)	65.10 (6.98)	**6.71 (4.48)**	16.15 (6.88)
FairVGNN-EOSP (Ours)	**70.09 (2.18)**	**74.76 (3.36)**	**71.07 (1.70)**	9.30 (4.32)	**11.06 (6.13)**
CAF [17]	66.68 (4.58)	71.60 (4.85)	**70.74 (2.59)**	8.60 (4.32)	**7.45 (6.09)**
CAF-EOSP (Ours)	**66.69 (6.20)**	**73.56 (3.33)**	70.38 (3.77)	**4.30 (4.86)**	7.47 (8.56)
FairGB [40]	**68.22 (3.23)**	**74.00 (2.55)**	**71.64 (2.66)**	13.85 (9.15)	**10.34 (8.55)**
FairGB-EOSP (Ours)	67.53 (2.35)	72.05 (2.69)	69.30 (2.75)	**11.67 (5.64)**	13.34 (8.47)
DAB-GCN [38]	74.07 (3.80)	80.51 (4.27)	76.51 (2.17)	12.30 (6.13)	12.03 (7.00)
DAB-GCN-EOSP (Ours)	**74.34 (4.96)**	**80.99 (2.65)**	**76.52 (3.58)**	**10.37 (4.75)**	**11.10 (8.15)**
Average Improvement	**+2.27**	**+2.24**	**+1.99**	**+0.97**	**+2.43**

Table 5: Evaluation of Node Classification and Fairness Performance on the Pokec-n Dataset. The best result between methods with and without EOSP is bold.

Method	BACC(↑)	AUC(↑)	F1(↑)	$\Delta_{SP}(\downarrow)$	$\Delta_{EO}(\downarrow)$
GCN [43]	64.65 (1.24)	69.21 (1.40)	61.46 (3.26)	1.85 (1.01)	2.37 (1.58)
GCN-EOSP (Ours)	**66.30 (0.72)**	**70.26 (0.70)**	**62.76 (1.03)**	**1.45 (0.75)**	**2.15 (1.50)**
SAGE [45]	64.36 (2.44)	69.44 (2.31)	**60.50** (4.27)	**1.56 (0.96)**	**2.24 (1.95)**
SAGE-EOSP (Ours)	**67.04 (0.81)**	**71.86 (0.69)**	60.47 (3.92)	1.55 (1.30)	3.03 (1.90)
FairGNN [12]	68.34 (0.96)	72.84 (0.98)	63.98 (1.12)	4.18 (3.56)	5.10 (4.51)
FairGNN-EOSP (Ours)	**68.45 (1.31)**	**73.19 (0.90)**	**64.87 (1.81)**	**2.90 (1.17)**	**3.24 (2.40)**
FairVGNN	66.38 (1.21)	**70.17 (0.45)**	58.91 (3.91)	4.23 (1.42)	5.44 (2.51)
FairVGNN-EOSP (Ours)	66.09 (0.78)	69.71 (0.75)	**59.89 (1.54)**	**3.03 (1.39)**	**4.08 (2.04)**
CAF [17]	62.43 (1.65)	65.76 (1.80)	**61.46 (3.22)**	1.64 (0.53)	**1.78 (1.81)**
CAF-EOSP (Ours)	**63.56 (2.32)**	**67.91 (2.19)**	61.38 (2.61)	**1.19 (0.41)**	2.65 (1.14)
FairGB [40]	65.08 (2.45)	70.11 (1.60)	**61.15 (5.91)**	2.13 (0.60)	**2.76 (1.01)**
FairGB-EOSP (Ours)	**67.14 (1.00)**	**71.05 (1.49)**	60.65 (2.76)	**1.40 (0.66)**	4.15 (1.17)
DAB-GCN [38]	62.03 (1.91)	66.85 (1.96)	**60.64 (3.60)**	**3.54 (2.73)**	4.15 (2.22)
DAB-GCN-EOSP (Ours)	**62.47 (1.62)**	**67.09 (1.93)**	59.97 (3.24)	3.68 (2.59)	**4.02 (2.71)**
Average Improvement	**+1.10**	**+0.81**	**+0.36**	**+0.56**	**+0.08**

five datasets. Results demonstrate that EOSP generally improves fairness metrics (Δ_{SP} and Δ_{EO}) while maintaining classification performance (BACC, AUC, F1) for most models. However, in a few cases, EOSP does not uniformly improve fairness metrics; for example, with SAGE-EOSP on the German dataset, Δ_{SP} increased from 6.98 to 11.63 compared to the baseline, while Δ_{EO} improved modestly from 11.23 to 10.37, alongside increased predictive performance. Although it performs better than the baseline, this raises the question of whether EOSP may not be equally effective for all models or if other factors contribute to its suboptimal performance in this case.

Due to computational constraints, we employ Bayesian hyperparameter optimization (15 tri-

als vs. 400 exhaustive) with hyperparameters α and β selected using the validation score defined in Eq. (13). Figure 1 shows the convergence trajectory for SAGE-EOSP on the German dataset using a single train/validation/test split, revealing an initial plateau (trials 15-20) followed by optimal discovery at trial 25 and subsequent stability. This demonstrates efficient near-optimal solution finding (16× faster than exhaustive search) and shows that additional trials enable overcoming initial limitations.

Our analysis reveals that extended optimization enables SAGE-EOSP on the German dataset to achieve superior fairness-accuracy-stability trade-offs. While the initial 15-trial configuration exhibited increased Δ_{SP}, optimization over 25 trials yields balanced improvements: Δ_{SP} improves from 11.63 to 8.37 (baseline: 6.93) and Δ_{EO} from 11.23 to 8.84, while maintaining competitive accuracy (BACC: 61.09 vs. 60.34 baseline) and dramatically enhancing stability (F1 std: 23.20→3.55). Figure 2 confirms consistent improvements across trials, demonstrating EOSP's robustness.

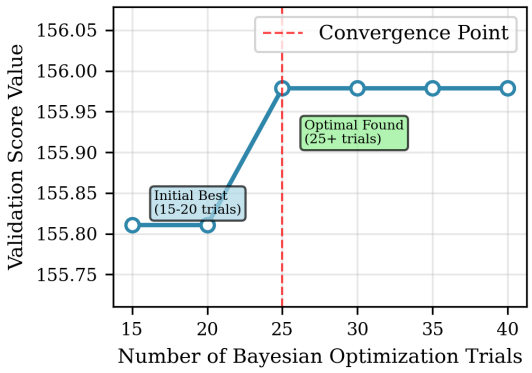


Figure 1: Bayesian Optimization Convergence

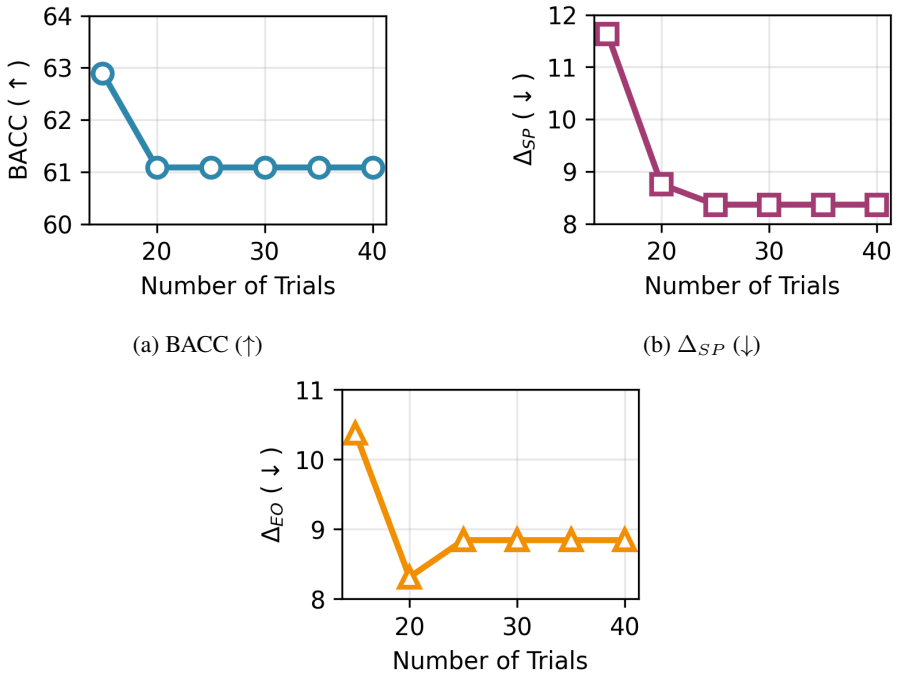
6.2 RQ2: Hyper-Parameter Analysis

Our proposed EOSP framework introduces two key hyperparameters: α, which controls the influence of the equal opportunity loss, and β, which controls the statistical parity loss contribution. **While both relate to fairness objectives, they impact model accuracy differently, creating a complex trade-off space that requires careful balancing.**

For optimal model configuration in our main experiments, we utilized Bayesian hyperparameter optimization to efficiently explore a discretized two-dimensional parameter space (defined as a rectangular grid) and to identify configurations that effectively balance the competing objectives.

However, to systematically characterize the sensitivity and interaction effects of these parameters **on both fairness and accuracy**, we conducted a comprehensive grid search with both hyperparameters varying across $\{0.02, 0.03, 0.04, 0.05, 0.5, 0.6, \ldots, 1.0\}$, covering both fine-grained low values and broader high values to capture their effects.

Figure 3 presents the hyperparameter ablation study on the German dataset, illustrating how variations in α and β affect both predictive performance (BACC, F1, AUC) and fairness metrics (Δ_{EO}, Δ_{SP}). The grid analysis reveals several important patterns: the configuration $\alpha = 0.02$, $\beta = 0.8$ achieves the highest balanced accuracy (67.43) with strong fairness properties ($\Delta_{SP} =$



(a) BACC (↑)

(b) Δ_{SP} (↓)

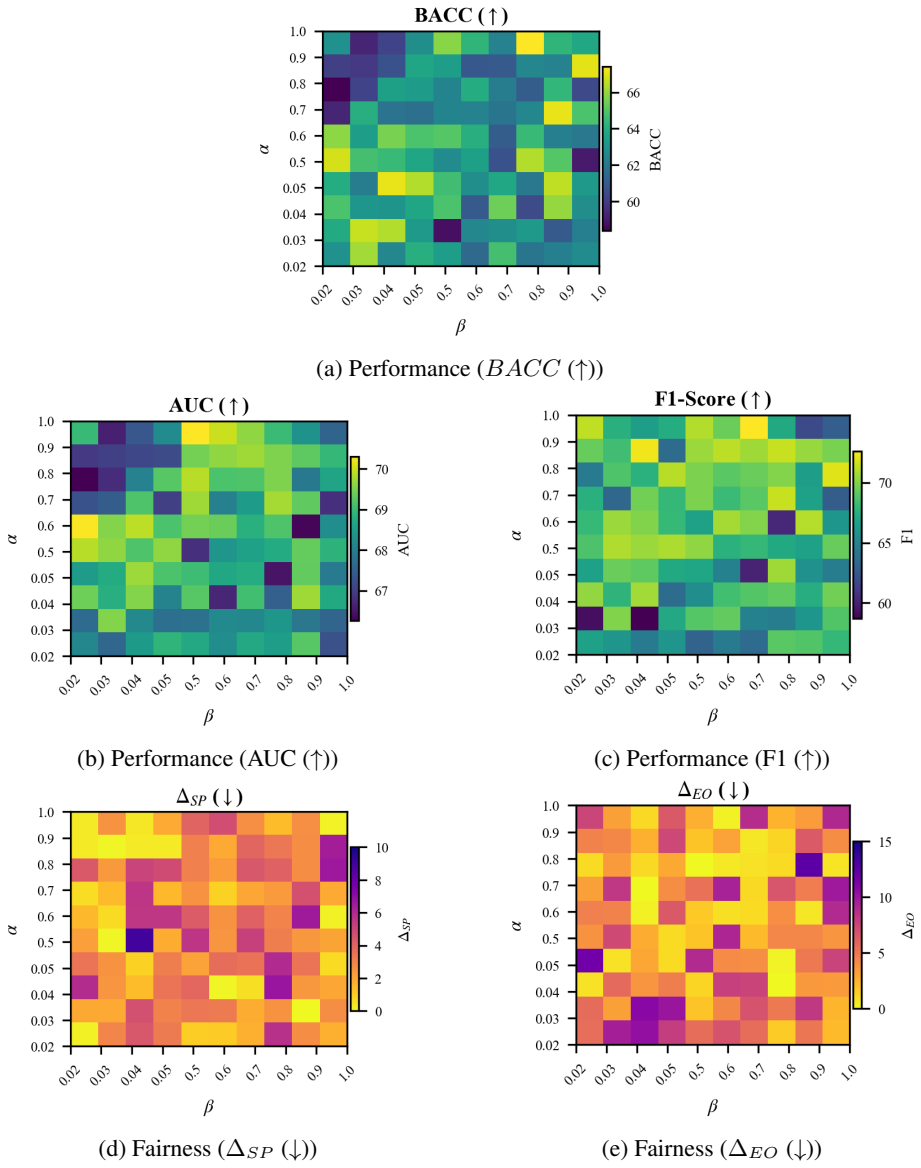
(c) Δ_{EO} (↓)

Figure 2: Bayesian optimization for SAGE-EOSP on the German dataset: (a) Balanced accuracy decreases slightly but stabilizes after 25 trials; (b) Statistical parity (Δ_{SP}) shows significant improvement, decreasing from 11.63 to 8.37; (c) Equal opportunity (Δ_{EO}) converges to stable values. All metrics reach convergence within 25 trials.

1.41, $\Delta_{EO} = 2.51$). However, the optimal fairness metrics occur at distinct operating points: perfect statistical parity ($\Delta_{SP} = 0.00$) is achieved at $\alpha = 0.9$, $\beta = 0.9$ (BACC = 60.95, AUC = 67.86, F1 = 66.00, $\Delta_{EO} = 8.17$), while optimal equal opportunity ($\Delta_{EO} = 0.03$) occurs at $\alpha = 0.04$, $\beta = 0.5$ (BACC = 62.48, AUC = 69.87, F1 = 69.87, $\Delta_{SP} = 3.12$). This demonstrates that no single configuration simultaneously optimizes all objectives, highlighting the challenge in identifying optimal hyperparameters for balancing the accuracy-fairness trade-off.

This complexity motivates our two-stage approach. First, we define a comprehensive validation metric to jointly evaluate accuracy–fairness trade-offs. This score is given in Eq. (13). Second, we employ Bayesian hyperparameter optimization to efficiently explore the high-dimensional parameter space and identify configurations that effectively balance these competing objectives. This approach is essential, as performing an exhaustive search across the entire hyperparameter space is computationally infeasible.

Figure 4 presents the results of our Bayesian hyperparameter optimization of the hybrid score in Eq. (13), showing the relationship between fairness metrics (Δ_{EO}, Δ_{SP}) and balanced accuracy across different α and β configurations. The analysis reveals that Bayesian optimization successfully navigated the complex fairness-accuracy trade-off space, identifying an optimal configuration ($\alpha = 0.08$, $\beta = 0.01$) where minimal fairness regularization achieves the best



(a) Performance ($BACC$ (↑))

(b) Performance (AUC (↑))

(c) Performance (F1 (↑))

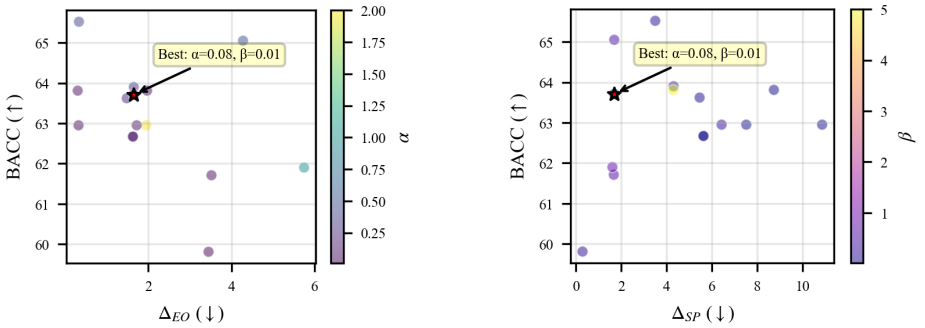
(d) Fairness (Δ_{SP} (↓))

(e) Fairness (Δ_{EO} (↓))

Figure 3: Hyper-parameter analysis on the German dataset, evaluating the effect of varying α and β on predictive performance and fairness metrics.

balance. This suggests that gentle fairness constraints are sufficient to guide the model toward equitable solutions without compromising predictive performance.

The success of Bayesian optimization in this context underscores its value for fairness-aware hyperparameter tuning, efficiently identifying subtle configurations that would be challenging to discover through exhaustive grid search or manual experimentation.



(a) Effect of (α) (b) Effect of (β)

Figure 4: Hyperparameter analysis on the German dataset using Bayesian optimization, evaluating the effects of the fairness coefficients α (equal opportunity) and β (statistical parity) on the accuracy-fairness trade-off.

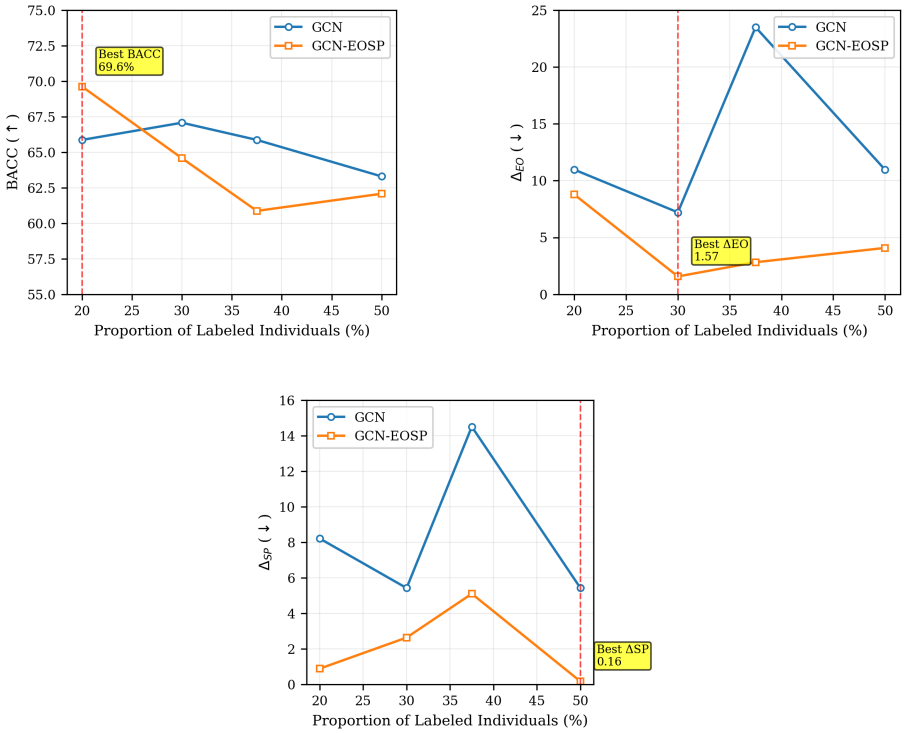


Figure 5: Impact of labeled sensitive attribute proportion on GCN vs GCN-EOSP performance across three metrics.

6.3 RQ3: Impact of Labeled Sensitive Attribute Availability

To answer RQ3, we investigate how the number of labeled sensitive attributes affects our proposed EOSP framework. We evaluate GCN and GCN-EOSP on the NBA dataset, varying the proportion of labeled sensitive attributes across $\{20\%, 30\%, 37.5\%, 50\%\}$ while maintaining an equivalent number of labeled individuals. Recall that for each labeled sample, both the class label and the corresponding sensitive attribute are available.

Figure 5 reveals that GCN-EOSP consistently outperforms vanilla GCN on fairness metrics. Notably, EOSP achieves **near-perfect statistical parity** ($\Delta_{SP} = 0.16$) at 50% labeled attributes, representing a 97% improvement over GCN, while maintaining competitive accuracy. The optimal configuration varies by metric: balanced accuracy peaks at 20% labels (BACC = 69.62), statistical parity optimizes at 50% labels ($\Delta_{SP} = 0.16$), and equal opportunity achieves its best trade-off at 30% labels ($\Delta_{EO} = 1.57$).

These results demonstrate that while both models benefit from increased labeled data, EOSP provides **substantial fairness improvements** at every data scale. Even with minimal labeled sensitive attributes (20% of individuals), EOSP improves balanced accuracy by 5.7% while significantly reducing fairness violations. This confirms EOSP's effectiveness across diverse real-world scenarios where labeled sensitive attribute availability may be limited.

6.4 RQ4: Time Complexity Analysis

To assess the additional training time complexity introduced by the proposed model-agnostic EOSP method compared to baseline methods, we report the training times for all methods, with and without EOSP, on the German dataset in Table 6. According to the results in Table 6, the additional training time introduced by EOSP is less than one second, making the overhead negligible.

Table 6: Evaluation of Training Time Performance on the German Dataset.

Method	**Time (Seconds)**	**Method-Fairness**	**Time (Seconds)**
GCN	1.93	GCN-EOSP	2.52
SAGE	1.94	SAGE-EOSP	2.56
FairGNN	1.39	FairGNN-EOSP	1.99
FairVGNN	47.01	FairVGNN-EOSP	49.93
CAF	10.30	CAF-EOSP	11.75
FairGB	2.21	FairGB-EOSP	3.07
DAB-GCN	5.32	DAB-GCN-EOSP	6.24

7 Conclusion

Fairness is a central principle in the European guidelines for high-risk AI systems, which emphasize the need for effective methods to avoid bias. However, the accuracy of algorithms is still crucial, and many fairness techniques lead to an undesirable loss of accuracy. In this paper, we present a novel, model-agnostic approach to improve fairness through equal opportunity and statistical parity regularizations. Our approach is applicable to any node classification model. It effectively mitigates bias while providing higher accuracy than conventional fairness methods. Moreover, when applied to a given model the siza of that model remains the same since no additional design is added.

A key advantage of our method is its practical feasibility: it requires only access to sensitive attributes and class labels of labeled nodes during training. This is particularly valuable for real-world applications, where full coverage of sensitive attribute labels is often not feasible due to data collection issues and privacy concerns. Our approach fits well with these constraints and provides a realistic solution to mitigate bias. Moreover, our method is easy to configure, with only two adjustable parameters to balance the trade-off between accuracy and fairness, making it both practical and accessible for various applications.

References

[1] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Computing Surveys*, 56(7):1–38, 2024.

[2] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[3] Mary Madden, Amanda Lenhart, Sandra Cortesi, Urs Gasser, Maeve Duggan, Aaron Smith, and Meredith Beaton. Teens, social media, and privacy. *Pew Research Center*, 21(1055):2–86, 2013.

[4] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80, 2008.

[5] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In *The world wide web conference*, pages 417–426, 2019.

[6] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. Graph neural networks in recommender systems: a survey. *ACM Computing Surveys*, 55(5):1–37, 2022.

[7] Luca Oneto and Silvia Chiappa. Fairness in machine learning. In *Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019)*, pages 155–196. Springer, 2020.

[8] Eustasio Del Barrio, Paula Gordaliza, and Jean-Michel Loubes. Review of mathematical frameworks for fairness in machine learning. *arXiv preprint arXiv:2005.13755*, 2020.

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[10] Yu Wang, Yuying Zhao, Yushun Dong, Huiyuan Chen, Jundong Li, and Tyler Derr. Improving fairness in graph neural networks via mitigating sensitive attribute leakage. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1938–1948, 2022.

[11] Maarten Buyl and Tijl De Bie. Debayes: a bayesian method for debiasing network embeddings. In *International Conference on Machine Learning*, pages 1220–1229. PMLR, 2020.

[12] Enyan Dai and Suhang Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 680–688, 2021.

[13] Yushun Dong, Ninghao Liu, Brian Jalaian, and Jundong Li. Edits: Modeling and mitigating data bias for graph neural networks. In *Proceedings of the ACM Web Conference 2022*, pages 1259–1269, 2022.

[14] Tahleen Rahman, Bartlomiej Surma, Michael Backes, and Yang Zhang. Fairwalk: Towards fair graph embedding. 2019.

[15] Chirag Agarwal, Himabindu Lakkaraju, and Marinka Zitnik. Towards a unified framework for fair and stable graph representation learning. In *Uncertainty in Artificial Intelligence*, pages 2114–2124. PMLR, 2021.

[16] Jing Ma, Ruocheng Guo, Mengting Wan, Longqi Yang, Aidong Zhang, and Jundong Li. Learning fair node representations with graph counterfactual fairness. February 2022.

[17] Zhimeng Guo, Jialiang Li, Teng Xiao, Yao Ma, and Suhang Wang. Towards fair graph neural networks via graph counterfactual. October 2023.

[18] YanMing Hu, TianChi Liao, JiaLong Chen, Jing Bian, ZiBin Zheng, and Chuan Chen. Migrate demographic group for fair graph neural networks. *Neural Networks*, 175:106264, 2024.

[19] Guixian Zhang, Debo Cheng, Guan Yuan, and Shichao Zhang. Learning fair representations via rebalancing graph structure. *Information Processing & Management*, 61(1):103570, 2024.

[20] April Chen, Ryan A Rossi, Namyong Park, Puja Trivedi, Yu Wang, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–23, 2024.

[21] Indro Spinelli, Simone Scardapane, Amir Hussain, and Aurelio Uncini. Fairdrop: Biased edge dropout for enhancing fairness in graph representation learning. *IEEE Transactions on Artificial Intelligence*, 3(3):344–354, 2021.

[22] Zemin Liu, Trung-Kien Nguyen, and Yuan Fang. On generalized degree fairness in graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4525–4533, 2023.

[23] O Deniz Kose, Yanning Shen, and Gonzalo Mateos. Fairness-aware graph filter design. In *2023 57th Asilomar Conference on Signals, Systems, and Computers*, pages 330–334. IEEE, 2023.

[24] Jian Kang, Jingrui He, Ross Maciejewski, and Hanghang Tong. Inform: Individual fairness on graph mining. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 379–389, 2020.

[25] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[26] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, pages 214–226, 2012.

[27] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *International Conference on Artificial Intelligence and Statistics*, pages 962–970. PMLR, 2017.

[28] Laurent Risser, Alberto González Sanz, Quentin Vincenot, and Jean-Michel Loubes. Tackling algorithmic bias in neural-network classifiers using wasserstein-2 regularization. *J. Math. Imaging Vis.*, 64(6):672–689, 2022.

[29] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

[30] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International conference on machine learning*, pages 2357–2365. PMLR, 2019.

[31] Suresh Venkatasubramanian, Carlos Scheidegger, Sorelle Friedler, and Aaron Clauset. Fairness in networks: social capital, information access, and interventions. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4078–4079, 2021.

[32] Jian Kang and Hanghang Tong. Fair graph mining. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4849–4852, 2021.

[33] Jian Kang and Hanghang Tong. Algorithmic fairness on graphs: Methods and trends. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4798–4799, 2022.

[34] Ahmad Khajehnejad, Moein Khajehnejad, Mahmoudreza Babaei, Krishna P Gummadi, Adrian Weller, and Baharan Mirzasoleiman. Crosswalk: Fairness-enhanced node representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11963–11970, 2022.

[35] Charlotte Laclau, Ievgen Redko, Manvi Choudhary, and Christine Largeron. All of the fairness for edge prediction with optimal transport. In *International Conference on Artificial Intelligence and Statistics*, pages 1774–1782. PMLR, 2021.

[36] Yushun Dong, Song Wang, Jing Ma, Ninghao Liu, and Jundong Li. Interpreting unfairness in graph neural networks via training node attribution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7441–7449, 2023.

[37] Yuchang Zhu, Jintang Li, Zibin Zheng, and Liang Chen. Fair graph representation learning via sensitive attribute disentanglement. In *Proceedings of the ACM Web Conference 2024*, pages 1182–1192, 2024.

[38] Yeon-Chang Lee, Hojung Shin, and Sang-Wook Kim. Disentangling, amplifying, and debiasing: Learning disentangled representations for fair graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 12013–12021, 2025.

[39] Chuxun Liu, Debo Cheng, Qingfeng Chen, Jiuyong Li, Lin Liu, and Rongyao Hu. Rethinking fair graph representation learning via structural rebalancing. *Available at SSRN 5320563*.

[40] Zhixun Li, Yushun Dong, Qiang Liu, and Jeffrey Xu Yu. Rethinking fair graph neural networks from re-balancing. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 1736–1745, 2024.

[41] Chengyu Li, Debo Cheng, Guixian Zhang, and Shichao Zhang. Contrastive learning for fair graph representations via counterfactual graph augmentation. *Knowledge-Based Systems*, 305:112635, 2024.

[42] John Adrian Bondy and Uppaluri Siva Ramachandra Murty. *Graph theory*. Springer Publishing Company, Incorporated, 2008.

[43] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[44] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks?, 2019.

[45] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. *Advances in neural information processing systems*, 30, 2017.

[46] Zhiyuan Liu and Jie Zhou. Graph attention networks. In *Introduction to graph neural networks*, pages 39–41. Springer, 2022.

[47] April Chen, Ryan A Rossi, Namyong Park, Puja Trivedi, Yu Wang, Tong Yu, Sungchul Kim, Franck Dernoncourt, and Nesreen K Ahmed. Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, 2023.

[48] MaryBeth Defrance and Tijl De Bie. Maximal combinations of fairness definitions. *Journal of Artificial Intelligence Research*, 82:1495–1579, 2025.

[49] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631, 2019.

[50] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[51] Kareem L Jordan and Tina L Freiburger. The effect of race/ethnicity on sentencing: Examining sentence type, jail length, and prison length. *Journal of Ethnicity in Criminal Justice*, 13(3):179–196, 2015.

[52] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, 36(2):2473–2480, 2009.

[53] Lubos Takac and Michal Zabovsky. Data analysis in public social networks. In *International scientific conference and international workshop present day trends of innovations*, volume 1, 2012.

[54] Yushun Dong, Jing Ma, Song Wang, Chen Chen, and Jundong Li. Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 35(10):10583–10602, 2023.

Appendix

Proof of Combination of Equal Opportunity and Statistical Parity

In this section, we reproduce the proof from [48] demonstrating the simultaneous satisfaction of both Equal Opportunity (EO) and Statistical Parity (SP). For completeness, we formalize these fairness definitions using the confusion matrix variables (Table 7) for two sensitive attribute groups, a and b.

Table 7: Symbolic representation of a confusion matrix.

	Predicted	
Actual	**Positive**	**Negative**
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

We normalize the confusion matrices such that the four entries in each matrix sum to 1, which simplifies subsequent calculations. The confusion matrix values for each sensitive attribute group are constrained by the underlying dataset characteristics:

$$
\begin{cases}
TP_a + FN_a = 1 - x, \\
TN_a + FP_a = x, \\
TP_b + FN_b = 1 - y, \\
TN_b + FP_b = y, \\
0 \leq TP_a \leq 1 - x, \quad 0 \leq TP_b \leq 1 - y, \\
0 \leq TN_a \leq x, \quad 0 \leq TN_b \leq y, \\
0 \leq FP_a \leq x, \quad 0 \leq FP_b \leq y, \\
0 \leq FN_a \leq 1 - x, \quad 0 \leq FN_b \leq 1 - y.
\end{cases} \tag{14}
$$

In Equation 14, the quantities $1-x$ and $1-y$ correspond to the base rates of groups a and b, respectively, where the base rate is defined as the proportion of positive individuals within each group's population.

To analyze the compatibility of Equal Opportunity and Statistical Parity, we first define what it means for a combination of fairness definitions to be possible, following [48].

Definition 1 (Possibility of Fairness Combinations)**.** *A combination of two fairness definitions is deemed **possible** if, after combining their constraints, for all pairs of base rates* $(1-x, 1-y)$ *belonging to a subset of* $[0,1]^2$ *with non-zero Lebesgue measure, all elements of both confusion matrices can still take values within subsets of* $[0,1]$ *that also have non-zero Lebesgue measure.*

We now express Equal Opportunity (EO) and Statistical Parity (SP) using the confusion matrix variables from Table 7.

Equal Opportunity (EO) (Equation 8) requires that the True Positive Rate (TPR) is equal across groups:

$$
\mathrm{TPR}_a = \mathrm{TPR}_b \quad \Longleftrightarrow \quad \frac{TP_a}{TP_a + FN_a} = \frac{TP_b}{TP_b + FN_b}. \tag{15}
$$

Using the base rate constraints from Equation 14, where $TP_a + FN_a = 1 - x$ and $TP_b + FN_b = 1 - y$, the EO condition simplifies to:

$$
\frac{TP_a}{TP_a + FN_a} = \frac{TP_b}{TP_b + FN_b} \quad \Longleftrightarrow \quad \frac{TP_a}{1-x} = \frac{TP_b}{1-y} \tag{16}
$$

Statistical Parity (SP) (Equation 6) requires that the probability of a positive prediction is equal across groups:

$$
\frac{TP_a + FP_a}{TP_a + FP_a + TN_a + FN_a} = \frac{TP_b + FP_b}{TP_b + FP_b + TN_b + FN_b} \tag{17}
$$

Since the confusion matrices are normalized such that $TP_g + FP_g + TN_g + FN_g = 1$ for $g \in \{a, b\}$, this reduces to:

$$
TP_a + FP_a = TP_b + FP_b \tag{18}
$$

Combining Equal Opportunity (Equation 16) and Statistical Parity (Equation 18) yields:

$$\begin{cases} FN_a = 1 - x - \dfrac{1-x}{1-y} TP_b, \\ TN_a = x - \dfrac{x-y}{1-y} TP_b - FP_b, \\ FN_b = 1 - y - TP_b, \\ TN_b = y - FP_b, \\ TP_a = \dfrac{1-x}{1-y} TP_b, \\ FP_a = \dfrac{x-y}{1-y} TP_b + FP_b. \end{cases} \tag{19}$$

Combining this system with the inequalities from Equation 14 allows us to compute feasible ranges for the free variables and the base rates:

$$\begin{cases} \dfrac{1}{1-y} TP_b \leq 1, \\ \dfrac{1-x}{1-y} TP_b \geq 0, \\ FP_b \leq x + \dfrac{-x+y}{1-y} TP_b, \\ \dfrac{-x+y}{1-y} TP_b \leq FP_b, \\ TP_b \leq 1 - y, \\ TP_b \geq 0, \\ FP_b \leq y, \\ 0 \leq \dfrac{1-x}{1-y}, \\ \dfrac{1}{1-y} TP_b \leq 1, \\ \dfrac{y-x}{1-y} \leq \dfrac{FP_b}{TP_b}, \\ FP_b + \dfrac{x-y}{1-y} TP_b \leq x \end{cases} \iff \begin{cases} \dfrac{y-x}{1-y} \leq \dfrac{FP_b}{TP_b}, \\ FP_b + \dfrac{x-y}{1-y} TP_b \leq x. \end{cases} \tag{20}$$

This system of inequalities shows a specific feasible range for the free variables FP_b and TP_b, while the base rates $1 - x$ and $1 - y$ remain unconstrained. Therefore, the Lebesgue measure is non-zero for each free variable over a set of base rates with a non-zero two-dimensional Lebesgue measure. Consequently, the combination of equal opportunity and statistical parity is possible.