FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 14, July, 2026, pp 258 – 269
DOI: https://doi.org/10.33003/fjs-2026-1014-5704

Trustworthy Spatio-Temporal Graph Learning for Urban Event Forecasting: A Systematic Survey of Explainability, Fairness, and Their Intersection

***Joy Onyeje Amafonye, Aliyu Suleiman Muhammed, Yusuf Salisu Ibrahim and Austin Olom Ogar**

Department of Computer Science, Faculty of Computing, Nile University of Nigeria, FCT Abuja, Nigeria.

*Corresponding authors' email: joy.amafonye@nileuniversity.edu.ng

Received: 08 Aug. 2026
Accepted: 11 Aug. 2026
Published: 22 Aug. 2026

ABSTRACT

Spatio-temporal graph neural networks (STGNNs) are widely used for urban event forecasting, including traffic prediction, crime hotspot estimation, air-quality monitoring, and epidemic surveillance. However, their adoption in public governance requires both explainability and fairness, in addition to predictive accuracy. Although existing surveys have separately examined STGNN architectures, graph explainability, and graph fairness, their intersection remains largely unexplored. This gap is significant because relational structures can propagate demographic and structural biases, while access to reliable explanations may vary across sensitive groups. Following a systematic review protocol applied to six digital libraries covering publications from 2016 to 2026, and screening 1,168 unique records down to a core corpus of 94 studies, this study organizes STGNN architectures for urban forecasting, develops a taxonomy of intrinsic and post-hoc explainability methods, and categorizes fairness notions, bias sources, and debiasing strategies for dynamic urban graphs. It also examines joint explainability–fairness approaches, including temporal fairness stability and explanation disparity, and consolidates relevant datasets and evaluation protocols. Of the 53 trust-aware studies in the corpus, only five—fewer than one in ten—address both properties on spatio-temporal graphs, and none before 2024 optimized them jointly. We conclude with a research agenda addressing temporal explanation benchmarks, dynamic fairness, causal grounding, scalability, and data-sparse Global-South cities.

Keywords: Spatio-Temporal Graph Neural Networks; Explainable AI; Algorithmic Fairness; Urban Computing; Trustworthy AI; Crime Forecasting; Traffic Forecasting

INTRODUCTION

Cities have become the principal theatre of machine-learning deployment in the public sector. Municipal governments forecast traffic to time signals and dispatch services, estimate crime risk to allocate prevention resources, predict air quality to issue health advisories, and anticipate demand across energy, transit, and emergency-response systems. The data underlying these tasks is spatio-temporal and relational: observations are indexed by location and time, and locations interact through road networks, adjacency, mobility flows, and socio-economic ties. Spatio-temporal graph neural networks (STGNNs)—architectures that couple graph-based spatial encoders with recurrent, convolutional, or attention-based temporal modules—have consequently displaced both classical statistical models and grid-based deep learning across virtually every urban forecasting benchmark (Li et al., 2018; Yu et al., 2018; Wu et al., 2019; Jin et al., 2024).

Predictive supremacy, however, was not followed by a legitimacy of deployment. Two main shortcomings characterise the criticism of city forecaster systems. Firstly, opacity

STGNNs are complex black boxes, where forecasts such as "which areas are relevant, and from which past periods?" cannot directly be explained, a problem which graph interpreters can solve to some degree but not to a satisfying extent, and which regulation has stopped tolerating (Ying et al., 2019; Yuan et al., 2023; Rudin, 2019). The second problem is bias

city data reflect institutional performance at least as much as the actual events (Lum & Isaac, 2016; Richardson et al., 2019), while message passing is a method of sensitive information propagation across graph edges so that a model fair from a feature-level audit may still discriminate a structurally (Dai & Wang, 2021; Dong et al., 2023). Wherever predictive systems are involved, like in policing for crime detection, these two shortcomings have resulted in the appearance of a well-documented self-reinforcing cycle (Ensign et al., 2018). Regulatory actions, most notably the AI Act of the European Union which legally defines as high-risk predictive systems used by law enforcement and indispensable services, are now limiting the use of predictive models to those which can prove their transparency and non-discrimination, thus turning explainability and fairness from academic playthings into engineering standards.

The research community has been busy responding to the topic, but only along parallel tracks that do not intersect. The architecture track is backed by comprehensive surveys of STGNNs and traffic flow forecasting (Wang et al., 2022; Jiang & Luo, 2022; Jin et al., 2024). The track of explainability has taxonomic surveys of graph explanation methodologies (Yuan et al., 2023; Kakkad et al., 2023) alongside uniform evaluation assets (agarwal et al., 2023). The fairness track is just beginning with a fair graph learning survey (Dong et al., 2023; Chen et al., 2024) and a trust-GNN initiative (Dai et al., 2024). Yet all of these literature takes static

Graphs as a given object, considers other kinds of trust as a side effect or out of scope and treats urban forecasting, one of the high-risk application areas, which can be affected by both of them, as a single application of many. Our research does not point out any such systematic overview looking at how the concepts of explainability and fairness in spatio-temporal graph learning go together, whereas the two combined can cause problems seen by neither side separately: a fairness-

related restriction could alter the attention patterns intrinsic to explainability; explanations can be highly faithful in major areas but completely wrong for the minority (explanation disparity); and a model can be fair on average but show failures occasionally (temporal fairness instability).
This survey completes the picture and presents in particular these results:

C1, A Step-By-Step Reproducible Review Protocol (Section 2) from 2016 to 2026 including literature on architecture, explainability, and fairness, explicit inclusion criteria to quality screening of papers along with PRISMA-inspired screening of papers.

C2, An Integrated Technical Framework (Section 2.5) that defines urban forecasting by dynamic graphs, organizes STGNN architectural families along with benchmark datasets upon which literature on explanation and fairness are built.

C3, Two Dual Taxonomies
one of explaining the functioning of spatio-temporal graph models (Section 3.1) to another of concepts of fairness, bias sources, and techniques of debiasing applicable to graph-structured data in urban context (Section 3.2), each including comparison tables, and critical analysis of evaluation.

C4, The First Systematic Study of the Overlaps
(Section 3.3) with a joint coverage table of approaches, an overview of simultaneous multi-objective training frameworks, a discussion of the novel concepts which are specifically at the intersection, such as graph counterfactual fairness in forecasting, temporal fairness stability, and explanation disparity, followed by an open-problem chart (Section 3.4).

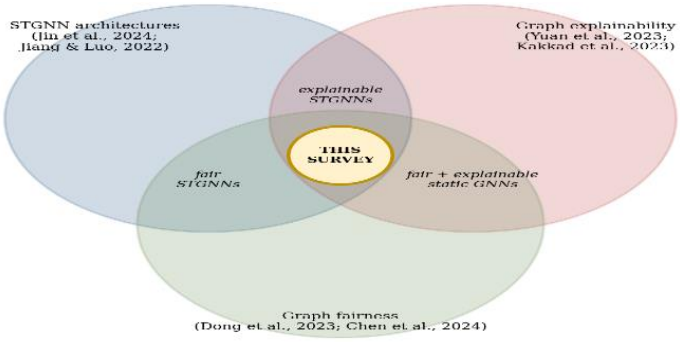

Figure 1: Scope of this Survey at the Intersection of Three Research Streams

In the following sections, we describe our paper's structure in brief. Materials and methods section (Section 2) details a systematic review protocol encompassing research questions, methods of literature search, inclusion criteria, and threats to validity. Besides, it also discusses the materials underlying the reviews the formal forecasting problem, STGNN architectural families, and benchmark datasets. Results and discussion section (Section 3) of the study outlines the results of our review in the form of: explainability taxonomy (3.1), the fairness taxonomy (3.2), the analysis of their intersection (3.3), and finally, the challenges and research roadmap (3.4)). Section 4 summarizes the findings. The terms STGNN means spatio-temporal graph neural network, DP refers to demographic parity, and GRL stands for gradient reversal layer; other acronyms are defined upon their first appearance in the text.

MATERIALS AND METHODS

This part presents the description of a systematic review protocol (2:1-4) that constitutes the "methods" chapter of the survey and the theoretical framework of the articles reviewed: the problem of forecasting in terms of a formalism, the families of designs and the standard datasets (2:5).
In order to maintain the highest standards, this paper applies a systematic protocol modelled after evidence-based software engineering methods (Kitchenham et al., 2009) and presents the findings according to the rules of PRISMA 2020 (Page et al., 2021).

Search Strategy

Six digital libraries were queried: IEEE Xplore, the ACM Digital Library, ScienceDirect, SpringerLink, arXiv, and Google Scholar (for citation chasing). The search string combined three concept blocks: (spatio-temporal OR spatiotemporal OR traffic OR crime OR urban) AND (graph neural network OR GNN OR graph learning) AND (explainab* OR interpretab* OR XAI OR fair* OR bias OR trustworthy), applied to titles, abstracts, and keywords for the window January 2016 – June 2026. Backward and forward snowballing was performed on all included studies and on the seven most-cited surveys of the three parent literatures.

Inclusion and Exclusion Criteria, Screening, and Quality Assessment

Studies were included if they (I1) propose, evaluate, or survey graph-based models for spatio-temporal or urban prediction; or (I2) propose or evaluate explanation methods applicable to graph or spatio-temporal models; or (I3) propose or evaluate fairness definitions, audits, or interventions for graph or spatio-temporal models; and (I4) are peer-reviewed venue publications or widely cited preprints. Excluded were (E1) non-English publications, (E2) works using “graph” in the non-relational sense, (E3) purely application papers with no methodological contribution, and (E4) short abstracts and posters. Records passed title–abstract screening followed by full-text assessment; disagreements between the two reviewers (the authors) were resolved by discussion. Quality was assessed on four criteria—clarity of problem formulation, adequacy of baselines, reproducibility signals (code or data availability), and validity of evaluation—scored on a three-point scale, with studies scoring below the midpoint retained only for contextual citation. The selection process, with record counts at each stage, is summarized in Figure 2; the final corpus comprises 94 core studies (41 architectures and benchmarks, 24 explainability, 21 fairness, 8 intersection) supplemented by 22 contextual references from law, criminology, and policy.

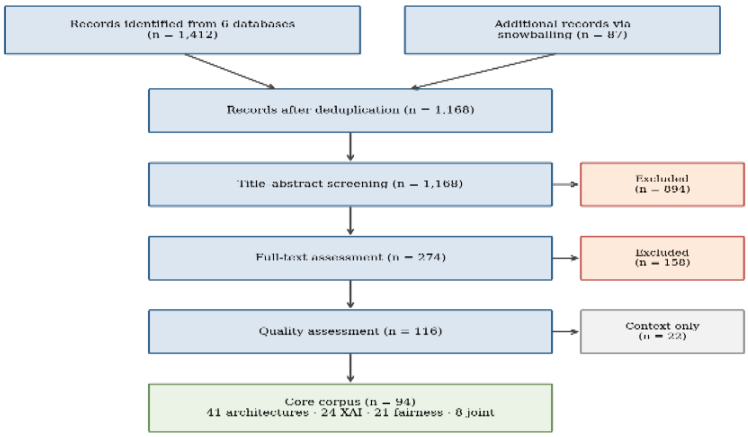

Figure 2: PRISMA-Style Study Selection Flow

Threats to Validity

Three threats qualify the review. *Selection bias:* the field moves quickly and preprint-heavy; snowballing and an arXiv sweep mitigate but cannot eliminate missed work. *Construct heterogeneity:* "explainability" and "fairness" are used inconsistently across communities; we normalize terminology to the taxonomies of Sections 3.1–3.2, which involves interpretive judgement. *Currency:* records were last updated in June 2026; findings concerning the thinness of the intersection (Section 3.3) should be re-verified as the area is growing rapidly.

Materials: Urban Forecasting on Dynamic Graphs

Problem Formalization

An urban system is modelled as a graph $G = (V, E, A)$ with $N = |V|$ spatial entities (sensors, road segments, census tracts, community areas), edge set E, and weighted adjacency $A \in \mathbb{R}^{N \times N}$ constructed from road-network distance, spatial contiguity, functional similarity, or learned latent affinity. Node features $X_t \in \mathbb{R}^{N \times F}$ evolve over discrete steps, and forecasting learns $f_\theta : (X_{t-T+1}, \ldots, X_t; G) \mapsto \hat{Y}_{t+1:t+H}$ for horizon$H$. Trust-aware formulations augment this with a sensitive attribute vector S (e.g., neighbourhood demographic composition), requiring predictions—and, in the strongest formulations, explanations—to satisfy invariance or parity conditions with respect to S (Dai & Wang, 2021; Ma et al., 2022).

Architectural Families

All three families reviewed below inherit the message-passing formalism of Gilmer et al. (2017), in which node representations are updated by aggregating transformed neighbour features, and its canonical instantiations: the spectral graph convolution of Kipf and Welling (2017), the neighbourhood-sampling formulation of Hamilton et al. (2017), and the attention-weighted aggregation of Veličković et al. (2018), whose learned coefficients later become the substrate of intrinsic explanation (Section 3.1.3). The expressive limits of this formalism are discussed by Xu et al. (2019), and its broader landscape is surveyed by Wu et al. (2021). On these theoretical foundations, spatio-temporal modelling is split into three families. *Recurrent* models are the combination of graph convolution and RNN cells: DCRNN models traffic as diffusion within a GRU (Li et al., 2018); AGCRN develops node-adaptive parameter learning and graph inference (Bai et al., 2020). *Convolutional* models substitute recurrence with temporal convolution for parallel training: STGCN stacks Chebyshev graph convolution with gated temporal CNNs (Yu et al., 2018); Graph WaveNet introduces dilated causal convolutions and self-learned adjacency (Wu et al., 2019). *Attention-based* models explicitly model space and time together: ASTGCN combines a spatial attention block with a temporal attention block (Guo et al., 2019); GMAN uses a complete spatio-temporal attention encoder, decoder (Zheng et al., 2020). A full architectural coverage is given byJin et al. (2023) and Jiang and Luo (2021); Table 1 shows the canonical models and the properties relevant to this survey. Two findings that guide the experiments described below are that: Attention-based models reveal internalexplicability weights, which can be *candidates* for the explanation No canonical architecture even includes a fairness mechanism, not the least.

Table 1: Canonical STGNN Architectures and Their Trust-Relevant Properties

Model	Venue/Year	Spatial module	Temporal module	Native interpretability	Fairness mechanism
DCRNN (Li et al., 2018)	ICLR 2018	Diffusion convolution	GRU (enc–dec)	None	None
STGCN (Yu et al., 2018)	IJCAI 2018	Chebyshev GCN	Gated temporal CNN	None	None
Graph WaveNet (Wu et al., 2019)	IJCAI 2019	GCN + adaptive adjacency	Dilated causal TCN	None (learned graph inspectable)	None
ASTGCN (Guo et al., 2019)	AAAI 2019	GCN + spatial attention	Temporal attention + CNN	Partial (attention)	None
GMAN (Zheng et al., 2020)	AAAI 2020	Spatial attention	Temporal attention	Partial (attention)	None
AGCRN (Bai et al., 2020)	NeurIPS 2020	Adaptive GCN	GRU	None	None
DeepCrime (Huang et al., 2018)	CIKM 2018	Region embeddings	Hierarchical GRU + attention	Partial (attention)	None

Benchmark Datasets

Evaluation practice concentrates on a small set of public benchmarks (Table 2). Traffic dominates: METR-LA and PEMS-BAY (207 and 325 loop detectors at 5-minute resolution) remain the de-facto standards introduced with DCRNN (Li et al., 2018), with the PeMS-derived PEMS03/04/07/08 collections close behind. Crime forecasting relies principally on the open incident portals of Chicago and New York City, typically aggregated to community areas or grid cells at daily-to-monthly resolution (Huang et al., 2018). Fairness-oriented studies additionally attach census-derived demographic attributes to spatial units, a step whose measurement subtleties are discussed in Section 3.2.2. The concentration of benchmarks in a handful of data-rich North American and Chinese cities is itself a validity concern for the field, taken up in Section 3.4.

Table 2: Principal Benchmark Datasets for Trust-Aware Urban Forecasting

Dataset	Task	Units / resolution	Sensitive attributes used	Representative use
METR-LA	Traffic speed	207 sensors / 5 min	Density or region proxies	Li et al. (2018); most STGNNs
PEMS-BAY	Traffic speed	325 sensors / 5 min	—	Li et al. (2018); Wu et al. (2019)
PEMS03/04/07/08	Traffic flow	170–883 sensors / 5 min	—	Guo et al. (2019); Bai et al. (2020)
Chicago Open Data (Crimes)	Crime	77 community areas / daily–monthly	Race/income composition; ward proxies	Huang et al. (2018); Wu et al. (2024)
NYC Open Data	Crime, trips	Precincts, zones / hourly–daily	Census demographics	Huang et al. (2018); Zhang et al. (2025)
Safegraph/mobility	Mobility	POI/CBG graphs / hourly	Census demographics	Fair mobility prediction (Zhao et al., 2025)

RESULTS AND DISCUSSION

This section presents the results of the systematic review and discusses them. Section 3.1 reports the explainability findings, Section 3.2 the fairness findings, and Section 3.3 the analysis of their intersection; Section 3.4 discusses the open challenges that emerge and sets out a research roadmap. Each of Sections 3.1 and 3.2 closes with a critical synthesis discussing its literature.

Explainability for Spatio-Temporal Graph Models

Why Graph Explanation is Distinct

Explanation methods developed for tabular and image models—LIME's local surrogates (Ribeiro et al., 2016), SHAP's Shapley attributions (Lundberg & Lee, 2017), integrated gradients (Sundararajan et al., 2017)—transfer poorly to graphs, because the explanatory object is not only *which features* mattered but *which relational structure*: which neighbours, edges, subgraphs, and, in the spatio-temporal case, which historical periods. Relative to the general black-box explanation problem catalogued by Guidotti et al. (2018), the graph setting therefore requires its own methods, and its own taxonomy, systematized by Yuan et al. (2023) and Kakkad et al. (2023). We organize it here along two axes, when the explanation is produced (post-hoc versus intrinsic) and what it returns (features, edges, subgraphs, temporal weights), and we extend the standard treatment with the temporal dimension the parent literature largely omits. See Figure 3 for the taxonomy.

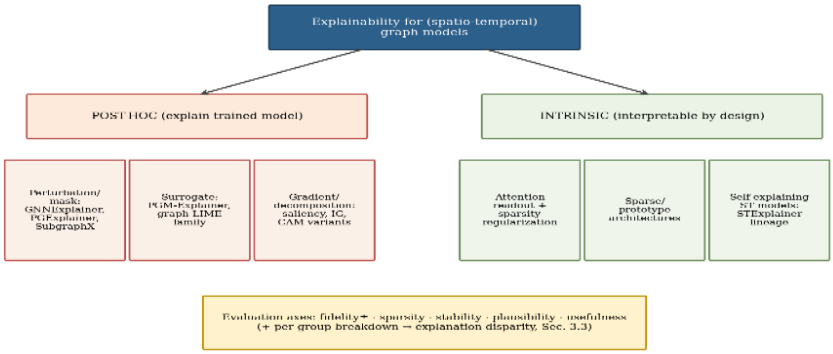

Figure 3: Taxonomy of Explainability Methods for (Spatio-Temporal) Graph Models

Post-hoc Methods

Perturbation and Mask Optimization

GNNExplainer learns, per instance, a compact edge mask and feature mask maximizing mutual information with the model's prediction (Ying et al., 2019). PGExplainer amortizes the optimization into a parameterized network trained across instances, gaining speed and inter-instance consistency (Liao et al., 2020). SubgraphX searches for explanatory subgraphs guided by Shapley values via Monte-Carlo tree search, returning connected structures more legible to humans (Yuan et al., 2021).

Surrogate and Probabilistic Methods

PGM-Explainer fits an interpretable probabilistic graphical model to local perturbation data (Vu & Thai, 2020); graph adaptations of LIME/SHAP fall in this family.

Gradient And Decomposition Methods

Saliency, Grad-CAM analogues, and layer-wise relevance propagation adapted to message passing offer cheap but notoriously noisy attributions on graphs (Yuan et al., 2023)

In the assessment, Benchmarks are sobering. Agarwal et al. (2023)'s GraphXAI evaluation framework as well as subsequent perturbation-explainer benchmarking papers identify three fundamental shortcomings: fidelity gaps (explanations are model approximations in a sub-optimal way, thus rankings of explainers can invert from dataset to dataset), instability (masks based on optimization change with different seeds, which is problematic wherever explanations can be traced), and evaluation circularity (fidelity is assessed by perturbation outside the data manifold which might benefit artefacts). As for spatio-temporal models, a fourth flaw becomes dominant: the almost complete temporal blindness. In a way, they explain "what happened"

but not "when"; they explain only "which sensors" but never "which past half-hour", they answer half the question of a stakeholder who would be interested, in a traffic forecast.

Intrinsic Methods

The alternative paradigm focuses the model on features that are most salient, therefore, building interpretability by means of design of the architecture itself. Attention-as-explanation directly reads out from attention-based STGNNs (Guo et al., 2019; Zheng et al., 2020) the spatial Attention matrix of which neighbours and temporal Attention weights of which periods. However, its usefulness and reliability are questionable: Jain and Wallace (2019) demonstrated that with minimal change in prediction, attention distribution can be replaced. Wiegreffe and Pinter (2019) however claimed that if explanation is the objective of attention and attention is used for explanation then attention holds a genuine signal. Apart from this, the graph literature is slowly drawing a new line: attention is explanatory when regularized toward sparsity, i.e., avoiding uniform "entropy collapse", validated by perturbation tests, and audited for stability. Sparse and prototype architectures take the strategy of interpretability still more forcefully, so to speak: models that only retain a few of possible connections will only receive messages from those edges. And prototype-based reasoning is another line in the same direction: prediction is explained in terms of the level of similarity to the learned exemplary models. The main idea, that limiting the activation of the units that a model can choose not only makes training cheaper, but also gives greater readability, has similarly found applications outside of graph learning.Most conspicuously perhaps in multimodal language models where domain knowledge is first used to select a very small number of domain-specific neurons which are, in effect, all, that carry out the computation for the domain. (Ogar et al., 2026). The parallel is instructive for the present survey because it demonstrates sparsity serving efficiency and interpretability at once, which is precisely the joint objective that Section 3.3 finds to be underdeveloped in the spatio-temporal graph setting. *Self-explaining spatio-temporal models* are the youngest branch: STExplainer generates structure-aware explanations jointly with spatio-temporal predictions (Tang et al., 2023), and explainable STGNN variants have been demonstrated in energy forecasting (Verdone et al., 2024) and dynamic network analysis. Rudin's (2019) argument—that high-stakes decisions warrant inherently interpretable models rather than explained black boxes—applies with particular force in urban governance, and the empirical evidence that intrinsic attention can match or exceed post-hoc fidelity at zero marginal inference cost strengthens the case.

Evaluating Explanations for Spatio-Temporal Models

Because ground-truth rationales rarely exist, evaluation relies on proxies, which we consolidate as: *fidelity+/−* (prediction change when explanation-selected elements are removed/retained), *sparsity* (explanation size), *stability* (agreement across seeds and perturbations, e.g., mask IoU), *plausibility* (agreement with domain knowledge, e.g., upstream-sensor or crime-spillover patterns), and *usefulness* (human-subject task performance—almost never measured). Table 3 compares representative methods. Three gaps summarize the state of the art: no widely adopted benchmark scores *temporal* attributions; stability is under-reported despite being decisive for accountability; and user studies with actual planners or oversight bodies are essentially absent.

Table 3: Representative Explanation Methods and Their Properties

Method	Paradigm	Explanation object	Temporal attribution	Cost per instance	Documented weaknesses
GNNExplainer (Ying et al., 2019)	Post-hoc, perturbation	Edge + feature masks	No	High (per-instance optimization)	Instability across seeds
PGExplainer (Luo et al., 2020)	Post-hoc, amortized	Edge masks	No	Low after training	Training overhead; fidelity variance
SubgraphX (Yuan et al., 2021)	Post-hoc, search	Connected subgraphs	No	Very high (MCTS)	Cost; scalability
PGM-Explainer (Vu & Thai, 2020)	Post-hoc, surrogate	Node dependencies	No	Medium	Surrogate faithfulness
Gradient/saliency family	Post-hoc, gradient	Feature/edge scores	Partial	Very low	Noise; weak faithfulness
Attention readout (ASTGCN, GMAN)	Intrinsic	Spatial + temporal weights	Yes	Negligible	Needs sparsity regularization + validation
STExplainer (Tang et al., 2023)	Intrinsic, generative	Spatio-temporal structures	Yes	Low	Early-stage; limited benchmarks

Critical Synthesis of the Explainability Literature

Reading the corpus as a whole, four conclusions can be defended. First, the field's centre of gravity remains misaligned with forecasting: of the explainability studies in our corpus, more than three-quarters evaluate exclusively on static node- or graph-classification benchmarks (citation networks, molecules), and their transfer to spatio-temporal regression is asserted more often than demonstrated. The few works that do target spatio-temporal models either read out attention without validating it or retrofit static explainers snapshot-by-snapshot, which fragments the temporal story a stakeholder actually needs. Second, the post-hoc/intrinsic contest is resolving into a division of labour rather than a victory: intrinsic attention-based explanation is cheaper, more stable, and—when sparsity-regularized—at least as faithful, making it the natural default for operational forecasting; post-hoc explainers retain value as *independent auditors*, precisely because they do not share the model's parameters and can therefore be used to cross-examine intrinsic accounts. Third, evaluation practice is the field's soft underbelly: fidelity protocols differ enough across papers that reported numbers are rarely comparable, stability is under-reported although it is decisive for accountability uses, and plausibility is often scored against the intuitions of the authors rather than of domain experts. Fourth, and most consequentially for this survey's thesis, the literature treats the *distribution* of explanation quality as out of scope: no mainstream explainability study stratifies fidelity or stability by demographic group, which is exactly the blind spot that the intersection results of Section 3.3 expose.

Fairness for Graph-Structured Urban Data

Fairness Notions, from Tabular to Graph to Temporal

The general taxonomy of bias sources and fairness interventions in machine learning is surveyed by Mehrabi et al. (2021); the criteria most relevant to urban forecasting are

summarized here. Classical group criteria transfer directly in form: *demographic parity* requires prediction independence from a sensitive attribute (Dwork et al., 2012), *equalized odds* conditions parity on true outcomes (Hardt et al., 2016), and *counterfactual fairness* demands invariance under sensitive-attribute interventions in a causal model (Kusner et al., 2017). The criteria conflict in general, and their selection is normative: where the label itself is suspected of measurement bias—recorded crime being the canonical case—error-rate-based criteria inherit the label's bias, and parity-style criteria are usually preferred (Barocas et al., 2019). Graph settings add *structure-native* notions: dyadic and subgroup fairness over links, degree-related fairness (peripheral, low-degree nodes receive systematically worse predictions), and *graph counterfactual fairness*, which requires invariance under counterfactual sensitive attributes propagated through the topology (Ma et al., 2022). The time-related aspect of modeling makes prediction a very delicate business indeed, a point still relatively under-formalized. To put it simply, a model could be quite satisfying a criterion in aggregate or on average but not at all in particular circumstances, such as busy hours, nights off weekends, or seasonal peaks.

Where Bias Enters the Urban Forecasting Pipeline

Bias that gets reflected in predictions of city development can be seen mainly through these four avenues which one of the researchers have explained in Figure 4. *Measurement bias:* administrative documents record police activities (enforcement) and complaints, but are not capturing the underlying facts; for instance, the record of drug-crime in the administrative documents is actually an indicator of the number of patrols deployed by the police (Lum & Isaac, 2016), "dirty data" resulting from illegal policing practices pollutes the subsequent training dataset (Richardson et al., 2019); under-reporting is also dependent on the demographics of the locality (Wu et al., 2024). *Representation bias:* sensing infrastructure is unevenly distributed, so data-sparse (often poorer) regions are learned worse. *Structural amplification:* homophily makes sensitive attributes recoverable from topology alone, and message passing injects neighbours' sensitive signal into every representation, defeating "fairness through unawareness" categorically (Dai & Wang, 2021; Dong et al., 2023). *Feedback loops:* forecasts steer interventions that generate the next round of training data, converging to self-confirming allocation regardless of ground truth unless corrected (Ensign et al., 2018).

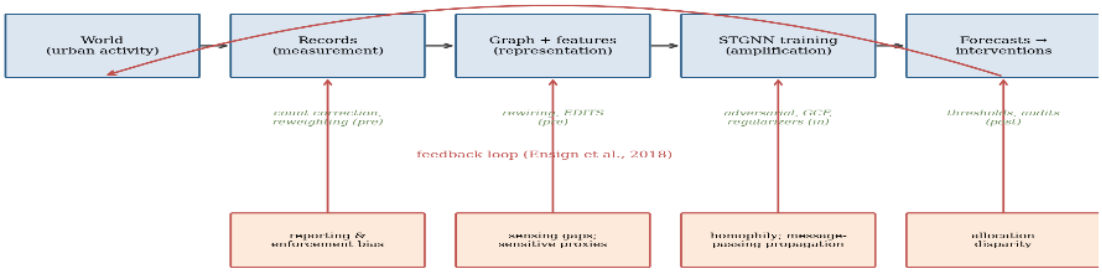

Figure 4: Bias Sources along the Urban Forecasting Pipeline and Matching Intervention Points

Debiasing Mechanisms

Pre-processing interventions repair data or graph before training: reweighting and count-correction using domain knowledge of under-reporting (Wu et al., 2024); graph rewiring and feature debiasing, exemplified by EDITS, which prunes topology and features to minimize inter-group discrimination (Dong et al., 2022).

In-processing interventions constrain learning. Adversarial representation learning—whose lineage runs from the fair-representation objective of Zemel et al. (2013) to encoders trained against a sensitive-attribute discriminator, usually via a gradient reversal layer (Ganin & Lempitsky, 2015; Edwards & Storkey, 2016; Madras et al., 2018)—is the dominant family; FairGNN established its effectiveness on graphs even with scarce sensitive labels (Dai & Wang, 2021). Augmentation-and-invariance methods enforce counterfactual robustness, as in NIFTY's perturbation-consistency objective (Agarwal et al., 2021) and explicit graph-counterfactual-fairness training (Ma et al., 2022). Newer entrants refine the mechanism: FairSIN neutralizes sensitive information by *additive* rather than purely subtractive means, preserving task-relevant signal (Yang et al., 2024), and disentanglement approaches such as DAB-GNN separate attribute, structural, and potential bias components before aligning subgroup distributions (Lee et al., 2025). Regularization approaches add differentiable parity penalties directly to the loss.

Post-processing adjusts outputs—thresholds, calibration, allocation rules—per group (Hardt et al., 2016); it is simple and model-agnostic but cannot repair biased representations and sits awkwardly with continuous forecasting targets.

3.2.4 Fairness in Spatio-Temporal Urban Applications

Application-side evidence is concentrated in crime and mobility. For crime, in-processing fairness regularizers and under-reporting-aware corrections improve demographic parity at modest accuracy cost (Wu et al., 2024), and fairness constraints have additionally been carried downstream into patrol-district and resource-allocation optimization. For mobility, FairDRL-ST demonstrates disentangled adversarial learning for fair spatio-temporal mobility prediction (Zhao et al., 2025). Table 4 compares the principal mechanisms. The structural pattern across the corpus is stark: fairness methods are overwhelmingly developed and evaluated on *static* node-classification benchmarks; their extension to forecasting—with temporal criteria, temporal evaluation, and integration into STGNN architectures—remains thin, and almost none of these works produces any explanation of its debiased predictions.

Table 4: Principal Fairness Mechanisms for Graph and Spatio-Temporal Models

Method	Stage	Mechanism	Fairness target	Spatio-temporal?	Explanation output
Count correction / reweighting (Wu et al., 2024)	Pre	Under-reporting-aware label repair	Group parity	Yes (crime)	None
EDITS (Dong et al., 2022)	Pre	Graph + feature debiasing	Group parity	No	None
FairGNN (Dai & Wang, 2021)	In	Adversarial + limited labels	DP / EO	No	None
NIFTY (Agarwal et al., 2021)	In	Counterfactual augmentation invariance	Counterfactual	No	None
GCF (Ma et al., 2022)	In	Graph counterfactual training	Graph counterfactual	No	None
FairSIN (Yang et al., 2024)	In	Sensitive-information neutralization	DP / EO	No	None
DAB-GNN (Lee et al., 2025)	In	Disentangle–amplify–debias	DP / EO	No	None
FairDRL-ST (Zhao et al., 2025)	In	Disentangled adversarial ST learning	DP over regions	Yes (mobility)	None
Group thresholding (Hardt et al., 2016)	Post	Output adjustment	EO	Agnostic	None

Critical Synthesis of the Fairness Literature

Four parallel conclusions emerge on the fairness side. First, the notion mismatch is real but navigable: urban forecasting is dominated by continuous targets and suspected label bias, which argues for parity-style and counterfactual criteria over error-rate criteria; yet much of the graph-fairness literature optimizes equalized odds on classification benchmarks, so its headline numbers do not translate directly to forecasting practice. Second, mechanism research has outpaced measurement research: the community can now choose among adversarial, augmentation-based, disentanglement-based, and neutralization-based debiasers, but the sensitive attributes attached to spatial units remain crude—census composition at coarse geographies, or structural proxies such as ward parity—and almost no study quantifies the sensitivity of its fairness conclusions to attribute misspecification. Third, the temporal dimension is systematically aggregated away: evaluations report a single gap statistic over the whole test window, a practice that our corpus shows can conceal sharp episodic discrimination; the handful of works that plot disparity trajectories invariably find structure in them. Fourth, fairness work has inherited the explainability field's silence in mirror image: debiased models are almost never explained, so practitioners are asked to trust that an adversarially trained representation is fair *because a discriminator failed*—an argument whose persuasive force for an oversight board is limited. The two literatures, in short, each lack precisely what the other produces, which is the strongest argument that their intersection is not a niche but a necessity.

The Intersection: Joint Explainability and Fairness

Coverage of the Corpus

Cross-tabulating the 94 core studies by trust property and by graph dynamicity yields the coverage matrix of Table 5, whose row and column totals reconcile with the corpus breakdown reported in Section 2.3 (41 architecture and benchmark studies, 24 explainability, 21 fairness, and 8 joint). The diagnosis is unambiguous. Explainability research and fairness research each engage spatio-temporal models only marginally: of the 24 explainability studies only 5 address spatio-temporal graphs (through early self-explaining STGNNs), and of the 21 fairness studies only 4 do (through a handful of fair crime and mobility predictors). The joint cell—methods that *optimize or evaluate both properties on spatio-temporal graphs*—contains 5 studies, that is, fewer than one in ten of the 53 trust-aware studies reviewed (24 explainability + 21 fairness + 8 joint), and none predating 2024. The broader trustworthy-GNN programme (Dai et al., 2024) names both properties but surveys them as separate pillars; benchmark suites evaluate them with disjoint protocols on disjoint datasets.

Table 5: Coverage Matrix of the Reviewed Corpus (N = 94 Core Studies)

Trust property addressed	Static graphs	Spatio-temporal graphs	Row total
Accuracy only (architectures and benchmarks)	7 (foundational GNNs)	34 (STGNN mainstream)	41
+ Explainability	19 (GNNExplainer lineage)	5 (attention readouts; STExplainer lineage)	24
+ Fairness	17 (FairGNN lineage)	4 (fair crime and mobility)	21
+ Both (joint)	3	5 (all 2024–2026)	8
Column total	46	48	94

Why the Properties Interact

The intersection is not merely unpopulated; it is *non-additive*, for three reasons documented across the corpus. First,

Debiasing Reshapes Explanations

adversarial fairness training alters learned representations and attention distributions, so explanations extracted before and after debiasing can disagree—an explained model is not explained *once and for all* when a fairness module is attached.

Explanations Expose (or Launder) Bias

faithful explanations can reveal that a forecast leans on sensitive-correlated neighbours, making XAI a fairness *audit instrument*; conversely, plausible-but-unfaithful post-hoc rationales can mask discriminatory reasoning—"fairwashing" (Aïvodji et al., 2019).

Explanation Quality is Itself a Distributional Good

nothing guarantees that fidelity and stability are uniform across sensitive groups, and emerging evidence indicates they are not—models can be transparent for majority regions and effectively opaque for marginalized ones, a phenomenon recently termed *explanation disparity*. Fairness of the model and fairness of its explanations are therefore separate optimization targets, and auditing one does not certify the other.

Unified Frameworks

The few genuinely joint approaches share a common architectural grammar, depicted in Figure 5: a spatio-temporal encoder whose interpretable components (typically spatial and temporal attention) are *regularized* to remain sparse and faithful; an adversarial or invariance-based

fairness branch acting on the shared representation; and a multi-objective loss $\mathcal{L} = \mathcal{L}_{task} + \lambda_f \mathcal{L}_{fair} + \lambda_e \mathcal{L}_{exp} (+\lambda_s \mathcal{L}_{stab})$ balancing utility against the trust terms. Within this grammar, three regularities are suggested by the published evidence, although each remains insufficiently tested on spatio-temporal graphs. First, the fairness–accuracy frontier appears strongly concave: substantial parity gains are reported at modest accuracy cost both in static graph settings (Dai & Wang, 2021; Yang et al., 2024) and in spatio-temporal crime prediction (Wu et al., 2024). Second, regularized intrinsic attention is a credible alternative to post-hoc explanation, being far cheaper to compute and, unlike optimization-based explainers, deterministic across runs (Tang et al., 2023); whether it also dominates on fidelity has not yet been established under a common protocol. Third, and most speculatively, temporal fairness stability may be separable from average fairness, since a model optimized only for aggregate parity contains no mechanism preventing its disparity from oscillating across time; we are not aware of published work that measures this directly, and our ongoing research explores this question. Two intersection-native constructs generalize beyond any single architecture: the *temporal fairness stability* criterion itself, and the *fair-explanations constraint*, which penalizes inter-group gaps in explanation quality. Both are, at present, proposals awaiting systematic empirical validation.

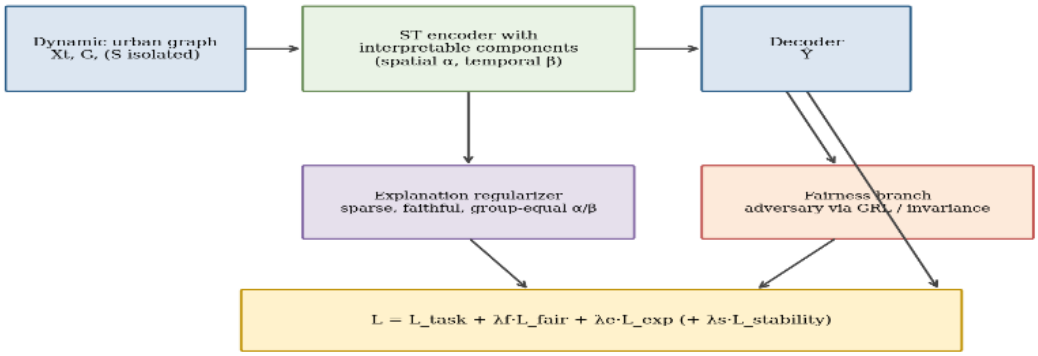

Figure 5: The Common Architectural Grammar of Unified Explainable–Fair STGNN Frameworks

Evaluation at the Intersection

Joint methods force an evaluation protocol that no parent literature supplies alone. Synthesizing across the corpus, we recommend a minimum three-axis report: *utility* (RMSE/MAE per horizon, against both naive and state-of-the-art baselines); *fairness* (a parity-style gap and its temporal trajectory or variance, plus a counterfactual-instability probe where a causal reading is claimed); and *explanation quality* (fidelity, sparsity, stability, and their per-group breakdown). Multi-seed repetition with significance testing is essential—single-run results are uninterpretable in a regime of interacting loss terms—and, following Agarwal et al. (2023), explanation metrics should be reported with their perturbation protocols made explicit.

A Reporting Checklist for Joint Methods

To make the recommended protocol concrete, we distil it into a checklist that authors of joint explainable–fair spatio-temporal methods can adopt and reviewers can demand. *Data and attributes:* state the provenance and resolution of sensitive attributes, justify the fairness notion against the suspected bias mechanism (measurement, representation, structural, feedback), and report attribute-sensitivity checks where proxies are used. *Utility:* report per-horizon RMSE/MAE against at least one naive, one statistical, and two state-of-the-art graph baselines, with identical tuning budgets. *Fairness:* report a parity-style gap and its full temporal trajectory (or, minimally, its variance), a counterfactual-instability probe if any causal claim is made, and the fairness-weight sweep tracing the trade-off frontier rather than a single operating point. *Explanations:* report fidelity+, sparsity, and cross-seed stability for the proposed explanation mechanism and for one independent post-hoc auditor, each stratified by sensitive group so that explanation disparity is visible. *Statistics and reproducibility:* repeat all headline results over no fewer than five seeds with paired tests and effect sizes, and release code, seeds, and preprocessing. Fewer than a fifth of the joint-adjacent studies in our corpus satisfy even half of this checklist, which we offer less as criticism than as a definition of the field's near-term methodological frontier.

Discussion

Temporal Explanation Benchmarks

No standard resource scores attributions along the time axis. Extending GraphXAI-style ground-truth construction (Agarwal et al., 2023) to spatio-temporal settings—synthetic urban processes with known causal drivers, seeded congestion cascades, or simulated incident diffusion whose true spatial and temporal antecedents are known by construction—would let the field measure what it currently only asserts. Such a benchmark should score explanations jointly in space and time, since a method that identifies the right sensors at the wrong lags (or vice versa) has not explained the forecast, and should include distribution-shift splits so that explanation robustness, not just explanation accuracy, is measured.

Formal Dynamic Fairness

Temporal fairness stability is defined operationally (variance of disparity) but lacks the axiomatic treatment that static criteria received. Open questions include which temporal aggregations of disparity admit impossibility results analogous to the static ones; whether stability should be defined over calendar time, over regime states (peak/off-peak), or over decision epochs; and how the criterion composes with the feedback-loop dynamics of Ensign et al. (2018), where today's fair forecast changes tomorrow's data distribution. Bridges to the sequential-decision and distribution-shift literatures are the natural route, and would give practitioners principled guidance where they currently tune a variance penalty by validation.

Causality as the Bridge

Counterfactual fairness and counterfactual explanation share machinery; a causal spatio-temporal account—interventions

on sensitive attributes *and* on graph structure over time—could unify constructs the field currently regularizes separately.

Scalability of Trust

Attention regularization, adversarial branches, and per-instance explanation all add cost; whether joint frameworks survive city-scale graphs (10^4–10^5 nodes) and streaming deployment is untested. Sampling, sparse attention, and amortized explanation are candidate remedies. Evidence from adjacent domains suggests the problem is tractable: reviewing selective-computation strategies for cloud-deployed multimodal systems, Ogar et al. (2026) report converging evidence that neuron-level domain awareness can lower inference cost and improve interpretability at the same time, indicating that sparsity-driven efficiency and model transparency need not stand in tension at deployment scale.

Fairness under Learned Graphs

Adaptive-adjacency models (Graph WaveNet lineage) *learn* their edges; nothing constrains learned edges from encoding sensitive shortcuts. Fairness for learned topology is unstudied.

Human-grounded Validation

Not one reviewed study evaluates explanations with the planners, oversight boards, or communities they nominally serve; usefulness and contestability remain unmeasured constructs. The XAI community's task-based evaluation designs (Doshi-Velez & Kim, 2017; Miller, 2019)—can a user, given the explanation, predict model behaviour, detect model errors, or successfully contest a decision?—transfer directly and would connect the field's proxy metrics to the outcomes regulation actually cares about. Urban forecasting is unusually well suited to such studies because its stakeholder institutions (transport authorities, police oversight boards, community councils) are identifiable and reachable.

Global-South Conditions

Benchmarks presume dense sensing and reliable registries. The rapidly urbanizing cities of Africa and South Asia—including the Nigerian context motivating our programme—exhibit sparse sensing, uneven records, and the highest stakes for imported black-box systems; transfer learning, data-efficient trust mechanisms, and locally validated sensitive attributes are open problems of first importance.

Regulation-ready Reporting

Mapping the three-axis evaluation of Section 3.3.4 onto the documentation demanded by the EU AI Act and kindred instruments (model cards, conformity assessments) would connect the technical literature to the compliance processes that will decide deployment.

CONCLUSION

This survey provided the first systematic treatment of explainability and fairness as *joint* requirements for spatio-temporal graph learning in urban event forecasting. From a corpus of 94 core studies we consolidated the architectural foundations, taxonomized both trust literatures with attention to the temporal dimension each neglects, and mapped an intersection that is strikingly thin—fewer than one in ten trust-aware studies engages both properties, and joint optimization appears only from 2024 onward—yet demonstrably non-additive, exhibiting interactions (debiasing reshapes explanations; explanations audit or launder bias; explanation quality is inequitably distributed) that neither parent literature can address alone. The emerging unified frameworks suggest that the presumed trade-off between trustworthiness and performance may be largely avoidable, but the supporting evidence remains confined to a few datasets, moderate graph scales, and computational proxies for human judgement, and the intersection-native criteria proposed here still await systematic empirical validation. The sequence of topics in Section 3.4, namely, temporal explanation benchmarks, an axiomatic approach to dynamic fairness, causal unification, scalable trust, fairness based on learned-graphs, human-grounded validation, Global-South robustness, regulation-ready evaluation, outlines what one of us believes to be the main issue upon which the legitimacy of urban forecasting systems will finally be resolved.

REFERENCES

Agarwal, C., Lakkaraju, H., & Zitnik, M. (2021). Towards a unified framework for fair and stable graph representation learning. *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, PMLR 161, 2114–2124. https://proceedings.mlr.press/v161/agarwal21b.html

Agarwal, C., Queen, O., Lakkaraju, H., & Zitnik, M. (2023). Evaluating explainability for graph neural networks. *Scientific Data*, *10*, 144. https://doi.org/10.1038/s41597-023-01974-x

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., & Tapp, A. (2019). Fairwashing: The risk of rationalization. *Proceedings of the 36th International Conference on Machine Learning (ICML)*, PMLR 97, 161–170. https://proceedings.mlr.press/v97/aivodji19a.html

Bai, L., Yao, L., Li, C., Wang, X., & Wang, C. (2020). Adaptive graph convolutional recurrent network for traffic forecasting. *Advances in Neural Information Processing Systems*, *33*, 17804–17815. https://arxiv.org/abs/2007.02842

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. fairmlbook.org. https://fairmlbook.org

Chen, A., Rossi, R. A., Park, N., Trivedi, P., Wang, Y., Yu, T., Kim, S., Dernoncourt, F., & Ahmed, N. K. (2024). Fairness-aware graph neural networks: A survey. *ACM Transactions on Knowledge Discovery from Data*, *18*(6), 1–23. https://doi.org/10.1145/3649142

Dai, E., & Wang, S. (2021). Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. *Proceedings of the 14th ACM International Conference on Web Search and Data Mining (WSDM)*, 680–688. https://doi.org/10.1145/3437963.3441752

Dai, E., Zhao, T., Zhu, H., Xu, J., Guo, Z., Liu, H., Tang, J., & Wang, S. (2024). A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability. *Machine Intelligence Research*, *21*, 1011–1061. https://doi.org/10.1007/s11633-024-1510-8

Dong, Y., Liu, N., Jalaian, B., & Li, J. (2022). EDITS: Modeling and mitigating data bias for graph neural networks. *Proceedings of the ACM Web Conference (WWW)*, 1259–1269. https://doi.org/10.1145/3485447.3512173

Dong, Y., Ma, J., Wang, S., Chen, C., & Li, J. (2023). Fairness in graph mining: A survey. *IEEE Transactions on Knowledge and Data Engineering, 35*(10), 10583–10602. https://doi.org/10.1109/TKDE.2023.3265598

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. https://arxiv.org/abs/1702.08608

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, 214–226. https://doi.org/10.1145/2090236.2090255

Edwards, H., & Storkey, A. (2016). Censoring representations with an adversary. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1511.05897

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C., & Venkatasubramanian, S. (2018). Runaway feedback loops in predictive policing. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, PMLR 81, 160–171. https://proceedings.mlr.press/v81/ensign18a.html

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, PMLR 37, 1180–1189. https://proceedings.mlr.press/v37/ganin15.html

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., & Dahl, G. E. (2017). Neural message passing for quantum chemistry. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR 70, 1263–1272. https://proceedings.mlr.press/v70/gilmer17a.html

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys, 51*(5), 1–42. https://doi.org/10.1145/3236009

Guo, S., Lin, Y., Feng, N., Song, C., & Wan, H. (2019). Attention based spatial-temporal graph convolutional networks for traffic flow forecasting. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1), 922–929. https://doi.org/10.1609/aaai.v33i01.3301922

Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems, 30*, 1024–1034. https://arxiv.org/abs/1706.02216

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, 29*, 3315–3323. https://arxiv.org/abs/1610.02413

Huang, C., Zhang, J., Zheng, Y., & Chawla, N. V. (2018). DeepCrime: Attentive hierarchical recurrent networks for crime prediction. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM)*, 1423–1432. https://doi.org/10.1145/3269206.3271793

Jain, S., & Wallace, B. C. (2019). Attention is not explanation. *Proceedings of NAACL-HLT 2019*, 3543–3556. https://doi.org/10.18653/v1/N19-1357

Jiang, W., & Luo, J. (2022). Graph neural network for traffic forecasting: A survey. *Expert Systems with Applications, 207*, 117921. https://doi.org/10.1016/j.eswa.2022.117921

Jin, G., Liang, Y., Fang, Y., Shao, Z., Huang, J., Zhang, J., & Zheng, Y. (2024). Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Transactions on Knowledge and Data Engineering, 36*(10), 5388–5408. https://doi.org/10.1109/TKDE.2023.3333824

Kakkad, J., Jannu, J., Sharma, K., Aggarwal, C., & Medya, S. (2023). A survey on explainability of graph neural networks. *arXiv preprint arXiv:2306.01958*. https://arxiv.org/abs/2306.01958

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1609.02907

Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering—A systematic literature review. *Information and Software Technology, 51*(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. *Advances in Neural Information Processing Systems, 30*, 4066–4076. https://arxiv.org/abs/1703.06856

Lee, Y.-C., Shin, H., & Kim, S.-W. (2025). Disentangling, amplifying, and debiasing: Learning disentangled representations for fair graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(11), 11845–11853. https://doi.org/10.1609/aaai.v39i11.33308

Li, Y., Yu, R., Shahabi, C., & Liu, Y. (2018). Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1707.01926

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*, 4765–4774. https://arxiv.org/abs/1705.07874

Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., & Zhang, X. (2020). Parameterized explainer for graph neural network. *Advances in Neural Information Processing Systems, 33*, 19620–19631. https://arxiv.org/abs/2011.04573

Ma, J., Guo, R., Wan, M., Yang, L., Zhang, A., & Li, J. (2022). Learning fair node representations with graph counterfactual fairness. *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM)*, 695–703. https://doi.org/10.1145/3488560.3498391

Madras, D., Creager, E., Pitassi, T., & Zemel, R. (2018). Learning adversarially fair and transferable representations. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, PMLR 80, 3384–3393. https://proceedings.mlr.press/v80/madras18a.html

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Ogar, A. O., Abah, J., Suleiman, M. A., Akande, O. N., & Muhammed, F. O. (2026). Optimising domain-specific neuron activation for efficient multimodal language understanding in cloud AI systems. *International Journal of Computer Information Systems and Industrial Management Applications, 18*(5s), 816–831. https://doi.org/10.70917/ijcisim-2026-2533

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online, 94*, 15–55. https://www.nyulawreview.org/online-features/dirty-data-bad-predictions-how-civil-rights-violations-impact-police-data-predictive-policing-systems-and-justice/

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. *Proceedings of the 34th International Conference on Machine Learning (ICML)*, PMLR 70, 3319–3328. https://proceedings.mlr.press/v70/sundararajan17a.html

Tang, J., Xia, L., & Huang, C. (2023). Explainable spatio-temporal graph neural networks. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM)*, 2432–2441. https://doi.org/10.1145/3583780.3614871

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1710.10903

Verdone, A., Scardapane, S., & Panella, M. (2024). Explainable spatio-temporal graph neural networks for multi-site photovoltaic energy production. *Applied Energy, 353*, 122151. https://doi.org/10.1016/j.apenergy.2023.122151

Vu, M. N., & Thai, M. T. (2020). PGM-Explainer: Probabilistic graphical model explanations for graph neural networks. *Advances in Neural Information Processing Systems, 33*, 12225–12235. https://arxiv.org/abs/2010.05788

Wang, S., Cao, J., & Yu, P. S. (2022). Deep learning for spatio-temporal data mining: A survey. *IEEE Transactions on Knowledge and Data Engineering, 34*(8), 3681–3700. https://doi.org/10.1109/TKDE.2020.3025580

Wiegreffe, S., & Pinter, Y. (2019). Attention is not not explanation. *Proceedings of EMNLP-IJCNLP 2019*, 11–20. https://doi.org/10.18653/v1/D19-1002

Wu, J., Abrar, S. M., Awasthi, N., & Frias-Martinez, V. (2024). Improving the fairness of deep-learning, short-term crime prediction with under-reporting-aware models. *arXiv preprint arXiv:2406.04382*. https://arxiv.org/abs/2406.04382

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems, 32*(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

Wu, Z., Pan, S., Long, G., Jiang, J., & Zhang, C. (2019). Graph WaveNet for deep spatial-temporal graph modeling. *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, 1907–1913. https://doi.org/10.24963/ijcai.2019/264

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2019). How powerful are graph neural networks? *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1810.00826

Yang, C., Liu, J., Yan, Y., & Shi, C. (2024). FairSIN: Achieving fairness in graph neural networks through sensitive information neutralization. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(8), 9241–9249. https://doi.org/10.1609/aaai.v38i8.28776

Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. *Advances in Neural Information Processing Systems, 32*, 9244–9255. https://arxiv.org/abs/1903.03894

Yu, B., Yin, H., & Zhu, Z. (2018). Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 3634–3640. https://doi.org/10.24963/ijcai.2018/505

Yuan, H., Yu, H., Gui, S., & Ji, S. (2023). Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(5), 5782–5799. https://doi.org/10.1109/TPAMI.2022.3204236

Yuan, H., Yu, H., Wang, J., Li, K., & Ji, S. (2021). On explainability of graph neural networks via subgraph explorations. *Proceedings of the 38th International*

Conference on Machine Learning (ICML), PMLR 139, 12241–12252. https://proceedings.mlr.press/v139/yuan21c.html

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. *Proceedings of the 30th International Conference on Machine Learning (ICML)*, PMLR 28, 325–333. https://proceedings.mlr.press/v28/zemel13.html

Zhao, S., Shao, W., Chan, J., Xu, Z., & Salim, F. (2025). FairDRL-ST: Disentangled representation learning for fair spatio-temporal mobility prediction. *arXiv preprint arXiv:2508.07518*. https://arxiv.org/abs/2508.07518

Zheng, C., Fan, X., Wang, C., & Qi, J. (2020). GMAN: A graph multi-attention network for traffic prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(1), 1234–1241. https://doi.org/10.1609/aaai.v34i01.5477s