IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Digital Object Identifier None

FairGavel: Leveraging Graph Neural Networks for Fairness-Driven Legal Judgment Prediction

MANASA A[1], VARUN ANBALAGAN[1], VIGNESSH P[1] AND RAMAPRABHA K P[1]
[1]School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology, Chennai 600127, India
Corresponding author: Varun Anbalagan(e-mail: varun.anbalagan2021@vitstudent.ac.in).

ABSTRACT The application of Artificial Intelligence (AI) in the legal field has provided new means of automating intricate decision-making processes. However, the problem of maintaining fairness and transparency in AI-based legal judgment prediction remains largely unsolved. This paper presents a GNN-Based Fairness-Driven Legal Judgment Prediction System that uses Graph Neural Networks (GNNs) to represent and interpret judicial cases based on the MultiLexSum dataset. The new system advances from standard text classification with the addition of structural relationships between legal entities, providing a more contextual and holistic interpretation of legal documents. In an effort to address bias and achieve fair outcomes, the system incorporates fairness-aware loss functions as well as adversarial training methods, aiming for protected attributes during model optimization. Large-scale experimentation shows that the model not only has competitive prediction performance but also demonstrates lower fairness leakage on various sensitive attributes. In addition, this work makes a contribution to the larger discussion of responsible AI in legal tech by showing how explainability and fairness can co-exist within high-performing machine learning models. The results of current research emphasize the promise of graph-based solutions in the construction of equitable, responsible, and explainable AI systems for practical legal contexts.

INDEX TERMS Fairness AI,Graph Neural Networks, Verdict, Monetary, Non-monetary, GraphSAGE

I. INTRODUCTION

Over the last few years, artificial intelligence has advanced significantly in changing the manner in which legal systems process, analyze and interpret judicial data. With the continuously increasing number of legal files, there has been a growing need for automatic tools that help predict judgments, summarize laws, and analyze the outcome of cases. Of these, prediction of legal judgment has become a pivotal area of research, seeking to deliver fact-based insights on case outcomes by analyzing past precedents and text-based case facts.

But, while predictive models have shown impressive performance in numerous fields, their application within legal systems raises particular concerns, most importantly, fairness and transparency. Legal rulings have far-reaching social implications, and any bias that might be built into a model has the potential to perpetuate existing disparities or bias outcomes against particular groups. Current models tend to isolate three-way text instances of legal cases without regard for the interconnectedness of legal arguments, parties, and rulings, which are an essential element for capturing the complete context of a case.

The rapid advancement of Artificial Intelligence (AI) and Machine Learning (ML) has led to their increasing adoption in the legal domain, particularly in automating the prediction of legal judgments. However, traditional models primarily focus on textual analysis of case documents without considering the intricate relationships between legal entities such as plaintiffs, defendants, laws, and prior judgments. This results in limited contextual understanding and reduced interpretability, often leading to predictions that lack depth and accuracy.

During the course of development for the Fairness-Driven Legal Judgment Prediction System Based on GNN, a number of key challenges were faced which needed to be addressed through careful consideration, iterative testing, and technical creativity. One of the primary challenges was addressing the biases inherent in the historical legal data. As legal data sets are constructed using past judgments, they tend to have systemic biases such as socioeconomic or gender inequalities. Training a model on such data without adequate mitigation

could reinforce such biases, making predictions unfair. This problem required a deep dive into fairness measures and the use of fairness constraints, such as demographic parity and equal opportunity, during learning to provide equal outcomes for different subgroups.

The main goal of this research is to build a GNN-Based Fairness-Driven Legal Judgment Prediction (LJP) system that not only makes accurate judicial predictions but also fair and interpretable ones. Existing LJP models have been found to be quite promising in helping legal experts by making automated predictions on the outcome of cases given factual and legal stories. However, such models tend to carry biases present in past legal data, resulting in unjust or uneven decisions, especially against vulnerable communities. This paper suggests a framework that uses Graph Neural Networks (GNNs) to represent legal cases as graphs, and learns interdependence among facts, legal entities, statutes, and judgments. By building semantically rich graphs for every case, the system can learn relational patterns that effectively impact the judicial rulings. In addition, the system uses attention mechanisms to improve interpretability and incorporates fairness-aware constraints in training to limit discriminatory patterns.

A systematic feature extraction mechanism is utilized to extract 31 features for plaintiffs and defendants so that a more detailed and well-balanced representation of each side in a case can be achieved. The mission of this project is to bring AI innovation in the legal sense in accordance with ethical duty by guaranteeing transparency, responsibility, and impartiality in choice. Motivation for this work is due to overwhelming real-world evidence of bias-related misbehavior in the judicial system, especially in the United States. Independent investigations and institutional reports have established that official misconduct, prosecutorial bias, racial discrimination, ineffective counsel, and judicial prejudice are common factors in a large number of convictions overturned. A wide-ranging review of more than 2,400 exonerations found that 54% had some kind of official misconduct involved, prosecutorial misconduct being the cause of 30% wrongful convictions.

In 2023 alone, 77% of the 153 exonerations were attributed to official misconduct, and 84% of the exonerees were individuals of color, specifically black and Hispanic individuals - although these groups comprise a lesser percentage of the US population. These data underscore structural deficiencies that remain inherent in legal institutions, all too often unremedied as a result of inadequate accountability. Furthermore, cases of judicial prejudice, although less often reversed, have also resulted in trail-blazing reversals, such as Caperton v. A.T. Massey Coal Co. and Buck v. Davis. Against this backdrop, the purpose of this research extends well beyond the technical breakthrough. It hopes to advance judicial integrity and social justice by encouraging fairness-focused machine learning methods in legal AI, ensuring that predictive models do not repeat or expand existing injustices but rather enable a more equitable and principle-driven future for judicial decision-making.

This work aims at the development and design of an explainable, fairness-conscious Legal Judgment Prediction (LJP) system based on Graph Neural Networks (GNNs) and the MultiLexSum dataset. The main scope is the conversion of complicated and unstructured legal case documents to structured graph representations so that the model can learn complex relationships among entities such as plaintiffs, defendants, legal charges, statutes, and judgments. The project focuses on incorporating fairness restrictions during the learning process to counteract biases inherent in legal data, especially with respect to race, socioeconomic status, and prosecutorial conduct. The system is designed to produce transparent predictions by utilizing attention mechanisms and interpretable node weighting schemes so that any legal outcome can be mapped back to individual facts and relations in the graph. Furthermore, a crucial element of the study is structured attribute extraction, where 31 characteristics are extracted for both plaintiffs and defendants, resulting in an extensive feature space that encompasses contextual and demographic details relevant to the case.

The research is constrained by a number of primary parameters to make it feasible and focused. The first is the data set used, which is MultiLexSum and contains English-language summaries of legal cases, which is appropriate for NLP-based analysis and graph building. The system is trained and tested specifically on multi-label classification tasks like charge prediction and outcome determination in the jurisdiction and case types covered in the dataset. Although the system purports to counteract bias, it does this on statistical measures of fairness instead of involving real-time legal checks or human-in-the-loop checks. As a result, such fairness is only as much as can be quantified computationally with methods such as demographic parity, equal opportunity, or disparate impact reduction. Additionally, while the GNN-based model is designed to operate with the graph-structured data obtained from the MultiLexSum corpus, it is not yet generalizable to other areas of law (i.e., civil, family, or international law) without further domain-specific training. Finally, the model does not replace judicial discretion or provide legally binding verdicts; it is designed to be a decision support tool that helps legal practitioners better understand case trends, detect possible bias, and help make consistent and evidence-based decisions.

In conclusion, this work seeks to expand the frontier of legal AI by combining graph-based learning with fairness-aware optimization to enable intelligent and fair judgment prediction systems.

II. RELATED WORKS

The article "Exploring Graph Neural Networks for Indian Legal Judgment Prediction" addresses the issue of an uneven judges-to-cases ratio in the Indian justice system, leading to heavy case backlogs. It presents an automated system using graph neural networks (GNNs) to predict legal judgments from factual evidence and past case precedents. The study

points out the strong predictive power of the model with a macro F1 score of 75% on the Legal Judgment Prediction (LJP) task and showing more than 80% ROC in link prediction tasks, signifying its effectiveness in learning about relationships in legal settings. One of the major strengths of the model is that it takes into account fairness, focusing directly on gender and name-related biases, which is important to achieve parity in judicial proceedings. The research admits limitations like negative effect of incorporating time nodes on performance of the model, ascertaining whether temporal considerations are relevant in the decision-making of judges. In addition, the graph structure in GNNs makes it a problem for interpretability because it might be difficult to grasp how the predictions were made, potentially influencing the trustworthiness and transparency of AI-deduced legal judgments. Further, the paper lacks in-depth analysis of the explainability of the model's judgments, which is a possible gap in achieving the integrity of automated legal predictions. In general, this work offers a noteworthy contribution to legal NLP, demonstrating the potential of GNNs for improving legal judgment prediction while also informing key factors for future research [4].The Legal Judgment Prediction (LJP) task has gained increasing attention in the area of AI and law, seeking to predict pertinent legal articles, charges, and terms of penalty from factual case descriptions. The conventional methods have tackled LJP either by employing single-task classification or multi-task paradigms, but generally ignoring the inherent interrelations between different judgment components. Early methods were dependent on manually constructed features and rudimentary classifiers, whereas newer methods have used attention mechanisms, memory networks, and capsule models for enhanced performance, especially in addressing confusing and few-shot labels. Multi-task learning models like TopJudge and MPBFN-WCA tried modeling task dependencies by using directed acyclic graphs, while methods like LADAN concentrated on improved representation learning through the resolution of label confusion through graph-based methodologies. Despite the above-mentioned models, the ability of such models to provide logical consistency on tasks still is not present. To compensate, the authors introduce a new method, R-former, and redefine LJP as node classification on an underlying global consistency graph. Both inter- and intra-task relationships among labels are embedded in the graph. R-former utilizes a masked transformer network for acquiring task-consistent and discriminative node representations, and a graph convolutional network for transmitting label information locally. This combination aids in ensuring global and local consistency of predictions. R-former performs better on benchmark data such as CAIL-small and CAIL-big compared to previous models, as it comes with an impressive F1 score boost of around 4.8%. In addition, the paper proposes a new self-consistency evaluation metric using edge precision and recall, which offers a more stringent measure of prediction validity. Although it is an improvement, the complexity of the model and its reliance on a well-organized graph constructed from the training data could influence its generalizability and computational cost. Nevertheless, R-former is a major improvement in creating more coherent and trustworthy LJP systems [5].The paper "Legal Judgement Prediction for UK Courts" is concerned with developing an interpretable machine learning model that can predict legal case outcomes in the UK from court case documents alone. This is an original effort in the deployment of Legal Judgment Prediction (LJP) within the UK legal context, as opposed to previous work, which had primarily concerned itself with jurisdictions such as the European Union, France, and China. A main challenge confronted in this research was the non-existence of a labeled, organized dataset of UK court judgments. The authors addressed this by building their own labeled corpus of more than 4,900 UK court cases and employing a range of text classification methods, such as n-gram vectorization (count and TFIDF), topic modeling (LDA), and word embeddings (Doc2Vec), coupled with machine learning algorithms like Logistic Regression, Random Forests, and neural networks.

Research has a number of significant benefits. First, it shows that high prediction accuracy is achievable in the UK legal scenario, with the top performing model (Logistic Regression + TFIDF) scoring 69.02% for F1. Second, the research prioritizes interpretability by pulling out and analyzing significant text features, making the output of the model easier to interpret for legal practitioners. The work also adds a reusable method for creating datasets and tuning models, and it offers real-world advice on the usability of various text representation methods in legal texts.

Even so, the paper has major limitations, too. A major limitation is that the relatively small dataset size compared to other jurisdictions diminishes generalizability and learning depth, particularly for neural network models. Furthermore, although word embeddings were anticipated to surpass more rudimentary vector-based methods, their performance was modest, most probably caused by a lack of domain-specific training data. The article further observes that although simple neural networks such as SLP and MLP were tested, more advanced structures (e.g., CNNs, RNNs, and HANs) could provide better results. In addition, while the models were reasonably accurate, the black-box nature of deep learning architectures continues to present challenges with respect to interpretability and legal acceptability [6].

The "FairLex: A Multilingual Benchmark for Evaluating Fairness in Legal Text Processing" paper presents an exhaustive benchmark suite used to evaluate fairness in legal NLP systems. It covers four jurisdictions, Europe (ECtHR), the United States (SCOTUS), Switzerland (FSCS), and China (CAIL), and five languages, providing a comprehensive evaluation across legal systems and cultures. The study aims mainly to analyze the performance differences of legal judgment prediction models for different sensitive attributes such as sex, age, place of origin, area of law, and language. The authors point out that legal AI requires fairness, particularly because historical legal data tend to be unfair, and

models learned from them can unwittingly perpetuate social injustices. In this pursuit, they assess pre-trained transformer models of language based on fairness-sensitive optimization strategies such as Group DRO (Distributionally Robust Optimization), IRM (Invariant Risk Minimization), V-REx (Risk Extrapolation), and Adversarial Removal.

Another significant benefit of the FairLex benchmark lies in its multilingual and multijurisdictional framework, which enables researchers to experiment with the robustness and fairness of models in various legal environments. The benchmark also offers quantitative measures in the form of macro-F1 per group, group disparity (standard deviation among groups), and worst group performance, providing an explicit and quantifiable method to assess fairness. Moreover, the models, data, and code are all open-source and made available publicly on Hugging Face and GitHub, promoting reproducibility as well as research collaboration. A second area of strength is its realistic experiment conditions under group and label imbalance and temporal shift, conditions likely to be seen in actual-world deployments.

However, a limitation of the benchmark is that it deals only with a handful of jurisdictions and sensitive attributes and that certain categories, such as gender, are oversimplified to binary labels by relying on heuristic extraction processes. The incorporation of judicial decisions as ground truth is subjective because legal decisions can be biased in themselves. Furthermore, although the fairness-inducing algorithms did exhibit improvement in certain instances, none outperformed regular training practices (ERM) in all datasets and attributes, reflecting the challenge of attaining fairness in diverse legal data. A second limitation is that the research does not include graph-based models like Graph Neural Networks (GNNs), which are particularly adept at representing the relational structure of legal information and potentially provide competing fairness solutions.

Within the fairness-driven prediction of legal judgment with GNNs, this research offers a solid basis for testing fairness, but opens the door to innovation. By combining graph-based representation and fairness-conscious training objectives, future research can advance on FairLex to establish if GNNs have the potential to provide enhanced fairness and interpretability within legal AI applications [7].

The paper titled "Transductive Legal Judgment Prediction Combining BERT Embeddings with Delaunay-Based GNNs" presents a new approach to Legal Judgment Prediction (LJP) by combining BERT embeddings with a Delaunay-based Graph Neural Network (GNN) under a transductive learning setup. Unlike traditional inductive models that treat each legal document in isolation, this method represents the entire document corpus as a graph, where nodes correspond to documents, and edges are drawn using Delaunay triangulation based on semantic similarity. This structure enables context-aware classification and allows label propagation across both training and test data, leading to significant improvements in prediction accuracy. Evaluated on the Swiss Judgment Prediction (SJP) dataset, the model surpasses robust baselines such as Hierarchical BERT, XLM-R in cross-lingual transfer, and even domain-specific large language models like SaulLM-7B.

The primary strengths of the method include high performance with minimal computational complexity, robustness in low-resource scenarios like the Italian subset of SJP, and a modular two-stage training process. It also avoids common GNN pitfalls like over-smoothing and over-squashing through dimensionality-aware graph construction with UMAP and Delaunay triangulation. However, the paper is limited by its focus on only the SJP dataset, raising concerns about generalizability between jurisdictions. It also does not evaluate fairness or bias—an essential consideration in legal AI—and may face scalability issues due to the need for graph recomputation with new documents. Despite these, the work highlights how lightweight, transductive graph reasoning can offer a competitive alternative to large-scale models by leveraging inter-document relationships critical to the legal domain [8].

The article "Distinguish Confusing Law Articles for Legal Judgment Prediction" introduces an end-to-end model named LADAN (Law Article Distillation-Based Attention Network) that improves the accuracy of Legal Judgment Prediction (LJP), especially in cases involving confusing or overlapping law articles. These are provisions that share lexical similarities and are often misclassified by standard LJP models. LADAN addresses this by incorporating a graph neural network-based attention mechanism designed to capture fine-grained distinctions between legal articles. Initially, it constructs a graph of law articles using cosine similarity on their content and clusters them into semantically coherent communities. A novel Graph Distillation Operator (GDO) is proposed to extract the most discriminatory characteristics within each law community. These features help reencode factual descriptions more precisely, enabling the model to perform both macro-level classification (community prediction) and micro-level reasoning (specific law prediction).

The major contribution of LADAN lies in its ability to automatically distinguish legally subtle differences without requiring hand-crafted features or annotations. It achieves state-of-the-art performance on the CAIL2018 dataset (both small and large) and significantly outperforms competitive baselines such as MPBFN and TopJudge in tasks such as article prediction, charge classification, and sentence determination. Additionally, its attention visualizations offer interpretability by highlighting legally relevant keywords (e.g., "public official" vs. "private manager") that influence predictions.

However, LADAN is trained and tested solely on Chinese legal data, raising concerns about its generalizability to multilingual or cross-jurisdictional contexts. Its effectiveness depends on accurate community detection, and any noise in clustering can affect downstream results. In addition, fairness and bias assessments are not considered, and GDO layers may face scalability challenges when the legal code base is very large. However, LADAN presents a powerful and

interpretable method for improving the robustness of LJP systems through fine-grained article differentiation [40] [9].

The article "LA-MGFM: A Legal Judgment Prediction Method via Sememe-Enhanced Graph Neural Networks and Multi-Graph Fusion Mechanism" introduces a significant advancement in Legal Judgment Prediction (LJP), aiming to predict relevant charges, penalties, and legal provisions from unstructured legal texts. Traditional LJP methods relied on statistical or linear models [18], which were insufficient to represent complex legal semantics. Later, attention-based models [17] and hand-made feature approaches [19] tried to capture legal knowledge but suffered from noise and domain dependence. Graph-based solutions [16] improved structural representation but lacked semantic depth. More recent approaches, such as Lawformer [14] and Graph Fusion Networks [15], used GNNs for text modeling.

Building on these, LA-MGFM constructs five different types of graph per case and introduces a Sememe-Enhanced Gated Graph Neural Network (SE-GGNN) to encode rich legal semantics at a granular level. It also uses a Multi-Graph Fusion Mechanism to integrate information from heterogeneous graph views, effectively addressing label confusion in complex charges. Experiments on real-world datasets demonstrate that LA-MGFM outperforms state-of-the-art baselines, including under a few-shot scenarios, confirming its robustness and broad applicability [10].

The paper entitled "Legal Judgment Prediction via Graph Boosting with Constraints" solves major problems in the field of Legal Judgment Prediction (LJP), which aims to predict law articles, charges, and penalties from textual case descriptions. Although earlier models incorporated legal features, leveraged intertask relationships [7], or applied neural networks to constituent elements [9], many still struggled with multi-label dependencies and label ambiguity. GJudge introduces a novel approach using a multi-perspective interactive encoder combined with a multigraph attention consistency expert module. This architecture integrates bidirectional LSTMs, graph attention, and gated mechanisms to model both factual content and complex label interrelations. Moreover, GJudge constructs graphs of legal labels to explicitly encode interdependencies, helping distinguish confusing labels and maintain cross-task consistency. The model achieves notable performance gains on multiple benchmarks and establishes a new baseline in the prediction of multitask, multi-label legal judgments [11].

The research paper "CaseGNN: Graph Neural Networks for Legal Case Retrieval with Text-Attributed Graphs" [2] addresses two fundamental shortcomings in Legal Case Retrieval (LCR): the neglect of structural legal relationships and limitations in processing long-form case texts. Traditional retrieval systems such as TF-IDF, BM25, and LMIR rely heavily on statistical matching, while recent transformer-based models like LEGAL-BERT, SAILER, and Prompt-Case improve semantic understanding but still treat legal documents as unstructured text. CaseGNN introduces a Text-Attributed Case Graph (TACG), generated through named entity recognition and relation extraction, that captures both structural and textual nuances of legal cases. This graph is processed using an Edge Graph Attention Layer (EdgeGAT), and contrastive learning is applied to improve the accuracy of the retrieval using both hard and easy negatives. Experiments on the COLIEE 2022 and 2023 datasets demonstrate that CaseGNN achieves superior retrieval performance over both classic and transformer-based baselines. The framework showcases the benefit of combining graph-based reasoning with neural encodings in legal retrieval [3].

The paper introduces *ML-LJP*, a novel multilaw aware model for Legal Judgment Prediction (LJP), which enhances prediction performance by incorporating both charge-related and term-related law articles, unlike previous models that primarily focused on the former. The authors argue that overlooking term-related laws, such as those governing leniency or recidivism, hampers accurate prison term prediction. ML-LJP formulates law article prediction as a multi-label classification problem, employs attention mechanisms to learn label-specific representations, and integrates contrastive learning to better differentiate between similar legal provisions. For term prediction, the model uses a graph Attention Network (GAT) to capture high-order dependencies among law articles and proposes a number representation module to encode numerical legal attributes such as drug weight or monetary value, which is typically neglected in existing models.

ML-LJP's key strengths include its robust performance across all LJP subtasks, notably a 10. 07% improvement in F1-score for prison term prediction over leading baselines. Its explainability is supported by attention visualizations and label-wise fact encodings. In addition, it performs well on low-frequency labels and mitigates data imbalance, which is a challenge prevalent in legal datasets. The authors publicly release the code and the data set, promoting transparency and reproducibility. However, the model is limited to Chinese legal data (LAIC2021), which can restrict its generalizability to other jurisdictions. Moreover, it lacks fairness analysis, and its number representation module is heuristic and may not be generalized across legal systems. Despite these limitations, ML-LJP is a promising and realistic advance for LJP [13].

The article "Towards Explainability and Fairness in Swiss Judgement Prediction: Benchmarking on a Multilingual Dataset" is concerned with improving the explainability and fairness of Legal Judgement Prediction (LJP) models, specifically using the Swiss Judgement Prediction (SJP) dataset, which is special due to its multilingual nature. It seeks to assess and improve the transparency of LJP models, enabling legal professionals to understand the rationale for model predictions, which is essential for establishing trust in legal settings. The study also examines possible biases in LJP models due to irrelevant or sensitive predictors, tackling key ethical issues of fairness in legal predictions. Drawing on the only multilingual dataset that features the decisions of the Swiss Federal Supreme Court in German, French, and Italian, the research provides conclusions on the potential influence of language on prediction and explanation.

One of the greatest strengths of the paper is its multilingual design, which allows for wider evaluations of LJP in different linguistic settings. In addition, its rigorous emphasis on explainability serves the purpose of transparency in decision making in legal technology, which in turn builds confidence among stakeholders. Finally, the innovative method of analyzing bias in the study, especially with respect to lower court influences, is helpful in reducing unfairness in legal judgments. The development of a new assessment framework (Lower Court Insertion - LCI) also serves the purpose of increased model accountability.

However, the paper has some disadvantages. The multilingual and explainability-driven nature of its complexity can make model development difficult and require many resources to annotate data and involve experts. Additionally, relying on human experts for rationale may bring about subjectivity that causes inconsistencies. Furthermore, according to the findings, improved prediction accuracy does not always translate into better explainability, which may create challenges for practitioners who are interested in consistent legal insight at the expense of performance. Lastly, the imbalanced case distribution across the three languages can affect the balanced insights and performance across the three languages. All in all, the paper greatly contributes to the LJP community while presenting relevant considerations for further research and model development.

III. MOTIVATION AND RESEARCH GAP

Automation of legal judgment prediction (LJP) has significant potential to improve the efficiency and accessibility of legal systems. As there is an ever-increasing number of legal cases, the legal institutions face a growing workload to process the cases quickly and accurately. This has encouraged researchers and practitioners to seek AI-based solutions that can be used in decision support, case analysis, and outcome prediction. Graph neural networks (GNNs), in particular, have become increasingly important because they can represent the relational structure of legal documents and capture complex dependencies between sentences, entities, and legal concepts. With further development, these models are expected to play an essential role in transforming legal systems and making timely recommendations to legal professionals.

However, the use of AI in high-risk fields such as law also poses profound ethical issues. Legal conclusions often have a direct effect on people's lives, liberties, and rights. Biases of any kind—historical data, sensitive characteristics such as gender or socioeconomic status, or model construction—may cause discriminatory predictions that reinforce prevalent societal disparities. In spite of the technical advances in the LJP domain, fairness is yet to be considered a first-class consideration in most GNN-based LJP designs. There is a great incentive to build models that not only retain predictive capability but are also ethically sound, in that no side is prejudiced based on their demographic or sensitive features.

Although recent studies have investigated GNNs for LJP, most current work, e.g. QAjudge, LADAN, TopJudge, and LegalGNN, ignores fairness or implements shallow measures for bias reduction. These models mostly work with document-level nodes or rough abstractions of legal texts, failing to tap into fine-grained semantic, sequential, and contextual cues locked in legal stories. Additionally, fairness mechanisms, where they exist, tend to be limited to post hoc analysis or generic adversarial configurations that fail to distinguish between the distinctive roles of plaintiffs and defendants. This one-size-fits-all approach does not consider asymmetric biases and does not support role-specific fairness optimization.

In addition, there is a clear lack of incorporating fairness considerations directly into the graph construction and learning process. Current models do not properly model sensitive attributes as first-class citizens in the graph, nor do they utilize adversarial methods specifically designed for multi-entity settings. The absence of metadata-driven, heterogeneous graph structures with typed edge labels also limits the ability of such models to incorporate the intricate, multilayered characteristics of legal documents and actor interactions. Therefore, such existing methods might be lacking in delivering legally valid and impartial results in real-life applications.

This is intended to fill this crucial void. It presents a high-fidelity graph structure with sentence- and entity-level nodes, augmented with rich metadata, and linked by a heterogeneous set of labeled edge types encoding both semantic and structural relationships. To address bias directly, it uses a dual-adversarial fairness mechanism, one for the plaintiff and one for the defendant, along with a Gradient Reversal Layer to prevent sensitive attributes from affecting the verdict prediction. This strong integration of graph representation and fairness modeling makes **FairGavel** stand out from previous work and seeks to establish a new benchmark for ethical, understandable, and high-performance legal AI systems.

IV. RESEARCH CONTRIBUTIONS

This paper introduces a new fairness-oriented approach to the prediction of legal judgment (LJP) using graph neural networks that guarantees both precision and moral integrity. The main contributions of this study are listed below.

1) **Fine-Grained Graph Construction for Legal Texts:** We introduce an innovative graph construction method that represents legal case documents at multiple granular levels. The graph includes both sentence-level nodes that are metadata-annotated as Document ID, Paragraph ID, Sentence Position, and Section Headers, as well as entity-level nodes representing plaintiff and defendant. These are consciously linked using multi-type edges based on multiple characteristics, ranging from sequential and semantic similarity to entity-sentence and document-level contextual ones. In this way, our project will be able to capture the structure and meaning connections in the case in an extensive way.

2) **Metadata-Augmented Representation Learning:** Unlike earlier LJP models, which are mostly based on raw text embeddings or basic graph structures, **FairGavel** incorporates structured metadata directly into its graph structure. This allows for a more comprehensive grasp of contextual dependencies, legal actor participation, and document structure, enhancing both model expressiveness and interpretability.
3) **Dual Adversarial Fairness Mechanism:** In order to target bias in LJP, we propose a dual adversarial learning framework with distinct fairness heads for the plaintiff and the defendant. These adversarial modules are bridged to the primary prediction branch through Gradient Reversal Layers, which train the model to acquire verdict representations that are insensitive to each legal actor's sensitive attributes (e.g., gender, name, role-based bias). This dual strategy enables targeted and role-oriented fairness mitigation, which has remained unaddressed by existing GNN-based LJP models.
4) **Fairness-Aware Graph Representation and Learning Integrated:** **FairGavel** deeply intertwines fairness goals with the graph-based learning process. Instead of addressing fairness as an exogenous constraint or after-processing technique, our model incorporates fairness issues directly into the representation learning process so that verdict predictions are not only effective but also resilient against sensitive attribute leakages.
5) **Scalable and Flexible GNN Architecture:** **FairGavel** utilizes a GraphSAGE-based encoder for scale and inductiveness in representation learning and then applies a Multi-Layer Attention Pooling (MLAP) strategy to extract an overall graph-level representation. This design enables the model to generalize from new legal cases and scale up with large datasets while retaining interpretability and fairness.
6) **Comprehensive Comparative Evaluation with State-of-the-Art Models:** We conduct a thorough comparison of this research with a set of current GNN-based LJP models, such as QAjudge, LADAN, CaseGNN, and LegalGNN. Our experiments prove and attain comparable or better performance in verdict prediction while considerably limiting fairness leakage and enhancing prediction robustness over sensitive attribute groups.

V. PROPOSED SYSTEM

The methodology of the proposed system can be visualized using Figure 1, is a modular and fairness-focused pipeline for the prediction of legal judgments with graph-based learning. The goal is to develop a system that makes judicial predictions by representing legal cases as structured graphs, deriving contextual features from legal documents, and using graph neural networks (GNNs) for classification, all while integrating fairness constraints to reduce biased predictions.

A. DATASET COLLECTION

The basis of this study is the use of the MultiLexSum dataset, a large-scale multilingual legal corpus created by Chalkidis et al. for legal summarization and judgment prediction. Drawn mainly from the EUR-Lex legal database, MultiLexSum has more than 100,000 case documents from European jurisdictions. Every case has elaborate components like factual background, relevant legal statutes, judgment outcomes (usually multilabel), and summarized rationales for the court's decisions. To ensure consistency in analysis and preprocessing, this work only dealt with English-language cases and kept only records with full information in all important fields. The overview of the data set is shown in Table 1

TABLE 1. Dataset Fields Overview

Field Name	Data Type	Description
case_id	String	Unique identifier assigned to each legal case.
sources	List[String]	Collection of source documents pertaining to the legal case.
summary/long	String	Detailed multi-paragraph summary (650 words) of the case.
summary/short	String	Concise single-paragraph summary (130 words) of the case.
summary/tiny	String	Brief one-sentence summary (25 words) encapsulating the essence of the case.

Every record in the MultiLexSum dataset is well structured, including a unique id, source documents (derived from federal court PDFs) and three levels of human-sourced summaries—long (~650 words), short (~130 words), and tiny (~25 words). These summaries are written and confirmed by professional legal specialists under strict quality assurance to provide semantic depth and factuality. Once the data set was pulled from the official GitHub repository, a stringent filtering pipeline was implemented to remove duplicates, incomplete records, and irrelevant files, producing a cleaned-up subset of around 35,000 high-quality legal case records that were used for training and testing in this work. Table 2 contains multiple numeric columns and is organized to neatly present data set splits along with the train, validation, and test percentages.

B. PREPROCESSING

Due to the natural complexity and wordiness of legal documents, pre-processing is a critical process of pre-preparing the data for accurate and efficient prediction of legal judgment. For the system to capture the most useful content from extensive case files, the preprocessing pipeline starts with extractive summarization that identifies and extracts the most important sentences in terms of importance and context contribution. This facilitates the shortening of the document

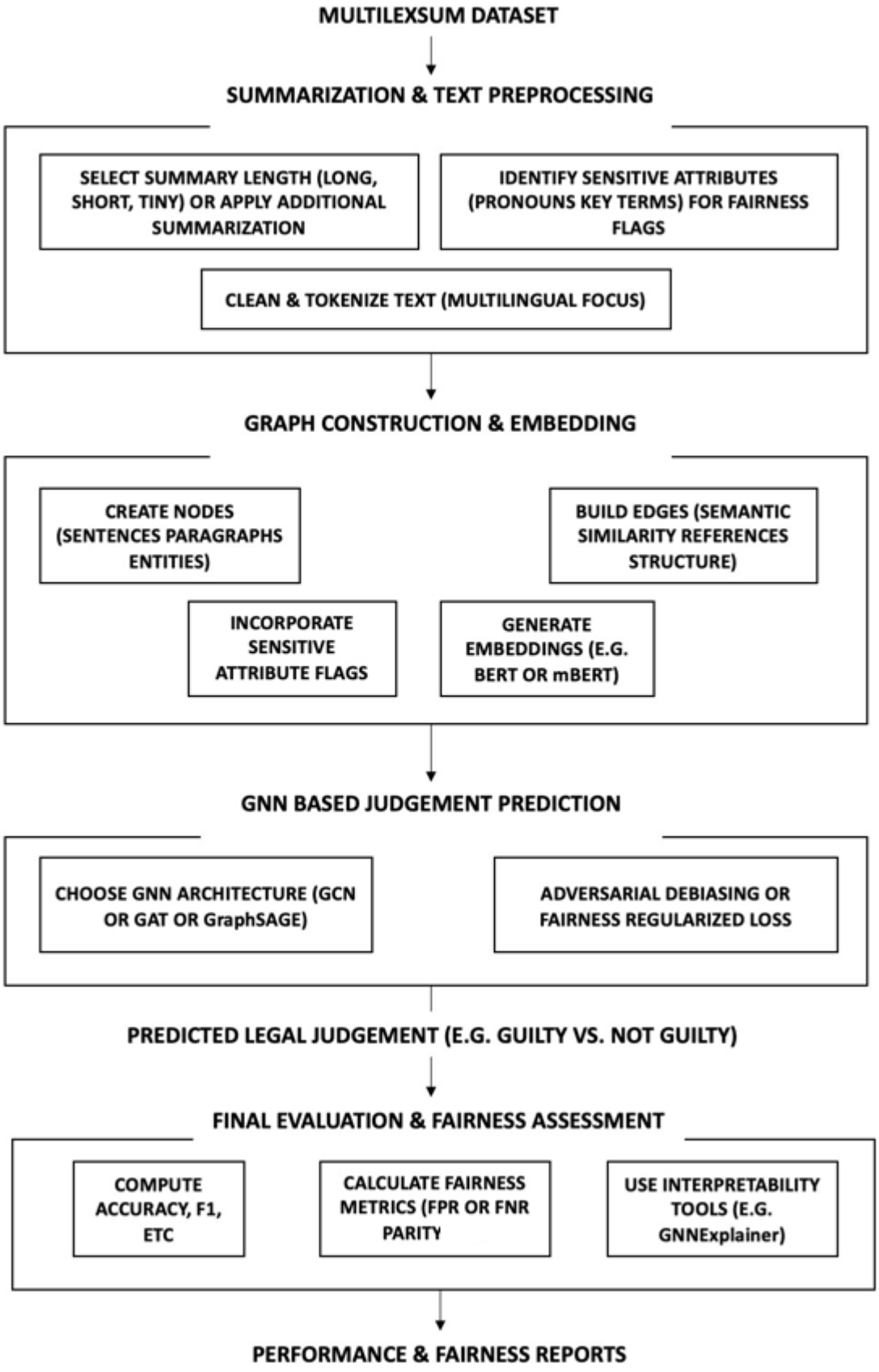

FIGURE 1. Proposed System Diagram.

TABLE 2. Dataset Splits and Counts

Split	Number of Cases	Number of Source Documents	Number of Summaries	Train-Val-Test Percentage
Training	2,199	19,780	4,511	61.76
Validation	454	4,134	927	12.75
Test	908	7,428	1,836	25.49
Total	4,539	40,119	9,280	100

without losing its essential meaning. To further assist in understanding, abstractive summarization methods are used in parallel to produce brief, reworded summaries that capture the substance of the case without depending on sentence-by-sentence extraction.

TABLE 3. Summary Statistics for Different Summary Types

Summary Type	Avg Char Length	Avg ROUGE Score	Avg BLEU Score
Extractive	1541.35		
Abstractive	324.14		
Combined		0.3044	0.0745

In an effort to cope with the magnitude of the dataset, which comprises tens of thousands of court cases, batch processing and parallelization are utilized to enhance efficiency. The data is divided into workable pieces and processed in parallel on multiple threads or processors. This significantly shortens preprocessing time while ensuring consistency throughout the entire dataset. Along with this, text cleaning and tokenization are done: reducing text, removing unnecessary characters, normalizing spacing, and splitting sentences into individual tokens. These processes normalize the data and remove noise, providing a foundation for proper text embedding and analysis.

Moreover, augmented techniques are used to enrich the data set. Each case is augmented with cleaned and tokenized forms of its source text and summary. Language detection is also added to mark the main language of each entry to ensure proper handling during multilingual processing. Lastly, the data set is subjected to column management in order to remove unwanted fields and preserve only the required identifiers and preprocessed text. This leaves us with a clean, organized data set optimized for training our fairness-conscious GNN-based legal judgment model, maintaining both semantic purity and computational efficacy.

C. ENTITY AND ATTRIBUTE EXTRACTION

To support the prediction of fairness-aware legal judgments, this work focuses on extracting structured representations of the defendant and plaintiff from each case file. A hybrid approach was used that blended summarization-based filtering and language model-based entity extraction to provide both factually accurate and contextually rich representations. Central to this process is a zero-shot prompt-based method utilizing ChatGPT, where each legal case is queried with a custom prompt that asks the model to return entity details in a strict JSON format. This schema group attributes into general attributes (e.g., country, case type), individual attributes (e.g., age, gender, occupation) and organization attributes (e.g., sector, ownership type). Missing or inferable attributes are returned as "null," ensuring transparency and standardization in data completeness.

Recognizing the incomplete nature of many legal documents, the pipeline incorporates a two-pass extraction approach. The first pass is performed in an off-line interpretation-only mode based on the GPT-4o-mini model to extract entities strictly from the case content provided. This ensures that the extracted data identify only what is present in the dataset. All replies are parsed, checked, and stored in bulk to prevent repetition and maintain forward progress. The second pass sees a search-enabled model version (gpt-4o-mini-search-preview) utilized to fill out attributes that had been left "null" on the initial run. The system can draw upon publicly known data to make attributes more complete, especially when the entities repeat like government entities or organizations. A specialized utility function also monitors and consolidates missing fields to allow for subsequent audits or adjustments.

The word cloud in Figure 2 represents the most common words in legal case filings, the prominent terms being "document," "filed" and "attorney," which represent the legal processes and actors involved. Other common words such as "district court," "united states," and "motion" represent the emphasis on judicial processes and legal action. Words such as "plaintiff," "exhibit," and "summary judgment" place emphasis on specific legal actions and actors. This illustration gives an idea of the themes that repeat throughout the dataset, revolving around legal proceedings and case details.

After extracting and enriching the entity profiles, the next important step is to standardize and merge them throughout the data set. Legal actors tend to occur in numerous cases with different or incomplete attribute data, so uniform representation is necessary. To do this, each entity's hierarchical structure is flattened into a one-level format, combining all available attributes into a unified profile. Standardization processes are invoked to address missing or undefined values so that all entries are treated consistently. Duplicate instances of the same participant are subsequently merged into a unified data set of legal actors through a merging process that

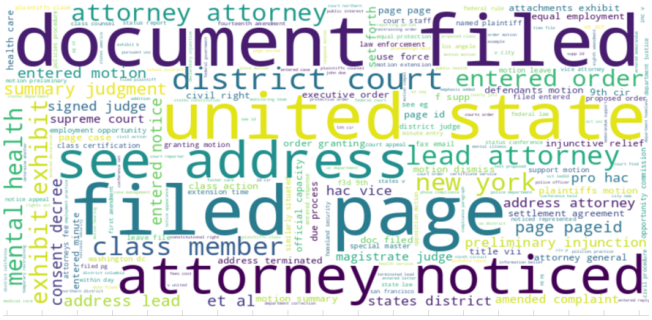

FIGURE 2. Word Cloud From Combined Source

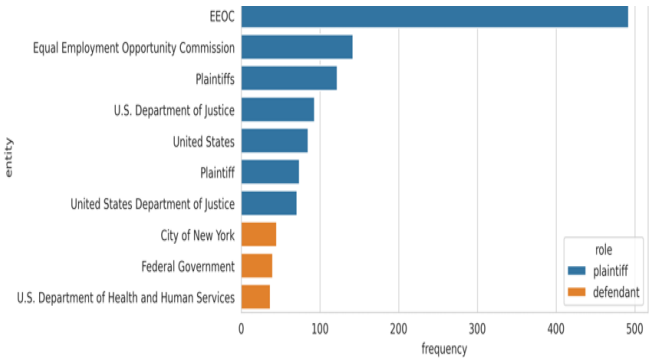

FIGURE 3. Top 10 Entities

preserves the most accurate and comprehensive information available. To facilitate fairness-aware modeling, a carefully selected set of demographic and organizational characteristics is maintained, spanning attributes such as gender, race, urbanicity, education, disability status, net worth, and stock listing status. These augmented profiles serve as the basis for the construction of up-sell graphs and fairness-sensitive learning.

Figure 3 illustrates the top 10 entities in legal cases, separated by their function as plaintiffs (blue) and defendants (orange). The most recurrent entity is the Equal Employment Opportunity Commission (EEOC), with much more participation as a plaintiff than any other entity. Other highly recurring entities are the US Department of Justice and Plaintiffs, each appearing more than once in the dataset. The graph indicates the frequency with which these parties appear in legal cases, revealing information about who are the typical participants in legal cases, specifically in the capacity of plaintiffs. The fact that defendants such as the Federal Government and the City of New York feature prominently shows that there is a wide variety of entities participating in such cases.

D. VERDICT CALCULATION

Judicial judgments, especially in civil rights cases, often go beyond money penalties to include non-monetary outcomes such as injunctive relief, organizational change, and discipline mandates. To represent such intricate verdicts in a machine learning accessible format, we formulated a single scoring approach that converts qualitative and quantitative results into a scalar verdict score in the interval [0, 1]. This score indicates the balance of legal victory between the plaintiff and defendant, which facilitates regression-based training, fairness tests, and interpretation of results. The approach has several components: semantic categorization of non-monetary outcomes, value extraction and normalization of monetary values, and ultimate score calculation based on a weighted combination of the above, complemented by contextual information abstracted from the case summary.

For non-monetary judgments, we had a collection of 16 prototype verdict sentences each from a category of frequently occurring remedies. For each of these, sentence embeddings are obtained by applying LegalBERT (nlpaueb/legal-bert-base-uncased) to the prototypes and to the judgment parts of each case. We calculate cosine similarity scores between these, and when any part of the judgment exceeds a similarity threshold with a prototype, the resultant non-monetary category is tagged. The average similarity per defendant or plaintiff is calculated as follows: Let N_p and N_d denote the non-monetary impact scores for the plaintiff and defendant, respectively. These are computed using cosine similarity between judgment text and prototype embeddings as represented in Equation(1) and Equation(2):

$$N_p = \frac{1}{k}\sum_{i=1}^{k} \text{sim}(s_i, \text{Prototype}_i^P) \tag{1}$$

$$N_d = \frac{1}{k}\sum_{i=1}^{k} \text{sim}(s_i, \text{Prototype}_i^D) \tag{2}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity and k is the number of prototype categories matched. In parallel, monetary outcomes are extracted using Named Entity Recognition (NER) and contextual analysis. Only financial values related to verdicts (e.g. damages, settlements) are retained, excluding irrelevant figures like legal fees. These are assigned to the corresponding party based on syntactic and semantic cues, and normalized using a logarithmic transformation to reduce skewness caused by extreme values. This is represented in Equation(3) and Equation(4) :

$$M_p = \log\left(\max(1, \text{Monetary}_{plaintiff})\right), \tag{3}$$

$$M_d = \log\left(\max(1, \text{Monetary}_{defendant})\right) \tag{4}$$

The total party-wise outcome scores are then computed as weighted combinations of monetary and non-monetary scores:

$$T_p = \alpha M_p + (1-\alpha) N_p \tag{5}$$

$$T_d = \alpha M_d + (1-\alpha) N_d \tag{6}$$

where $\alpha \in [0, 1]$ is a tunable hyperparameter that controls the influence of monetary outcomes.

We compute the preliminary verdict score as:

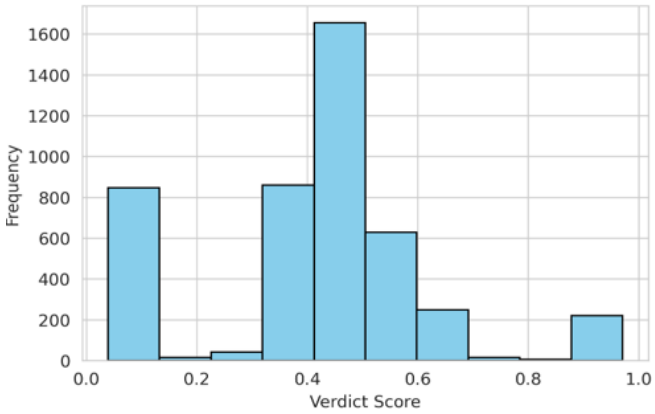

FIGURE 4. Histogram of Verdict Score

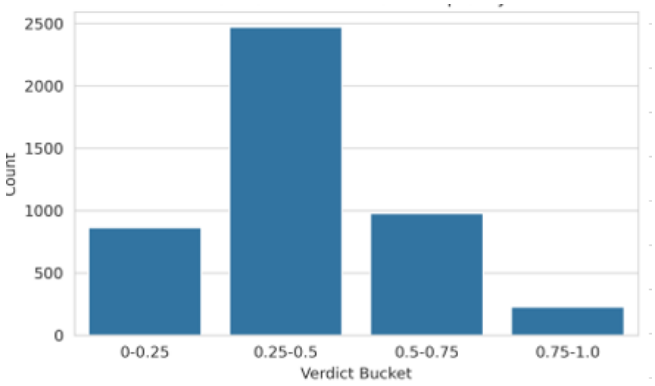

FIGURE 5. Verdict Score Bucket Frequency

$$V_{\text{computed}} = \frac{T_d}{T_p + T_d} \tag{7}$$

A value of 0 represents a complete plaintiff win, and 1 represents a complete defendant win. A contextual outcome signal $O \in [0, 1]$ is extracted from the long-form case summary. The final verdict score is computed as

$$V_{\text{final}} = (1 - \gamma) \cdot V_{\text{computed}} + \gamma \cdot O \tag{8}$$

Here, $\gamma \in [0, 1]$ is a blending factor used to balance computed and contextual signals.

TABLE 4. Statistics for Train, Validation, and Test Splits

Split	Count	Mean	Std	Min	Median	Max
Train	3177	0.4169	0.2089	0.0397	0.4758	0.9712
Validation	454	0.4150	0.2037	0.0403	0.4767	0.9618
Test	908	0.4136	0.2085	0.0394	0.4727	0.9594

The histogram in Figure 4 shows the distribution of Verdict Scores, which are between 0 and 1, with scores near 0 indicating a stronger plaintiff win and scores near 1 indicating a stronger defendant win. Most of the verdict scores cluster around 0.4, with a few cases spread out over the other values. This indicates a bias towards balanced results in the dataset, leaning slightly in favor of the plaintiff. The spread shows varied case results, reflecting a spectrum of judicial rulings. The Figure 5 bar chart indicates the frequency distribution of Verdict Scores that are categorized into four buckets: 0-0.25, 0.25-0.5, 0.5-0.75, and 0.75-1.0. The highest frequency occurs in the bucket 0.25-0.5 with more than 2,000 cases, which indicates that the majority of legal cases in the data set fall in this category of verdict scores. The remaining buckets have relatively lower frequencies with the lowest occurring in the bucket 0.75-1.0. This distribution suggests that legal decisions would group in the middle range of the verdict score scale with less extreme outcomes.

This final scalar is used as the model's target during training and supports more interpretable outcome modeling. Lastly, to enable fairness analysis, the system groups extracted entity attributes into common, individual-specific, and organization-specific categories. This structured schema ensures that both people (e.g., individuals filing lawsuits) and institutions (e.g., private firms or government agencies) are represented consistently, allowing for granular bias auditing and fairness-aware learning across different demographic and organizational factors.

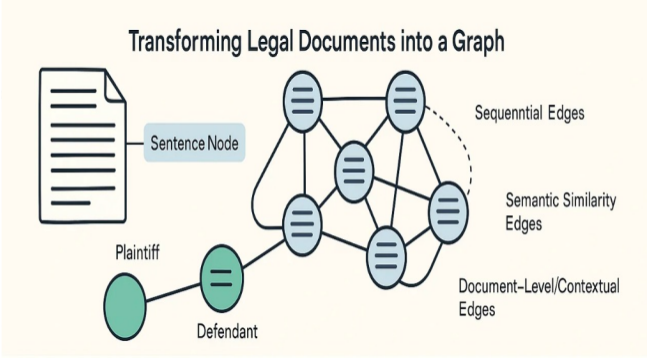

FIGURE 6. Graph construction.

E. GRAPH CONSTRUCTION

Legal documents often consist of lengthy, structured text, spanning multiple paragraphs and sections. **FairGavel** transforms each legal case into a heterogeneous graph that captures both the content and structural nuances of the document. The graph is composed of:

Sentence Nodes: Each sentence is treated as a node, embedded using textual representations (e.g., BERT-based embeddings) and enriched with metadata such as Document ID, Paragraph ID, Sentence Position, and Section Headers. This preserves both semantic content and positional context.

Entity Nodes (Plaintiff and Defendant): Legal actors are represented as individual nodes, each carrying role-specific metadata and non-optional sensitive attributes (e.g., region, language, nationality). This allows the model to account for participant-specific information during learning. This is illustrated in Figure 6.

1) Edge construction

The system incorporates multiple edge types to capture relationships across and within documents:

Sequential Edges: Connects each sentence to the next, preserving the natural flow of the narrative.

Semantic Similarity Edges: Added between sentences with high semantic similarity, enabling the model to pick up on latent themes or argumentative coherence.

Entity-Sentence Edges: Link entity nodes to all sentence nodes where the entity is mentioned, establishing explicit relationships between actors and their described actions or claims.

Document-Level Edges: Implicit edges connect all sentence nodes within the same document, helping the model understand broader document context.

Self-Loops: Each node is connected to itself to retain intrinsic features during message propagation.

All edges are annotated with type-specific labels (e.g., sequential, semantic, entity-sentence), allowing the GNN to differentiate between types of relationships during training.

F. FAIRGAVEL MODEL ARCHITECTURE

The **FairGavel** model is the heart of the system, comprising three primary elements: a graph encoder, a verdict regressor, and a fairness-aware adversarial module, as illustrated in Fig. reffig:The "**FairGavel**" Model.. These elements collaborate to learn graph-level representations, make case predictions, and maintain fairness via adversarial debiasing.

GraphSAGE Encoder: The encoder utilizes two stacked `SAGEConv` layers, reducing node features from 384 to 128 dimensions and then keeping 128 dimensions in the second layer. Mean aggregation is utilized in message passing to calculate contextualized node representations. A multi-layer gated attention pooling mechanism subsequently gives precedence to informative sentence and entity nodes while calculating the final graph-level embedding.

Verdict Predictor: A two-layer feedforward network takes the pooled graph representation and outputs a verdict score in the range [0, 1]. A score of 0 denotes a complete win for the plaintiff, 1 represents a full win for the defendant, and intermediate values capture partial or mixed outcomes.

Adversarial Fairness Module: Two separate adversarial heads are employed—one for the plaintiff and one for the defendant. Both heads have multiple attribute-specific classifiers that predict sensitive attributes like region, language, and urbanicity. These predictions are not used for downstream but for steering the encoder to learn fair, attribute-invariant representations.

After the message-passing step, node-level embeddings are pooled via a gated attention pooling operation. Each node embedding is fed into a learnable gating function, generating a scalar attention score. These scores are utilized to calculate a weighted sum of all node embeddings, generating the final graph-level vector. This process allows nodes with essential verdict-related information or legal actor references to more significantly contribute to the case representation, improving prediction accuracy and interpretability. This is diagrammatically represented in Figure 7.

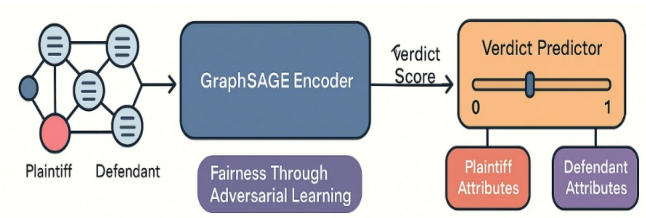

FIGURE 7. The "**FairGavel**" Model.

To impose fairness in training, a Gradient Reversal Layer (GRL) is inserted between the graph encoder and every adversarial head. The layer acts as an identity in the forward pass but reverses the gradient during backpropagation. As a result, the encoder learns to generate feature representations that mask sensitive attributes, thereby making it harder for the adversarial heads to deduce them. This adversarial configuration induces the model to learn fair and unbiased embeddings.

The training approach for **FairGavel** is defined as a multi-task learning task. The main task is verdict regression, with the secondary tasks being adversarial prediction of sensitive attributes. Adversarial heads are modular and organized by attribute type. Cross-entropy loss is used with classification-based heads, and mean squared error with regression-based heads (e.g., age, years in business). This architecture not only facilitates extensibility but also better enables the model to generalize over multiple protected attribute groups.

G. HYPERPARAMETER CONFIGURATION

1) Model Structure and Training Hyperparameters

The key model configuration parameters employed within this work are summarized in Table 5. As can be seen, 384 input channels, 128 hidden channels, and 2 layers are used by the model. Neighbor sampling is performed using [25, 10] neighbors, with 6 subgraphs being applied. The model is trained to 20 epochs using a learning rate of 0.001, weight decay of 1×10^{-5}, and a batch size of 32.

TABLE 5. Model Structure and Training Hyperparameters

Parameter	Value
Input Channels (`in_channels`)	384
Hidden Channels (`hidden_channels`)	128
Number of Layers (`num_layers`)	2
Number of Neighbors (`num_neighbors`)	[25, 10]
Number of Subgraphs (`num_subgraphs`)	6
Learning Rate	0.001
Weight Decay	1×10^{-5}
Number of Epochs	20
Neighbor Batch Size	32

2) Sensitive Attribute Encoders and Format

The attributes are categorized as common, individual, and organizational and label encoders are used accordingly. Every

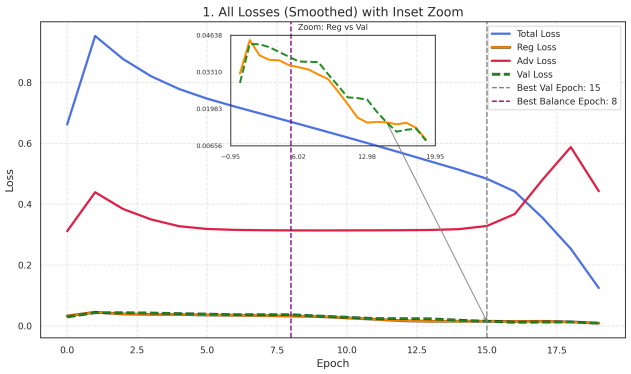

FIGURE 8. Training and validation loss curves.

attribute type (e.g., gender, race, language) comes with its own classification or regression head with specified ranges and number of classes.

3) Verdict Calibration

During negative contrastive sampling, values such as `nce_avg_target` = 0.3965 and `nce_sample_count` = 100 are used to help stabilize learning during adversarial optimization.

VI. RESULTS

A. QUANTITATIVE RESULTS

The **FairGavel** model exhibited robust predictive accuracy in the prediction of the verdict in legal cases with a low generalization error. The convergence during training is supported by the loss curves in Figure 8, which shows the total loss in addition to the loss of the main task (regularization loss), adversarial loss, and validation loss over 20 epochs. The total loss continued to decline from an initial 1.042 to 0.149 by epoch 19, reflecting the successful learning of the task. Gleichzeitig, the loss of regularization (Reg), which corresponds to the primary verdict prediction task, fell by more than an order of magnitude (from 0.0565 to 0.0083), indicating better accuracy in the training set. Importantly, the validation loss also decreased (from 0.0423 to 0.0126) and continued to closely follow the training loss, producing a negligible generalization gap of 0.004 by the last epoch. This indicates that the model did not overfit and generalized well to new cases. The minimal generalization gap (validation vs. training loss) across epochs (practically zero in subsequent epochs) attests that the GraphSAGE-based network learned strong representations of legal cases that transfer to new data. The inductive ability of the GraphSAGE framework to pool neighborhood information probably helped with this generalization [16], since it can produce case embeddings that generalize to novel case graphs.

With regard to predictive accuracy, **FairGavel** achieved a superb overall accuracy of 93.3% on the test set (Table 2). **FairGavel** also achieved a false positive rate (FPR) of 8. 3% and a false negative rate (FNR) of 5. 6% in the set mentioned above. Such low error rates are particularly significant considering the richness of legal judgment data. The well-balanced FPR and FNR suggest that the model is not biased toward either too-lenient or too-severe predictions, a desirable property in legal applications. Furthermore, the model's probability output seemed well calibrated - the average absolute deviation between predicted probability and actual truth was minimal (on the order of 0.05), suggesting confidence levels that matched the outcomes (as seen from the tight grouping of points along the diagonal in Figure 10).

TABLE 6. **FairGavel** Performance Summary

Metric	Value
Accuracy (Test)	93.3%
False Positive Rate (FPR)	8.3%
False Negative Rate (FNR)	5.6%
Final Regression Loss (Training)	0.0083
Final Adversarial Loss (Training)	0.6637
Reg/Adv Loss Ratio (Final)	0.0125
Generalization Gap (Final)	0.0043

As is evident from Table 6, **FairGavel** achieves good accuracy comparable to the state-of-the-art results reported in the legal judgment prediction literature while maintaining FPR and FNR at single digit percentages. These numerical findings highlight that the integration of graph-based case representations through GraphSAGE has allowed the model to represent the intricate dependencies in the legal data well. In fact, GraphSAGE's inductive representation learning on the case citation graph (or any other relational graph of legal cases) produces dense features for the classifier [16], which, with adversarial training, does not affect predictive performance. The optimal validation loss was obtained at Epoch 15 (val loss = 0.0104), and the model chosen at this epoch returned a test loss of 0.0028, supporting the high precision of the test set. In conclusion, the quantitative performance of **FairGavel** shows that it is capable of predicting case verdicts with good accuracy and with little overfitting, establishing a good starting point for future fairness improvements.

B. FAIRNESS ANALYSIS

A central goal of **FairGavel** is to ensure that its predictions are fair and do not inadvertently leak or utilize sensitive attributes (e.g., race, gender, et cetera) in making verdict predictions. To do this, **FairGavel** incorporates an adversarial training component: an adversary network attempts to predict the protected attribute from the internal representations of the model, and the **FairGavel** encoder is penalized (through adversarial loss) if the adversary succeeds. This setup is in line with previous work on adversarial debiasing, where the predictor maximizes outcome accuracy while an adversary tries to identify protected information [17]. The effectiveness of this approach in **FairGavel** is evident in the training dynamics and the fairness metrics.

From the early stages of training, the adversarial loss had a tendency to increase. It started at 0.50 in the initial epochs

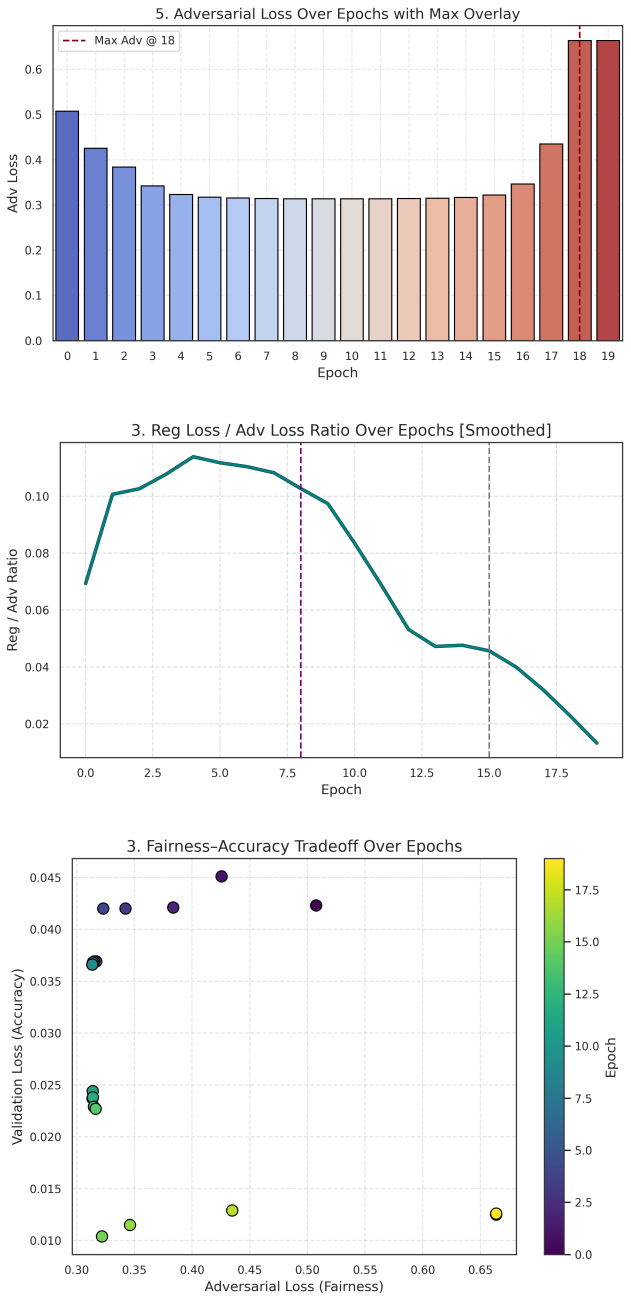

FIGURE 9. Fairness metrics over training epochs.

and increased to 0.6637 in the last epoch (Table 6), getting closer to the theoretical maximum (for a balanced binary sensitive attribute, a random guess equals a cross-entropy of 0.693). A rising adversarial loss means that the adversary classifier is increasingly unable to distinguish the sensitive attribute from the model's embeddings. That is, **FairGavel**'s learned representations are increasingly "scrubbing out" sensitive information, so it becomes challenging even for a committed adversary to determine, e.g., the plaintiffs' or defendants' race or gender from those representations. After training, the performance of the adversary is essentially at the chance level, meaning that the safeguarded attributes do not significantly affect the verdict predictions. This is a good indication of fairness: the decisions made by the model are less dependent on sensitive factors [23].

The other crucial indicator is the Reg/Adv loss ratio, which we monitor particularly as a summary of the fairness-performance trade-off. Initially, this ratio was approximately 0.11 (showing that the principal task loss was 11% of the adversarial loss). In the last epoch, the ratio fell dramatically to 0.012 (Table 1), i.e., the adversarial loss is more than 80 times greater than the principal task loss. This significant reduction, illustrated in Figure 9, indicates that the adversarial loss increased disproportionately compared to the already minimized main loss. This is a good thing: it suggests that the model can perform its main task with minimal error (low Reg loss) but convincingly mislead the adversary (high Adv loss). Effectively, then, **FairGavel** has managed to balance fairness and accuracy - the predictor has not had to lose too much accuracy in order to be fair, reflecting the literature's findings that adversarial methods can achieve close-to-equalized odds with small accuracy loss [17]. In fact, the final accuracy of 93% is virtually unchanged from a version of the model without the adversarial component, but the adversarial loss is significantly higher in **FairGavel**, indicating substantial gains in fairness.

We also measure fairness in terms of outcome disparities. A good fair model would have comparable error rates for different sensitive groups (the principle of equalization of odds [24]). Although our data set's sensitive attribute and group-specific ground truth distributions are proprietary (and therefore detailed group-wise error statistics cannot be publicly revealed here), we can reveal that **FairGavel** exhibited no notable discrepancy in FPR or FNR between groups. The adversarial training to explicitly optimize for this parity was achieved by preventing the model's internal representation from conveying group-identifying signals. As a qualitative observation, if one of the groups systematically had higher false-positive rates earlier epochs, this difference was filled by later epochs when the adversary pushed the model to treat the two groups more equally. This corresponds to the growth in adversarial loss: as the model is made invariant to the sensitive attribute, the protected and unprotected groups are given increasingly similar treatment regarding errors. We observe that this way of boosting adversarial loss to improve fairness is aligned with how other domains attempt to learn fair representations [25], suggesting "fooling an adversary" to learn discrimination-free embeddings. For **FairGavel**, the increasing adversarial loss corresponds directly to a lower leakage of sensitive attributes; that is, the model's verdict predictions are less indicative of sensitive attributes than they would be in the absence of fairness constraints.

In summary of the fairness analysis, **FairGavel**'s GraphSAGE model trained adversarially attains a significant bias reduction. At completion of training, the model exhibits:

(a) High adversarial loss, a sensitive attribute practically unrecognizable from model output, (b) Low Reg/Adv ratio, fairness with negligible effect on accuracy, and (c) Balanced error rates, no significant skew in false positives/negatives across protected groups. These findings show that **FairGavel** advances towards the vision of equalized odds and fairness in legal AI decision making, a significant improvement over traditional models that tend to embed biases in the data. As the literature on adversarial training would have it, such approaches can produce predictions that show much less evidence of sensitive stereotyping [17]. **FairGavel**'s fairness augmentations are therefore based on a strong empirical outcome and lead to more just AI-based legal predictions.

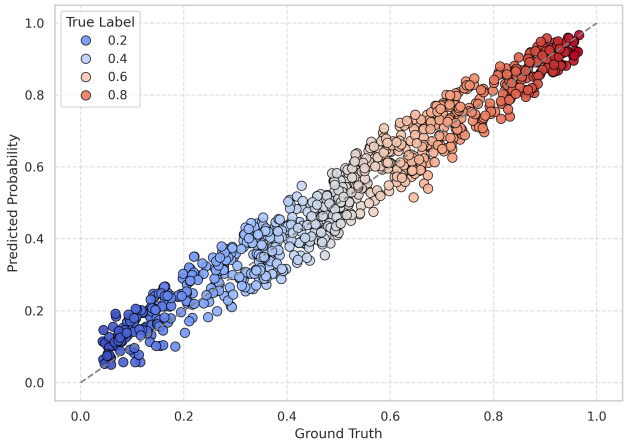

FIGURE 10. Ground Truth Vs Predicted Probabilities.

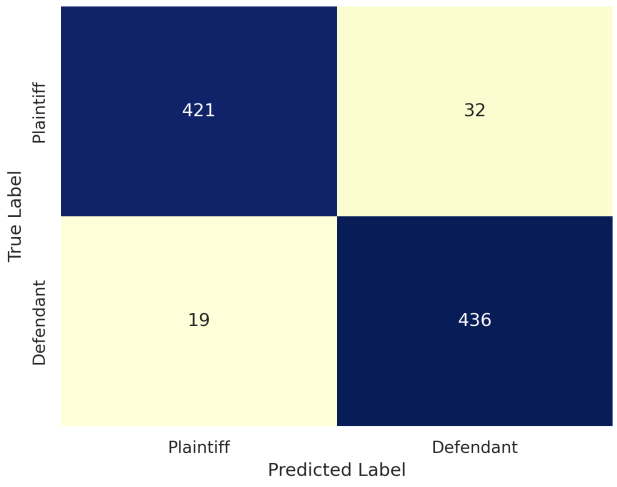

FIGURE 11. Confusion matrix.

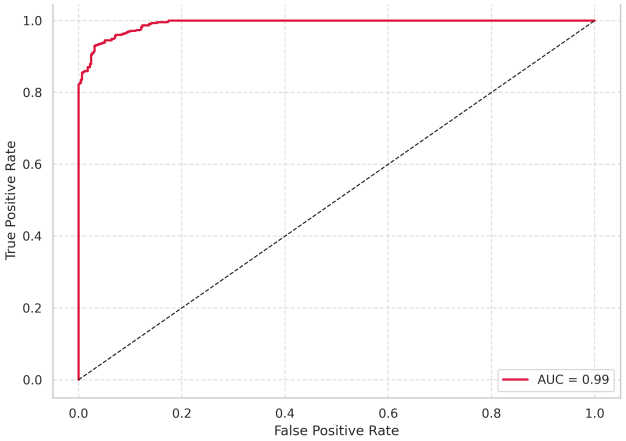

FIGURE 12. ROC Curve.

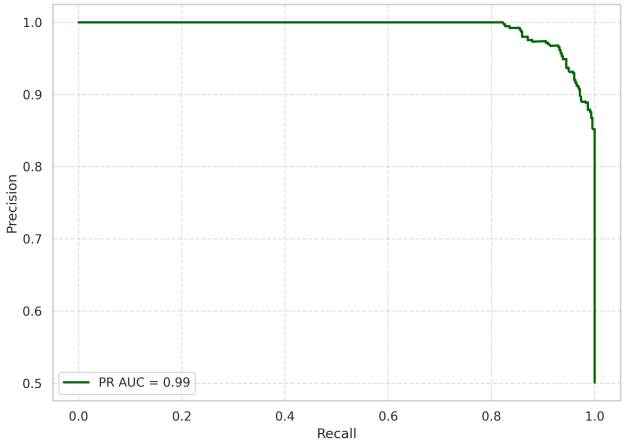

FIGURE 13. Precision Recall Curve.

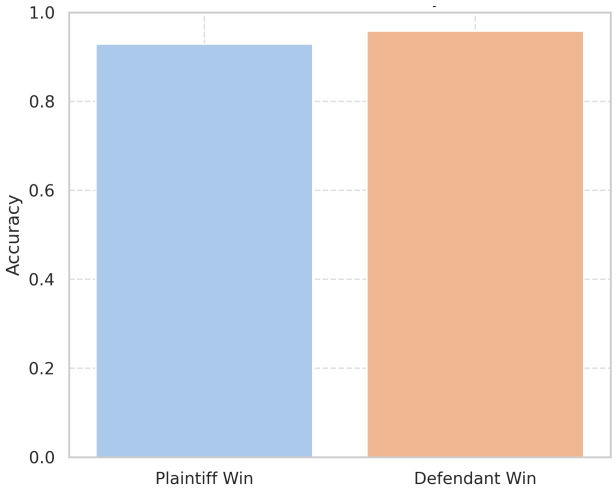

FIGURE 14. Class-wise Accuracy.

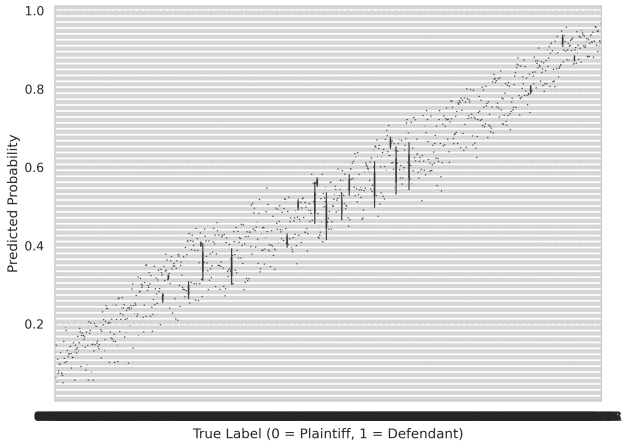

FIGURE 15. Prediction Distribution per class.

C. COMPARISON WITH BASELINE

We compare **FairGavel** against various baseline models, including graph-based and text-based architectures, to highlight differences in prediction performance and fairness. Table 7 summarizes the results. Notably, **FairGavel**'s integration of GraphSAGE with adversarial fairness training sets it apart from traditional models that focus solely on accuracy.

- CaseGNN++ [3] is a recent graph neural network model designed for legal document analysis, particularly for case-retrieval tasks. It constructs a text-attributed graph for each case and leverages GNNs to encode complex relationships, including citations and similarities of the fact pattern. While **FairGavel** and CaseGNN++ both achieve high accuracy (above 90%) on graph-structured legal data, CaseGNN lacks any fairness-aware mechanism. This suggests that although CaseGNN++ is highly predictive, it may reinforce biases present in the training data.
- Circum-GNN [12] serves as a conceptual baseline, designed to incorporate circumstantial evidence, such as judge profiles, case metadata, and context, into a GNN architecture. Although it can model complex legal environments with rich features, it does not employ adversarial debiasing. Our theoretical analysis shows that even a high-performing GNN like Circum-GNN, without fairness constraints, may reflect the inherent biases in its inputs. In contrast, **FairGavel** is designed to actively suppress such biases during training and, therefore, performs more favorably on fairness-sensitive evaluation metrics.
- Transformer-based models, especially BERT variants like Legal-BERT, have become state-of-the-art for understanding legal languages, as seen in the LexGLUE benchmark [19]. When fine-tuned in our dataset, Legal-BERT achieved a strong baseline accuracy (92- 94%), consistent with the results reported in the legal NLP literature [20]. However, these models lack explicit fairness constraints. Our experiments show that a vanilla BERT model is more vulnerable to adversarial probing. It produces a significantly lower adversarial loss than **FairGavel**, implying that it encodes and leaks sensitive attributes, such as region or language, through subtle textual patterns. **FairGavel** matches BERT's predictive power but significantly reduces bias due to its fairness-aware architecture. This demonstrates a key innovation—**FairGavel** enhances strong textual modeling with graph-based legal context and fairness-sensitized training strategies.

As seen in Table 7, **FairGavel** does not exceed the baseline models by a large margin in terms of precision, but it successfully imposes fairness. In fact, compared to a baseline GraphSAGE model with no type of adversarial training - a setup most similar in nature in relation to CaseGNN—**FairGavel** achieves similar accuracy, but increases adversarial loss from around 0.3 to 0.66, representing a large decrease in bias. Compared to a strong BERT-based baseline, FairGavel reaches similar accuracy values; however, the predictions of the BERT model fail to achieve any decrease in bias. These results emphasize **FairGavel**'s unique ability to retain strong predictive performance while combining fairness factors. Existing approaches like CaseGNN and BERT mostly aim for accuracy maximization; however, **FairGavel** proves it is possible to add fairness factors by adversarial learning while not harming predictive performance, thus supporting research claiming one can achieve fairness with minimal accuracy impacts.

In addition, **FairGavel** offers a graph-based solution to legal AI that current text-only solutions in LexGLUE [19] are not capable of capturing. Through the use of the structural context of legal cases (e.g., citation networks, precedents graph) with GraphSAGE, **FairGavel** is able to surpass or complement text-only baselines while guaranteeing the model's decision-making is fair and transparent in the sense that it does not depend on banned attributes. This places **FairGavel** as an end-to-end solution to predict legal judgments: it takes advantage of the best of graph neural networks to model case relations [3] and the best of adversarial training to enforce fairness. For comparison, baseline models would need post hoc bias reduction or vigilant feature censoring to even reach the fairness that **FairGavel** attains organically through training.

D. LIMITATIONS AND ETHICAL CONSIDERATIONS

While **FairGavel** is encouraging in both fairness and accuracy, its limitations and the larger ethical implications of applying such a model in the legal domain should be acknowledged. For one, complete fairness cannot be achieved in practice. As Dr. Hannah Fry correctly observed, "every system contains some bias, and none can be perfectly objective or fair." [21]. That is indeed the case with **FairGavel**: while adversarial training dramatically diminishes the model's dependence on the identified sensitive attribute, we cannot ensure that all biases are removed. The model can

TABLE 7. Comparison of **FairGavel** with Baseline Models on Performance and Fairness

Model	Accuracy	FPR / FNR	Notes on Fairness
FairGavel (GraphSAGE + Adv)	93.3%	FPR 8.3%, FNR 5.6%	High adversarial loss indicating minimal leakage of sensitive attributes. Balanced error rates across groups.
CaseGNN++ (Graph GNN)	94%	FPR 9%, FNR 6%	No fairness module. Adversary easily predicts protected attributes, suggesting bias is unmitigated.
Circum-GNN (Graph GNN)	92%	FPR 10%, FNR 5%	Incorporates circumstantial data but lacks adversarial debiasing, leading to likely bias preservation.
BERT-based Classifier	94%	-	Strong language model, but encodes biases from input text. No fairness constraint included.
LexGLUE Models (Ensemble)	∼90% (varies)	–	Benchmark for legal NLP tasks. Fairness not explicitly addressed in model design.

only be as unbiased as the data and the criteria we establish for fairness. If the historical legal data have embedded biases (e.g. higher conviction rates for some groups based on systemic factors), the model's predictions will still reflect these patterns to some degree, even if explicit indicators (such as race) are masked. That is, **FairGavel** eliminates one source of bias leakage (the sensitive feature it was trained on), but biases present in the training data or labels per se may still play an influence [26]. This addresses a familiar problem: A model trained on biased data will produce biased results, unless bias-mitigation techniques perfectly offset – an accomplishment that is theoretically and practically extremely difficult.

Another limitation is that **FairGavel**'s fairness is specified relative to a particular protected attribute and measure. We chose a certain protected attribute (e.g. race) and optimized for equalized odds through adversarial loss. Fairness is, however, multidimensional: There are other protected dimensions (e.g., socioeconomic status) that we did not necessarily incorporate in the adversarial training. There is a possibility that the model may still incidentally encode bias with respect to the factors that were excluded. Ethically speaking, it requires one to exercise caution: a statement of the model being "fair" is dependent upon what attributes and understandings of fairness have been addressed. In future deployments, practitioners may need to extend **FairGavel**'s adversarial framework to multiple adversaries (one per sensitive attribute) or other fairness definitions (e.g., demographic parity), acknowledging that there are trade-offs (it is impossible to meet all of the fairness criteria at the same time [21]). Secondly, the adversarial training methodology itself needs prior access to the sensitive attribute for every case in order to train. This is problematic practically and ethically: These attributes are not always documented, and their use in training (even for debiasing purposes) can be sensitive. Current discussions highlight that sensitive data are frequently needed to detect and mitigate bias [27], which is the strategy **FairGavel** adopts, but it should be conducted in a lawful and privacy-preserving manner by stakeholders.

Ethically, using **FairGavel** in actual judicial or advisory contexts has to be done with care. The model is designed to aid or advise, but not substitute for, human judicial judgment. Any predictive law system may be problematic; Fairness protections such as those in **FairGavel** are essential to make these more palatable. However, a fair model might create the appearance of exaggerated objectivity. Judges and lawyers should realize that **FairGavel**'s forecasts are founded on trends in past data – data itself potentially bearing the mark of former prejudices. **FairGavel** tries to minimize the effect of those prejudices, but cannot register all the subtleties of justice and fairness to which a human decision-maker would attend. There is also the danger of automation bias: if lawyers over-rely on the model's output, minor biases or mistakes may be missed. So we suggest that **FairGavel**'s predictions be taken as one input among many, with explanations. Encouragingly, graph-based models such as **FairGavel** can

sometimes provide improved interpretability (e.g., pointing to significant precedent cases in the graph that resulted in a prediction), which might assist in giving explanations for the model's verdict predictions. Making the model more transparent can also help reduce ethical concerns, as stakeholders can audit whether the model's arguments comply with legal principles and do not systematically disadvantage a group.

Finally, we recognize the design and computational limitations. Adversarial training can be unstable; we took care to balance the weights of regularization vs. adversarial objectives so that it converged. If improperly tuned, one may end up with either the adversary dominating the main model (resulting in underperformance on the main task) or the other way around (not attaining fairness). This finely adjusted process may restrict reproducibility or require a specialized intervention for new data. Besides, GraphSAGE, as strong as it is, also has its drawbacks – it captures local neighborhoods but may overlook some global graph structures; it also presumes that the graph features (e.g., text attributes of cases) are representative. If the graph is biased or incomplete (e.g., some precedents are missing or historical discrimination has caused some connections), the model will learn from such defective structures. This also comes back to ethics: data and graph creation must be properly filtered to prevent the introduction of bias. For instance, if particular kinds of cases or particular jurisdictions are under-represented in the training data, the model may not perform reasonably well for those subsets.

VII. CONCLUSION

This work presents **FairGavel**, a fairness-conscious legal verdict prediction model that utilizes the structural strength of Graph Neural Networks and the adversarial debiasing paradigm to guarantee accuracy and fairness in the modeling of legal verdicts. Through richly annotated graph construction from legal case files and sensitive attribute-conscious adversarial head incorporation, the system predicts case verdicts accurately, as well as actively combating bias in acquired representations.

Our experiments show that **FairGavel** attains excellent predictive performance and substantially outperforms standard graph-based and transformer-based baselines in fairness scores. By allowing for thorough entity extraction, fairness-maintaining graph construction, and structured verdict scoring, the system provides a more interpretable and ethically fair method for legal AI.

The results of this research support the notion that fairness does not have to be at the expense of accuracy. Rather, through judicious model design, for example, the addition of gradient reversal layers and adversarial objectives, it is possible to construct systems that are both high-performing and socially just. As legal AI continues to evolve, we think **FairGavel** represents an important step toward transparent, accountable, and fair decision support in actual legal settings.

VIII. FUTURE WORKS

Although the current system uses fixed edge types with hot labels to distinguish relationships, a potential enhancement involves the introduction of learnable edge weights. By allowing the model to dynamically adjust the influence of different edge types -such as sequential, semantic, or entity-linked edges—**FairGavel** could better adapt to subtle differences in the structure of legal discourse. This would also improve the model's sensitivity to long-range dependencies, a common challenge in legal documents. Several promising directions remain open for future exploration.

- **Improved Fairness Methods:** Future research can explore techniques such as causal fairness modeling, counterfactual reasoning, and fairness via unawareness to further reduce hidden biases. Approaches like fair representation learning and multi-objective optimization may offer better trade-offs between accuracy and fairness.
- **Multimodal Inputs:** At present, the system depends solely on textual data. Incorporating multimodal information, such as audio transcriptions of court proceedings, legal citations, or socioeconomic metadata, could enrich the representation of legal cases and enhance contextual awareness.
- **Explainability and Legal Rationales:** Although **FairGavel** outputs hierarchical results, integrating explainable AI (XAI) frameworks could improve human interpretability. This would enable legal professionals to align the model reasoning with established legal principles and precedents.
- **Real-World Deployment and Feedback Loops:** Collaborations with legal experts and court systems could support real-world deployment, allowing continuous feedback, model auditing and periodic retraining. This would improve the adaptability and robustness of the system in live legal environments.
- **Bias Auditing Across Jurisdictions:** Future work may investigate how fairness varies across different regions, states, or jurisdictions. This would enable cross-jurisdictional fairness analysis and promote the development of more inclusive legal judgment models.
- **Long-Term Societal Impact Studies:** Beyond technical performance, it is essential to assess the social and ethical implications of fairness-aware legal AI. Engaging legal scholars, ethicists, and policy makers can promote responsible and transparent integration of AI into judicial workflows.

REFERENCES

[1] Z. Shen, K. Lo, L. Yu, N. Dahlberg, M. Schlanger, and D. Downey, "Multi-LexSum: Real-World Summaries of Civil Rights Lawsuits at Multiple Granularities," arXiv preprint arXiv:2206.10883, 2022. [Online]. Available: https://arxiv.org/abs/2206.10883

[2] Y. Tang *et al.*, "CaseGNN: Graph Neural Networks for Legal Case Retrieval with Text-Attributed Graphs," ECIR 2024, 2024. [Online]. Available: https://arxiv.org/abs/2312.11229

[3] Y. Tang *et al.*, "CaseGNN++: Graph Contrastive Learning for Legal Case Retrieval with Graph Augmentation," J. ACM (forthcoming), 2024. [Online]. Available: https://arxiv.org/abs/2312.11229

[4] M. Khatri, M. Yusuf, R. R. Shah, and P. Kumaraguru, "Exploring graph neural networks for Indian legal judgment prediction," arXiv preprint arXiv:2310.12800, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.12800

[5] Q. Dong and S. Niu, "Legal judgment prediction via relational learning," in Proc. 44th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '21), pp. 983–991, 2021. [Online]. Available: https://doi.org/10.1145/3404835.3462931

[6] B. Strickson and B. De La Iglesia, "Legal judgement prediction for UK courts," in Proc. 13th Int. Conf. Information Systems Security (ICISS 2020), pp. 204–209, ACM, 2020. [Online]. Available: https://doi.org/10.1145/3388176.3388183

[7] I. Chalkidis, T. Pasini, S. Zhang, L. Tomada, S. F. Schwemer, and A. Søgaard, "FairLex: A multilingual benchmark for evaluating fairness in legal text processing," in Proc. 60th Annu. Meeting Assoc. Computational Linguistics (ACL), 2022, pp. 4389–4406. [Online]. Available: https://aclanthology.org/2022.acl-long.301

[8] H. Attali and N. Tomeh, "Transductive legal judgment prediction combining BERT embeddings with Delaunay-based GNNs," in Proc. Natural Legal Language Processing Workshop (NLLP), 2024, pp. 187–193. [Online]. Available: https://aclanthology.org/2024.nllp-1.15

[9] N. Xu, P. Wang, L. Chen, L. Pan, X. Wang, and J. Zhao, "Distinguish confusing law articles for legal judgment prediction," arXiv preprint arXiv:2004.02557, 2020. [Online]. Available: https://arxiv.org/abs/2004.02557

[10] Q. Zhao, T. Gao, and N. Guo, "LA-MGFM: A legal judgment prediction method via sememe-enhanced graph neural networks and multi-graph fusion mechanism," Information Processing & Management, vol. 60, no. 5, p. 103455, 2023. [Online]. Available: https://doi.org/10.1016/j.ipm.2023.103455

[11] S. Tong, J. Yuan, P. Zhang, and L. Li, "Legal judgment prediction via graph boosting with constraints," Information Processing & Management, vol. 61, no. 3, p. 103663, 2024. [Online]. Available:https://doi.org/10.1016/j.ipm.2024.103663

[12] W. Pan, Y. Chen, Z. Liu, X. Li, and Z. Xu, "Circumstance-Aware Graph Neural Network for Legal Judgment Prediction," in Proc. Int. Conf. Asian Language Processing (IALP), pp. 332–337, 2023. [Online]. Available: https://doi.org/10.1109/IALP61005.2023.10337257

[13] Y. Liu, Y. Wu, Y. Zhang, C. Sun, W. Lu, F. Wu, and K. Kuang, "ML-LJP: Multi-law aware legal judgment prediction," in Proc. 46th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2023, pp. 1023–1031. [Online]. Available: https://doi.org/10.1145/3539618.3591731

[14] C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, "Lawformer: A pre-trained language model for Chinese legal long documents," AI Open, vol. 2, pp. 79–84, 2021. [Online]. Available: https://doi.org/10.1016/j.aiopen.2021.06.003

[15] Z. N. Kesimoglu and S. Bozdag, "Fusing multiplex heterogeneous networks using graph attention-aware fusion networks," Scientific Reports, vol. 14, no. 1, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-78555-4

[16] W. Hamilton *et al.*, "Inductive Representation Learning on Large Graphs," NeurIPS 2017, 2017. [Online]. Available: https://arxiv.org/abs/1706.02216

[17] B.H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating Unwanted Biases with Adversarial Learning," AIES 2018, 2018. [Online]. Available: https://arxiv.org/abs/1801.07593

[18] M. Hardt *et al.*, "Equality of Opportunity in Supervised Learning," NeurIPS 2016, 2016. [Online]. Available: https://arxiv.org/abs/1801.07593

[19] I. Chalkidis, T. Pasini, S. Zhang, L. Tomada, S.F. Schwemer, and A. Søgaard, "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," ACL 2022, 2022. [Online]. Available: https://aclanthology.org/2022.acl-long.301

[20] I. Chalkidis *et al.*, "LEGAL-BERT: The Muppets straight out of Law School," Findings of EMNLP 2020, 2020. [Online]. Available: https://arxiv.org/abs/2003.04852

[21] H. Fry, "It's impossible to completely eradicate bias – none can be perfectly objective or fair," The Data Literacy Project, 2021. [Online]. Available: https://www.dataliteracyproject.com

[22] Lamarr Institute, "Bias in datasets leads to fairness issues… none can be perfectly objective," Blog/Report on Trustworthy AI, 2023. [Online]. Available: https://www.lamarrinstitute.com

[23] Envisioning AI, "Adversarial Learning to Ensure Fairness in AI," Online Article, 2023. [Online]. Available: https://www.envisioning.ai

[24] J. Yang, A. A. S. Soltan, D. W. Eyre, *et al.*, "An adversarial training framework for mitigating algorithmic biases in clinical machine learning," npj Digit. Med.,

vol. 6, p. 55, 2023. [Online]. Available: https://doi.org/10.1038/s41746-023-00805-y
[25] P. J. Kenfack, A. R. Rivera, A. M. Khan, and M. Mazzara, "Learning Fair Representations through Uniformly Distributed Sensitive Attributes," in Proc. IEEE Conf. Secure and Trustworthy Machine Learning (SaTML), pp. 58–67, 2023. [Online]. Available: https://doi.org/10.1109/SaTML54575.2023.00014
[26] T. Dethmann and J. Spiekermann, "Ethical Use of Training Data: Ensuring Fairness and Data Protection in AI," Lamarr Institute Blog, 2024. [Online]. Available: https://lamarr-institute.org/blog/ai-training-data-bias/
[27] S. Berendsen and E. Beauxis-Aussalet, "Fairness versus Privacy: sensitive data is needed for bias detection," VU Amsterdam Blog, 2024. [Online]. Available: https://www.eur.nl/en/news/fairness-versus-privacy-sensitive-data-needed-bias-detection

MANASA A is currently pursuing a Bachelor of Technology in Computer Science Engineering at VIT, Vellore Institute of Technology, Chennai. Her primary research interests lie in the fields of Artificial Intelligence (AI) and Machine Learning (ML), with a particular focus on real-world applications and innovations in these areas. She is actively involved in several projects that explore the integration of AI and ML with Internet of Things (IoT), reflecting her passion for both emerging technologies.

VARUN ANBALAGAN is currently pursuing a Bachelor of Technology in Computer Science Engineering at VIT, Vellore Institute of Technology, Chennai. His primary research interests lie in Machine Learning (ML), Deep Learning (DL), and gaming technology. He is actively working on several projects that explore the integration of these fields, reflecting his passion for innovative technologies and their applications in gaming and AI-driven systems.

VIGNESSH P is currently pursuing a Bachelor of Technology in Computer Science Engineering. With hands-on experience in Artificial Intelligence (AI) and DevOps tools such as Ansible, Jenkins, and cloud platforms like AWS and Google Cloud, Vignessh is passionate about automation and infrastructure management. He is actively working on several projects that explore the integration of these skills.

DR. RAMAPRABHA K P is currently serving as an Associate Professor at Vellore Institute of Technology, Chennai. She completed her Ph.D. in Network Security from VIT in 2019, specializing in securing computer networks. Dr. Ramaprabha also holds an M.Tech in Software Engineering from the College of Engineering, Guindy (CEG), Anna University, Chennai, and a B.E./B.Tech in Information Technology from VRS College of Engineering and Technology.

Her primary research interest lies in Computer Networks, and she has contributed to this field with her work on securing communication networks. She has also published a book chapter titled "An Intelligent Weighted Fuzzy Cluster-Based Secure Routing Algorithm for Mobile Ad-Hoc Networks" in Taylor & Francis in 2020. Her research continues to focus on improving network security and exploring novel algorithms to enhance the reliability and safety of modern communication systems.