Analyzing and Mitigating Bias for Vulnerable Road Users by Addressing Class Imbalance in Datasets

Katare, Dewant; Noguero, David Solans; Park, Souneil; Kourtellis, Nicolas; Janssen, Marijn; Ding, Aaron Yi

DOI
10.1109/OJITS.2025.3564558

Publication date
2025

Document Version
Final published version

Published in
IEEE Open Journal of Intelligent Transportation Systems

IEEE Open Journal of
Intelligent Transportation
Systems

Received 14 August 2024; revised 16 January 2025 and 11 March 2025; accepted 16 April 2025. Date of publication 25 April 2025;
date of current version 15 May 2025.

Digital Object Identifier 10.1109/OJITS.2025.3564558

Analyzing and Mitigating Bias for Vulnerable Road Users by Addressing Class Imbalance in Datasets

DEWANT KATARE [1] (Graduate Student Member, IEEE), DAVID SOLANS NOGUERO [2], SOUNEIL PARK [2], NICOLAS KOURTELLIS [2] (Member, IEEE), MARIJN JANSSEN [1], AND AARON YI DING [1] (Member, IEEE)

[1]Department of Engineering, Systems and Services, Delft University of Technology, 2600 AA Delft, The Netherlands
[2]Telefónica R&D, Telefónica, 08019 Barcelona, Spain

CORRESPONDING AUTHOR: D. KATARE (e-mail: d.katare@tudelft.nl)

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska Curie Grant under Agreement 956090 (APROPOS); in part by the SPATIAL Project under Grant No 101021808; in part by the CONCORDIA Project under Grant 830927; and in part by the Smart Networks and Services Joint Undertaking (SNS JU) through the European Union's Horizon Europe Research and Innovation Programme under Grant 101096435 (CONFIDENTIAL6G).

ABSTRACT Vulnerable road users (VRUs), including pedestrians, cyclists, and motorcyclists, account for approximately 50% of road traffic fatalities globally, as per the World Health Organization. In these scenarios, the accuracy and fairness of perception applications used in autonomous driving become critical to reduce such risks. For machine learning models, performing object classification and detection tasks, the focus has been on improving accuracy and enhancing model performance metrics; however, issues such as biases inherited in models, statistical imbalances and disparities within the datasets are often overlooked. Our research addresses these issues by exploring class imbalances among vulnerable road users by focusing on class distribution analysis, evaluating model performance, and bias impact assessment. Using popular CNN models and Vision Transformers (ViTs) with the nuScenes dataset, our performance evaluation shows detection disparities for underrepresented classes. Compared to related work, we focus on metric-specific and cost-sensitive learning for model optimization and bias mitigation, which includes data augmentation and resampling. Using the proposed mitigation approaches, we see improvement in IoU(%) and NDS(%) metrics from 71.3 to 75.6 and 80.6 to 83.7 for the CNN model. Similarly, for ViT, we observe improvement in IoU and NDS metrics from 74.9 to 79.2 and 83.8 to 87.1. This research contributes to developing reliable models while addressing inclusiveness for minority classes in datasets. Code can be accessed at: BiasDet.

INDEX TERMS Behaviour metrics, class imbalance, data disparities, cost-sensitive learning, sample representation, object detection, vision models.

I. INTRODUCTION

THE SAFETY of vulnerable road users (VRUs) such as pedestrians, cyclists, and motorcyclists in the vehicular ecosystem remains a challenge and global concern. World Health Organization's statistics describe the risk faced by these groups in traffic-related incidents.[1] As per the WHO report, VRUs account for more than 50% of road traffic fatalities worldwide, with developing countries having even higher numbers [1], [2]. The report also describes that in the urban driving environments, the VRUs represent around 70% of total fatalities. These statistics and analysis become concerning when the focus is on developing and deploying connected and automated vehicle systems and services using sensor suites and state-of-art object detection models. The correctness of vision and perception applications is measured using the model performance metrics, which include accuracy, precision (average, mean), recall, and intersection over union (IoU) [3], [4], [5]. Metrics such as robustness, uncertainty measure, and fairness of these vision applications are often overlooked. These metrics help to identify and quantify trustworthiness and AI model performance in

The review of this article was arranged by Associate Editor Chen Wang.
[1]https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

non-ideal conditions and real-world scenarios [6], [7], [8], [9], [10]. Research studies have covered the existence of biases, class imbalances and disparities in vision datasets, including autonomous driving datasets, which could lead to unfair object classification and false predictions of vulnerable road users (VRU) [2], [11].

Class imbalance generally refers to the unbalanced or unequal representation of different categories or classes within a dataset [12], [13], [14], [15]. The imbalance and class disparities often lead to over-representation for a type of class; for example, in popular datasets, the most represented class is vehicles and pedestrians, while classes such as cyclists, motorcyclists, and especially mobility-impaired individuals are classes that remain under-represented [16]. Such an imbalance can lead to perception systems that are less adept at recognizing minor yet vulnerable classes [16], [17], [18]. Studies have shown that this imbalance can lead to the inheritance of biases in AI models from the datasets, where models preferentially detect overrepresented classes [19]. Figure 1 shows a heatmap from our tests, where images on the left are input images, and gradients flow for a class are shown on the right. As shown in the image, the bicycle class is overlooked or has a false detection.

The consequences of such an imbalance and biases in the dataset cannot be overlooked, as it can lead to incidents like the self-driving car collision [20]. Ethical issues of biased autonomous driving systems are significant, as even with accurate perception applications and systems, there is a growing requirement for equitable AI systems that ensure the safety and fairness of all road users [21]. Also, regulatory bodies are increasingly focusing on how these technologies follow the safety standards, such as their ability to detect and respond to diverse and representative [22] scenarios in a given latency measure. Within this scope, we aim to address class imbalances and possibilities of bias inheritance in an AI model when trained with class disparities. For study and test purposes, we use two major datasets in the autonomous (vision) domain, nuScenes [23] and Waymo [24]. This research emphasizes the need for balanced, diverse and representative datasets that include a wide range of scenarios, reflecting real-world complexity. In this dataset imbalance context, our paper explores the following questions:

- How do class representation and disparities within datasets influence the accuracy and reliability of AI models in detecting vulnerable road users (VRUs)?
- In what ways do biases from under-represented classes in training datasets gets inherited in machine learning models for perception tasks?
- What are the impacts of class disparities on the model's performance metrics and the model's ability to perform across varied and unseen environments?

As compared to related works, which are based on the analysis of "F-1 Score, False Positives(FP) and False Negatives(FN)" for bias impact assessment tasks [3], [15], [19], [25], [26], our research focuses on behavioural metrics to study inclusiveness for minority classes within the datasets and class disparities between majority and minority samples by performing empirical tests on CNNs, DNNs and ViTs. This approach complements the previous work, adding value to the error analysis techniques. Our approach shows model learning processes and behaviour towards a particular class or representation during training. The key contributions are:

1) We explored behavioural metrics for CNNs [26] and adopted it for 3d features, which can be implemented on DNNs processing camera and lidar data. We further used layer-wise relevance propagation for the vision transformer to assess bias effectively.
2) We investigated the impact of class and sample disparities on weight distribution during the model learning process. We also evaluated the empirical performance using metric-specific and cost-sensitive learning.
3) We propose a framework integrating bias assessment, mitigation strategies, and model recalibration, ensuring model fairness and accuracy.

FIGURE 1. Results when one class is preferred during training, leading to the 'unseen' class being poorly detected.

II. RELATED WORK

Perception applications and tasks in autonomous driving have evolved through extensive research and development and progression in sensor technology, algorithmic processing, and machine learning models [5], [27]. Early perception systems relied on rule-based algorithms and limited sensor input, but the advancements in deep learning and improvements in LiDAR, radar, and camera technologies have provided next-generation solutions [28], [29], [30]. However, developing a supervised learning-based fair perception system with object classification and detection that can reliably

interpret diverse and unpredictable road scenarios remains an important challenge [31], [32]. For supervised learning, sample distribution within datasets plays an important role in training and validating autonomous driving systems. Initial vision datasets like KITTI [33] and Cityscapes [34] have provided foundations for advancing the field, specifically from 2d to 3d classification and detection. Recent popular datasets like nuScenes and Waymo have offered more diverse and complex model training and testing environments. These datasets include various annotations for vehicles, pedestrians, cyclists, and other classes, providing a rich foundation for developing and evaluating perception algorithms.

A. CLASS IMBALANCE

Data augmentation, re-sampling, and cost-sensitive learning are some known techniques that are used to address class imbalance [15], [17], [41]. The goal includes balancing the representation of classes in training data [42] and developing fair and accurate models across all categories [19], [43]. To detect imbalanced classes, a scenario-based simulation approach is proposed by Hahner et al. [44]. Targeting imbalanced representation, an approach that combines visual codebook generation with deep features and a non-linear Chi^2 SVM classifier, is proposed in [14]. This method tackles the issue of imbalanced datasets, where algorithms often fail to detect minority classes. The approach extracts low-level deep features using transfer learning with the ResNet-50 pre-trained model and k-means clustering to create a visual codebook. Each image is then represented as a Bag-of-Visual-Words (BOVW), derived from the histogram of visual words in the vocabulary. The Chi^2 SVM classifier is used for classifying these features, showing optimized performance in empirical analysis. This method shows better accuracy, F1-score, and AUC metrics results than state-of-the-art methods, validated on two datasets: Graz-02 and TF-Flowers [14].

He et al. [6], addresses the issue of fairness in cooperative driving strategies. By integrating fairness considerations into cooperative driving mechanisms used in congested on-ramps, the study shows via simulation results that balancing fairness and traffic efficiency is possible. Focusing on fairness in demand prediction for ride-hailing services, a socially-aware neural network (SA-Net) is introduced by Zheng et al. [8]. The study integrates socio-demographic data to improve prediction fairness. The SA-Net, supplemented by a bias-mitigation regularization, reduces prediction disparities between underprivileged and privileged communities, showing improvements in both the accuracy and fairness of estimation. Exploring the impact of data management in vehicular networks, Figueiredo et al. [43] propose an algorithm to optimize the inclusion of extra object data in Collective Perception Messages (CPMs). The algorithm improves the efficiency and effectiveness of data transmission within vehicular networks, showing potential for substantial improvements in object information transmission with minimal additional network delay.

With a focus on training methods to address the challenges of learning from imbalanced data, a new loss function that mitigates the impact of samples leading to overfitted decision boundaries is proposed in [45]. This loss function has been shown to enhance the performance of various imbalance learning methods. The proposed approach is robust and can be integrated with existing resampling, meta-learning, and cost-sensitive learning methods to tackle class imbalance problems [45]. An approach for predicting driver behaviour critical for safely integrating autonomous vehicles into human-dominated traffic is discussed in [25]. The proposed method addresses the research gaps in existing predictive models, which lack transparency (deep neural networks) or are not explainable (rule-based models). The authors introduce a model that embeds the Intelligent Driver Model (IDM), a rule-based approach, into deep neural networks. This hybrid model combines the long-term coherence and interpretability of rule-based models with the expressiveness of deep learning, aiming to accurately predict driver behaviour in complex scenarios like merging. The method is an attempt to bridge the gap between two modelling paradigms. It enhances the interpretability of neural network predictions while maintaining accuracy, a critical factor for real-time decision-making in autonomous driving. The model's transparency is particularly beneficial for debugging and understanding its predictions.

Existing tools and frameworks that address the issue of class imbalance to train unbiased ML models include AIF360 [31], an open-source library that includes multiple algorithms specifically designed to reduce bias in datasets and models. AIF360 includes techniques like reweighting, which modifies the weights of training instances to address the underrepresentation of a particular class or classes. Similarly, Fairlearn [32], another open-source toolkit, provides interactive visualizations and algorithms to explore and mitigate bias, focusing on balancing fairness and model performance. Ghosh et al. [37] provides a detailed overview of the challenges of class imbalance in machine learning. The author proposes various resampling and algorithmic adjustment techniques, such as SMOTE (synthetic minority over-sampling technique), and their effectiveness in managing class disparities. The analysis shows the need for more adaptive resampling methods tailored to specific machine-learning scenarios to enhance the performance and fairness of predictive models. Chen et al. [38], proposes an ensemble technique to address model learning challenges associated with imbalanced datasets. Balancing class influence during model training enhances model performance across various datasets. The paper's empirical test shows the ensemble strategy's effectiveness, though the authors discussed that extending this approach to more complex architectures could provide model learning knowledge. Table 1 compares related baseline work with our proposed method.

B. MODEL LEARNING AND REPRESENTATION

Wang et al. [35] introduce a deep attention-based method for imbaanced image classification (DAIIC), using an attention mechanism within a logistic regression framework to prioritize minority classes in prediction and feature

TABLE 1. Comparison of related work on bias identification and mitigation in vision datasets.

Study	Methodology	Bias Type	Mitigation Approach	Contribution
DAIIC [35]	Deep attention-based classification	Class imbalance	Cost-sensitive learning	Improved minority class detection and representation
Tejani et al. [36]	Theoretical framework	Systematic and cognitive biases	Diverse data representation and model monitoring	Framework to mitigate bias across AI development stages
Ghosh et al. [37]	Review and comparison	Class imbalance	Resampling and algorithmic adjustments	Highlights the need for adaptive resampling methods
Chen et al. [38]	Ensemble technique	Class imbalance	Balancing class influence during training	Enhances model performance across various datasets
SMOTE [39]	Synthetic minority over-sampling	Class imbalance	Generating synthetic samples	Advances in minority class representation
REVISE [19]	Dataset analysis tool	Object-based, gender-based	Preemptive dataset analysis	Identifying under representations in visual datasets
BAdd [40]	Bias addition	Spurious correlations	Bias inducing attributes with Model architecture	Learning bias-neutral representations
This Work	Behavioral metric analysis	Class imbalance	Cost-Sensitive learning and Data augmentation	Error analysis and Model learning representations

representation. It automatically determines misclassification costs to aid discriminative feature learning. Robust across different networks and datasets, the DAIIC method has proven effective, surpassing several benchmarks in single-label and multi-label contexts [35]. The background and related work in autonomous driving technologies, datasets, and the challenges of class imbalance and bias show the importance of this research area. Our study builds upon these foundations, aiming to contribute to the development of more equitable and reliable autonomous driving systems. By addressing class imbalance and bias in key datasets, we seek to enhance the safety and fairness of these technologies for all road users. In this scope, libraries such as AIF360 [31] also include state-of-the-art bias mitigation methodologies such as adversarial debiasing, which enhances the learning of more equitable representations by scoping the model to overlook protected classes or attributes. Similarly, Fairlean [32] offers a visualization suite that enables an in-depth examination of model predictions across different groups, highlighting potential biases and facilitating the application of mitigation strategies. Tejani et al. [36] explores the origin and impact of biases in imaging AI systems, proposing a framework to mitigate these biases. The paper identifies that biases can emerge at various stages of AI development, from data handling to model deployment, and suggests that these biases can contribute to health disparities. The recommended strategies for mitigation include ensuring diverse data representation and ongoing monitoring of models post-deployment. While providing a thorough theoretical discussion on systematic and cognitive biases, the paper suggests that real-world applications and empirical validations could further provide novel insights. Elreedy et al. [39] proposed SMOTE as an approach in addressing class imbalances by generating synthetic samples. This technique is complemented with random over-sampling in enhancing the minority class representation. The authors discuss both the theoretical framework and practical applications of SMOTE.

Chen et al. [46] provided an evaluation of bias mitigation techniques by discussing the performance of machine learning models. Through empirical testing, it was found that these methods often reduce the performance of models in an attempt to enhance fairness, with a decrease in machine learning performance observed in over half of the scenarios tested. This study shows the variability in the effectiveness of bias mitigation techniques, which largely depends on the specific machine learning tasks, models, and fairness metrics used. The research covered 17 bias mitigation methods, 11 performance metrics, 4 fairness metrics, and 20 types of fairness-performance trade-offs across various decision-making tasks. The paper's critical insight shows that no single method provides an end-to-end solution, highlighting the requirement of joint-optimized mitigation solutions specific to applications. However, the study also notes that improvement in fairness could potentially compromise model performance, suggesting a complex trade-off that must be achieved in practical applications. BAdd [40] introduces a method for reducing bias in machine learning by incorporating biased attributes into the training process to achieve a fair model representations. The test and analysis shows improvement in model accuracy across benchmarks with single and multi-attribute biases, such as FB-Biased-MNIST and CelebA, showing its efficiency over existing methods. Based on the subsection discussion and our scope of supervised learning, the current challenges in biases occurrence in the datasets and inheritance in AI models can be described as:

1) Class Disparities: Ensuring fair representation of classes and samples in training datasets to improve the fairness and accuracy of perception systems.

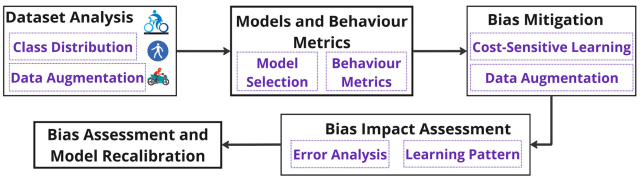

FIGURE 2. Proposed methodology for bias analysis.

2) Model Interpretability: Improving the understandability of complex machine learning models to facilitate debugging, understanding, and learning patterns.
3) Adaptive Performance: Developing systems that reliably interpret and react to dynamic, unpredictable environments such as varying weather conditions and traffic scenarios.
4) Regulatory and Ethical Compliance: Aligning with safety standards and ethical considerations, including addressing privacy, accountability, and socio-economic impacts.

III. METHODOLOGY

When an imbalanced dataset D with an unjustified distribution of classes $C_1, C_2, \ldots, C_n$ is used to train machine learning models, specifically neural networks and Vision Transformers (ViTs), the models are likely to inherit biases because of class representation in the dataset. These biases may arise due to imbalanced class representation and a predominance of normal conditions as compared to challenging real-world scenarios. For neural networks, we anticipate biases to be reflected in neurons' sensitivity and selectivity scores, while for vision transformers, biases may manifest in the attention allocation. Therefore, we propose a methodology consisting of statistical analysis for analyzing and mitigating biases using dataset analysis, bias impact assessment, bias mitigation, metrics behaviour analysis, and bias assessment and model re-calibration, as shown in Figure 2. The following subsections detail and discuss the concepts and terminologies used in our methods and tests.

A. DATASET AND CLASSES

In this paper, we use nuScenes as the case study, focusing on the Pedestrian, Cyclist, and Motorcyclist classes. However, the approach can be tested and validated on other driving datasets, e.g., the Waymo dataset [24], where the samples and classes can be varied.

A) Class Distribution Analysis: We perform statistical analysis on the frequency of the specified classes in the nuScenes and Waymo datasets to identify class imbalances. The datasets are analyzed to understand the context (urban versus rural settings, diverse weather and lighting conditions) in which these classes are represented, highlighting any underrepresented scenarios.

Pedestrian Class: nuScenes dataset includes a major proportion of pedestrian annotations, with 149,921 instances labelled as "human.pedestrian.adult" accounting for 21.61% of all annotations. This substantial representation highlights the importance of pedestrian detection in autonomous driving systems. However, other pedestrian sub-categories like children (1,934 annotations, 0.28%), construction workers (13,582 annotations, 1.96%), and individuals with personal mobility devices (2,281 annotations, 0.33%) are less represented. These differences shows gaps in dataset and subsamples diversity.

Cyclist Class: It is categorized as "vehicle.bicycle" and has around 17,060 annotations, comprising 2.46% of the total dataset. This relatively lower representation compared to pedestrians impacts the class weight calculation and, therefore, the model's ability to accurately detect cyclists, a safety concern in urban driving scenarios.

Motorcyclist Class: It is annotated as "vehicle.motorcycle" and has around 16,779 annotations, making up 2.42% of the dataset. This class is slightly underrepresented compared to pedestrians but is on par with cyclists. The lesser representation of cyclists and motorcyclists indicates an area for improvement in class balance, especially considering the different dynamics and associated risks.

The standard metrics used for model performance evaluation in nuScenes are *mean average Precision* (mAP), *Intersection over Union* (IoU), and *nuScenes Detection Score* (NDS). mAP and IoU are well-known metrics used in 3d object detection and segmentation tasks [23], and are calculated using the fundamentals of precision, recall, area of overlap and intersection over union respectively. Similarly, NDS combines several metrics, including the mAP, to provide a model performance score for the object detection models. The NDS metric is calculated as:

$$NDS = \frac{1}{10}\left[5 \cdot mAP + \sum_{mTP \in TP} (1 - \min(1, mTP))\right] \quad (1)$$

B) Data Augmentation/Re-sampling: Given the class representations, particularly for the less represented subcategories of pedestrians and overall lower representation of cyclists and motorcyclists, following strategies can be used for model training and evaluation.

Resampling Technique: Addressing class imbalances in the datasets for underrepresented classes, like cyclists and motorcyclists, is essential for unbiased training in autonomous driving models. We apply both "Random Oversampling" and "Undersampling" from resampling techniques". Random Oversampling duplicates instances from minority classes, ensuring models have sufficient data to learn from these critical yet less frequent road users. While, undersampling reduces the dominance of over-represented classes, like pedestrians, to prevent bias inheritance in models. Both methods are balanced to maintain data diversity [47].

Data Augmentation: Another key approach is the use of geometric transformations, such as rotating and flipping images. This technique represents various class orientations, which is crucial for autonomous driving where objects are

encountered from multiple directions. These transformations prevent the model from developing orientation-specific biases and enhance its ability to adapt to diverse real-world scenarios. Additionally, this approach contributes to increasing the dataset's diversity and size, thereby improving the model's robustness and generalization capabilities for more accurate and reliable perception tasks [48].

Combination of Strategies: For varied training, testing, and to increase the samples/sub-classes diversity, we combine resampling and data augmentation to balance the training dataset. This approach integrates the strengths of both Resampling (Random undersampling/Oversampling) and Transformations, such as rotating and flipping images, to optimize the representation and variability of data.

B. MODELS AND BEHAVIOURAL METRICS

We strategically use popular models such as ResNet18 [49], SqueezeNet [50] from the CNN, Centerpoint [51], FS3D [52] from DNN and ViT [29] from the transformer family for evaluation. The selection of these models is intended to provide an understanding of how different architectures perform on biased datasets and to compare their ability to generalize across various classes, especially after the implementation of bias mitigation. Using these models, we further gain insights across different neural network architectures, ensuring robustness and reliability in our findings.

Model Selection: ResNet18 [49] is a well-known architecture and performs fairly in perception tasks due to its ability to capture spatial hierarchies, but it can show a tendency of bias towards common textures and patterns. SqueezeNet [50], is a compact model and focuses on global image features but risks inheriting biases from non-diverse datasets [26]. Vision Transformers (ViTs) process images by analyzing sequences of patches, adeptly focusing on relevant image parts due to their attention mechanisms [29]. While ViTs are proficient at capturing dependencies, these models may struggle to focus on underrepresented classes/unique scenarios adequately. Each model's inherent strengths and learning mechanisms influence classification and learning, highlighting the need for tailored bias detection and mitigation strategies.

Behaviour Metrics: Sensitivity and selectivity scores of neurons are used to evaluate biases in models like ResNet and SqueezeNet. Sensitivity scores measure a neuron's response to changes in input from specific classes, indicating potential biases. For example, a ResNet neuron more sensitive to vehicles than cyclists might show a vehicle-detection bias. Selectivity scores, on the other hand, assess the specificity of a neuron's response to a class, with high selectivity indicating specialized recognition capabilities. This helps identify whether models effectively differentiate between classes or show biases, such as a score showing higher selectivity for pedestrians over motorcyclists, potentially indicating a feature recognition bias. The formula for the measures is described as follows:

$$\text{Sensitivity Score} = \partial a / \partial x \tag{2}$$

where a is activation of neuron, and x is the input feature.

$$\text{Selectivity Score} = \frac{a_c - a_{avg}}{\max(a_c, a_{avg})} \tag{3}$$

where a_c is activation of the neuron for the target class and a_{avg} is the average activation for all classes.

Traditional convolutional neural networks (CNNs) process 2D image data, thus applying sensitivity and selectivity metrics to measure the neural response per pixel or feature map does not require specific adaptation. However, 3D object detection deep neural networks (DNNs) such as CenterPoint and FS3D manage complex data types or modalities, including 3D point clouds and fused data from lidar and cameras. These models require metrics that can accommodate the unique 3D spatial relationships, intensities, and RGB data present in their inputs. The adaptation of sensitivity metrics for these 3D DNNs involves modifying the traditional approach to account for 3D spatial features or fusion features. This includes changes in point positions or attributes and their impact on detection confidence. Hereby, in this paper, we represent sensitivity scores for 3D data contexts as follows:

$$\text{Sensitivity Score (3D)} = \frac{\partial \text{Detection Confidence}}{\partial \text{Point Feature}} \tag{4}$$

This formula helps quantify how minor changes in input features influence the model's output, providing insights into potential biases or model dependencies. For selectivity, while CNNs evaluate activation differences due to traditional image features, 3D DNNs focus on distinguishing objects based on 3D spatial features or sensor fusion data. The selectivity score for 3D DNNs can be expressed as:

$$\text{Selectivity Score (3D)} = \frac{|a_{c,\,3D} - a_{avg,\,3D}|}{\max(a_{c,\,3D}, a_{avg,\,3D})} \tag{5}$$

where a_c is the activation for the target class, and a_{avg} is the average activation across all classes, adjusted for 3D data. The application of these metrics requires visualization techniques, such as 3D Grad-CAM or point cloud saliency mapping, which shows sensitivity and selectivity in a three-dimensional context. These visual tools are essential for understanding how models prioritize some features over others, potentially influencing their decision-making processes and dominance while learning.

In Vision Transformers (ViTs), attention map analysis is key for identifying model focus areas and potential decision-making biases. This involves extracting attention weights from each Transformer layer, which indicates the model's focus on different image parts. These heatmaps, created using the scaled dot-product attention mechanism, help determine if the model disproportionately focuses on some features,

potentially signalling bias. For ViTs the attention weights are calculated as:

$$A = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

where Q (Query), K (Key), and V (Value) are matrices derived from the input, and d_k is the dimension of the Key vectors. By analyzing these attention maps across different layers and attention heads, we can identify whether the model disproportionately focuses on certain features or misses critical aspects, potentially indicating bias.

Layer-wise Relevance Propagation (LRP) offers another approach, starting from the output layer and backpropagating the relevance score of the predicted class through the network. LRP redistributes each layer's relevance to its inputs, particularly in self-attention layers, where relevance among patches is allocated based on attention weights [53]. The relevance for a patch in a layer, highlighting the model's decision-making basis is calculated as:

$$R_j^{(l)} = \sum_i A_{ij}^{(l)} R_i^{(l+1)} \tag{7}$$

where $R_j^{(l)}$ is the relevance of patch j in layer l, $A_{ij}^{(l)}$ are the attention weights from patch j to patch i in the self-attention mechanism, and $R_i^{(l+1)}$ is the relevance in the next layer.

C. BIAS IMPACT ASSESSMENT

This section describes model errors (false positives and negatives) and examines learning patterns and bias inheritance, which are fundamentals for identifying potential biases in model performance.

A) Error Analysis: We focus on false positives (e.g., incorrectly identifying objects as belonging to a target class) and false negatives (e.g., failing to detect cyclists). We further use *Class-specific error* analysis, which includes evaluating how well the model detects pedestrians, cyclists, and motorcyclists under various scenarios. We also examine if there is a tendency to overlook vulnerable classes and how this impacts overall model performance. The correlation between these errors and previously identified biases, such as neurons' sensitivity or selectivity scores, is also explored to understand the underlying causes and guide bias mitigation strategies using *bias correlation*.

B) Learning Patterns and Bias Inheritance: Different model architectures lead to varied learning behaviours and potential biases. CNNs like ResNet may focus on textures and local patterns, which could introduce biases if such features are not uniformly present across all classes. This can result in misclassification, especially when distinguishing features are missing. In comparison, Vision Transformers (ViTs) may develop biases based on how attention is distributed across an image.

Bias Inheritance from Data: A model's performance can vary depending on the diversity/samples encountered in training. For instance, a model trained predominantly on urban pedestrian images might underperform in rural settings. Understanding this bias inheritance is crucial for ensuring that perception systems are adaptable to diverse real-world conditions and can guide necessary adjustments in data representation and model training.

Comparative Analysis: By comparing how different models like ResNet, SqueezeNet, and ViTs respond to the same dataset, we can uncover specific biases inherent in each architecture. This comparative approach helps identify which models are more prone to biases, helping categorise the most suitable model for bias-sensitive applications.

Interpretation and Visualization: Layer-wise Relevance Propagation (LRP) provides insights into what drives a model's decisions by highlighting influential input features. For example, if a model consistently focuses on irrelevant background features for decision-making, it indicates a bias that needs correction [54]. Analysis of these scores, using visual tools like heatmaps and attention maps, helps in interpreting these complex models, offering an understanding of their decision-making processes, and adjusting the training process to mitigate identified biases. Overall, this process involves the analysis of learning patterns, data bias inheritance, comparative model analysis, and visualization, which can help discover bias occurrence/inheritance in different model architectures.

D. BIAS MITIGATION TECHNIQUES

We explore the strategies mentioned below to mitigate biases inherited from sample representation and address these class disparities. These techniques are necessary for addressing the disparities and ensuring that the model's performance is equitable and reliable.

A) Cost-Sensitive Learning: Customized cost-sensitive learning offers a robust approach to mitigating these issues by adjusting the model's loss function according to the representation of each class. The class weights calculation and loss function are discussed below.

Class Weights Calculation: Based on the dataset statistics, we calculate the weights for the pedestrian w_p, cyclist w_c, and motorcyclist w_m classes using the inverse of their respective representation percentages. This approach assigns higher weights to underrepresented classes, emphasizing their importance during training. The calculated weights are:

$$w_p = \frac{1}{21.61\%}; w_c = \frac{1}{2.46\%}; w_m = \frac{1}{2.422\%}$$

These weights are then normalized to ensure they contribute proportionally and maintain stability during training.

Loss Function: We integrate these calculated weights into the model's loss function, utilizing a weighted multi-class cross-entropy loss. For each instance i, the loss is defined:

$$L_i = -w_p y_{ip} \log(p_{ip}) - w_c y_{ic} \log(p_{ic}) - w_m y_{im} \log(p_{im}) \tag{8}$$

Here w_p, w_c, and w_m are the weights for pedestrians, cyclists, and motorcyclists. y_{ix} is a binary indicator of

whether instance i belongs to a class x. p_{ix} is the model's predicted probability for instance i belongs to a class x.

Implementing Focal Loss [10] provides another practical mechanism that focuses on hard-to-classify instances by reducing the loss contribution from easily classified examples. This dynamic adjustment is controlled through a tunable focusing parameter, which can lead to faster and more effective learning on datasets where intra-class variation in example difficulty is significant. By mitigating the influence of numerous easy examples, Focal Loss can enhance the model's learning efficiency, allowing for a more nuanced understanding and handling of complex class imbalances. However, the Weighted loss functions, particularly weighted cross-entropy, offer a straightforward and easily implementable solution. By assigning weights inversely proportional to class frequencies, they directly compensate for imbalances, ensuring that minority classes have a proportionally higher influence during the training process. This method simplifies the implementation for the models selected for experimental evaluation and further enhances the stability and predictability of the training process, making it a robust choice for preliminary model training where simplicity and direct addressing of class imbalances are prioritized. Therefore, the selection between these two loss functions depends on the dataset's specific requirements and the desired balance between simplicity and dynamic learning efficiency.

Dynamic Adjustment Evaluation: To ensure continual model adaptability, we propose dynamically adjusting these weights based on performance metrics. This approach can be tested using a validation set to monitor performance improvements for underrepresented classes and to prevent overfitting. By using this approach, we aim to reduce the biases inherent in the AI model from the imbalanced dataset, ensuring an equitable and robust model performance across all classes, which is critical for the reliability and safety of perception applications.

Reweighing: In our methodology, we propose to use an extended version of reweighing algorithm [55], which extends the traditional class frequency adjustments by performing reweighting of the classes in a dataset that has a diverse representation of samples and classes. Recognizing the complexity of bias in autonomous driving scenarios, our weights are dynamically calculated not only based on the underrepresentation of classes but also considering factors such as error sensitivity and historical bias measures. For example, if historical data indicates that certain classes, such as cyclists, are consistently prone to higher false negative rates, our model significantly increases their weights. Additionally, we factor in the predictive importance of each class in the overall model performance, ensuring that critical classes influencing safety-critical decisions receive appropriate attention during training.

Cross-Metric Optimization: As the proposed method includes model learning observation using behavioural metrics and evaluation using performance metrics, in our method, we also consider an approach to enhance model performance by implementing cross-metric optimization, which extends the methodology focus of weighting(class) from balanced class representation to also including sensitivity, selectivity and model performance metrics in the weighting process. This approach allows for model tuning, debiasing and addressing performance through a choice of class representation and model calibration.

B) Data Augmentation Based on Model Analysis: We use a targeted data augmentation strategy to address biases identified in ViTs through attention map analysis and Layer-wise Relevance Propagation (LRP). This approach specifically addresses under-representation issues in cyclists and motorcyclists within the nuScenes dataset.

Attention-Guided Augmentation: Insights from attention maps guide this augmentation strategy. For instance, if the model frequently overlooks pedestrians under night or varied lighting conditions, then there is a need for more class samples highlighting these scenarios. Techniques such as zooming or adjusting brightness/contrast are used to emphasize these aspects. Similarly, additional variations of poses or orientations for motorcyclists are incorporated to improve model recognition abilities in these contexts.

LRP-Informed Sampling: Layer-wise Relevance Propagation provides an understanding of which features contribute most to the model's decisions. Using insights from LRP, the dataset can be augmented to enhance the representation of features critical for correct classifications. Augmentation for this case includes scenarios where the model typically misclassifies images, like partially occluded subjects or specific textures and patterns, enhancing the model's ability to differentiate between relevant and misleading features.

Implementation: This strategy involves iterative dataset refinement based on continuous model analysis. By aligning augmentation closely with model performance, we aim to mitigate biases, ensuring balanced and fair model performance. This is critical for perception systems, where the accurate detection of all road users, particularly those underrepresented like cyclists and motorcyclists.

E. BIAS ASSESSMENT AND MODEL RE-CALIBRATION

Effective bias mitigation requires continuous evaluation and adjustment of models. This section describes our approach to bias assessment and model re-calibration.

Bias Assessment: In this step, we use behaviour metrics, specifically sensitivity and selectivity scores for each class, to measure the impact of our mitigation efforts. Comparing these metrics before and after mitigation measures provides insights into changes in model behaviour. Additionally, we analyze error rates, particularly false positives and negatives, to assess improvements in the model's predictive accuracy across different classes. For Vision Transformers, attention map analysis is used to verify if attention distribution is now more balanced across various classes and scenarios.

Model Re-calibration: This involves dynamically adjusting class weights within the loss function, guided by real-time performance metrics. The step ensures the model stays optimized for any shifts in class representation or emerging imbalances. When new class samples or data become available, the model undergoes re-training to stay aligned with the evolving operational context. The last step includes an iterative refinement, monitoring and updating bias mitigation strategies based on the latest performance assessments.

IV. EXPERIMENTAL EVALUATION

This section covers training, testing and evaluation methods. The overall dataset comprises approximately 1.4 million images, with annotations across several classes. Our focus is on the Pedestrian (149,921), Cyclist (17,060), and Motorcyclist (16,779) classes, as it provides a rich source for analyzing biases in object detection.

Preprocessing Steps: Data preprocessing involved normalization of image pixel values to the [0,1] range, aligning annotations to ensure consistency, and resizing images to standard dimensions suitable for input requirements of chosen models.

Model Architectures: The study uses SqueezeNet and ResNet18 from the CNN family alongside Vision Transformer: ViT. SqueezeNet, known for its compressed size and robustness in feature extraction, and ResNet18, recognized for its efficiency in learning from residual connections, are expected to provide insights into traditional CNN performance. ViT, representing the Transformer family, is included to evaluate the effectiveness of its attention-based mechanism in handling class imbalances.

A. HYPERPARAMETER OPTIMIZATION

Initial hyperparameters were set as follows: learning rate of 0.001, batch size of 32, and weight decay of 0.0001 for CNN models. For ViT we configured the training with a batch size of 32 and an initial learning rate of 1e-3, following a linear decay to 1e-5. The model was trained for 30 epochs using the Adam optimizer. A dropout rate of 0.1 and a weight decay of 0.03 is applied to avoid overfitting.

Optimization Techniques: Hyperparameter tuning is performed using a grid search approach, systematically varying learning rates, batch sizes, and dropout rates to identify the optimal combination [18]. The optimal set of hyperparameters was chosen based on the highest average scores across these metrics as followed in [23], focusing on improvements in detecting classes.

B. TRAINING PROCESS

Our sample size is small, so CNN models were trained for 50 epochs using an Adam optimizer. A learning rate scheduler was used to reduce the learning rate by 10% every 10 epochs. Regularization techniques such as dropout and data augmentation (rotations and flipping) were used to prevent overfitting [56].

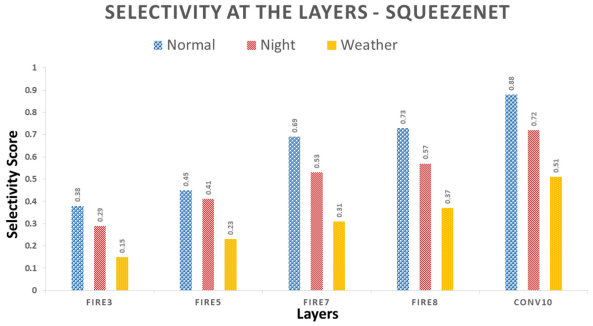

FIGURE 3. Layerwise selectivity for the Classes.

The dataset is divided into 70% training, 15% validation, and 15% test splits, ensuring a fair representation of scenarios in each set. Training is conducted on a high-performance computing cluster with a combination of NVIDIA Tesla V100 GPUs using Open MPI and PyTorch as the deep learning framework.

C. MODEL EVALUATION

Models were evaluated using average precision, mAP (mean average precision), %IoU and NDS (Intersection over Union, NuScenes Detection Score), alongside sensitivity and selectivity scores for class analysis.

Baseline Comparison: Model performances pre and post-implementation of bias mitigation strategies were compared. Baseline models were trained with respect to data conditions.

Class-specific Analysis: Performance metrics for each class were used to assess improvements in detecting pedestrians, cyclists, and motorcyclists. For analysis, attention is given to false positive or negative rate changes for these classes. The tests for CNN can be performed using layer-wise analysis and by varying data diversity where one class has a dominant presence.

Layer-wise Analysis: This analysis shows whether specific layers are biased towards particular classes, which provides strategies for bias mitigation. For instance, if early layers are biased towards detecting pedestrians more than cyclists or motorcyclists, there will be a need to adjust the training data with inclusiveness. Figure 3 shows Selectivity scores at various layers of SqueezeNet under different conditions: Normal, Night, and Weather. The plot shows that selectivity increases in deeper layers, with the highest scores observed under normal conditions and comparatively lower scores during Night and Weather scenarios. A similar analysis can be done at the epoch level, where performance metrics, training loss and accuracy can be analysed with behavioural metrics, thus allowing feature learning progression.

D. BIAS IMPACT ANALYSIS

Error Analysis: Figure 4 shows error analysis for the discussed classes. The ResNet18 model showed a higher tendency for false positives in all classes than the ViT model,

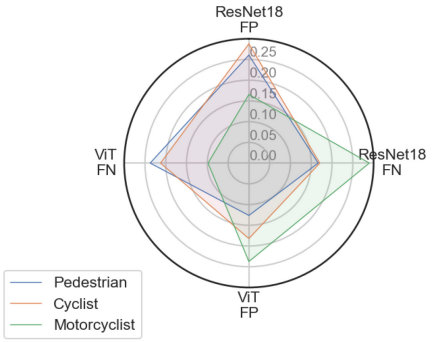

FIGURE 4. Visualization of class-wise error rate for ResNet and ViT according to False Negatives and False Positives in model validation.

FIGURE 5. Figure showing heatmap for all three classes.

TABLE 2. NDS and %IoU of VRU with ResNet18.

%IoU of Classes			NDS	Data Condition
Pedestrian	Cyclist	Motorcyclist		
86.1	53.6	76.4	72.0	Normal
78.5	8.3	40.7	42.5	Night
67.3	7.5	18.9	31.2	Weather
85.7	52.8	76.1	71.5	Rotated
72.1	12.4	21.4	35.3	Mixed

TABLE 3. NDS and %IoU for classes with ResNet18.

%IoU of Classes				NDS	Data Condition
Car	Pedestrian	Cyclist	Motorcyclist		
77.0	68.1	53.6	56.4	64.9	Normal
73.1	64.5	8.9	38.7	45.6	Night
72.8	61.3	7.8	20.9	39.2	Weather
76.7	67.9	52.8	56.1	62.5	Rotated
71.4	65.1	12.9	17.6	36.8	Mixed

TABLE 4. mAP of VRUs with SqueezeNet.

Pedestrian	Cyclist	Motorcyclist	Total	Data Condition
86.3	54.7	77.8	72.9	Normal
78.1	7.5	38.3	41.3	Night
67.0	8.3	19.7	31.6	Weather
85.9	52.3	75.2	71.1	Rotated
73.2	14.8	22.1	36.7	Mixed

suggesting a preference toward overprediction. Also, the ViT model shows fewer false positives but has a higher number of false negatives, particularly with Pedestrians and Cyclists, highlighting a less effective detection strategy for the class. The ViT model shows a more balanced result for the Motorcyclist class with significantly reduced false negatives than ResNet18, indicating more reliable detection.

Visualizations: Figure 5 shows heatmap visualization to represent the models' focus areas and any changes before the mitigation strategies. When trained with unbalanced classes, the tests show that both CNN models falsely classify motorcyclist and cyclist classes.

E. VRU ANALYSIS WITH VEHICLE CLASS

As one of the most represented class in the vision datasets is vehicle also sometimes used as Car, we further evaluate models and respective performance for detailed performance analysis by including vehicle class (Car) with vulnerable road users. As the sample representation of the Vehicle class in the dataset is around 36.6%, our class weight calculations are adjusted as:

$$w_v = \frac{1}{36.6\%}; w_p = \frac{1}{21.6\%}; w_c = \frac{1}{2.4\%}; w_m = \frac{1}{2.4\%}$$

F. DISCUSSION OF RESULTS

For measuring IoU, mAP and NDS scores mentioned in Table 2 and Table 4, we used a subset approach, where the representation of one class (e.g., pedestrian) is kept at 67% and the other two classes have equal representation. This ratio is adjusted during iterations until all classes have the same representation. The purpose is to capture metrics in a scenario where the weighted class functions are normalized. Table 2 shows the Intersection over Union (%IoU) for different classes and the NuScenes Detection Score (NDS) across various conditions for ResNet18. Pedestrian detection remains relatively high in all conditions, with %IoU above 70% except when combined adverse conditions are present. Cyclist detection suffers significantly in all but normal conditions, with %IoU dropping to as low as 7.5% during adverse weather. Motorcyclist detection is also affected by these conditions but to a lesser extent. The overall NDS reflects these trends, with the highest score of 72.0 in normal conditions and the lowest of 31.2 during adverse weather. The rotated condition only slightly impacts the %IoU for pedestrians and motorcyclists but more so for cyclists. Mixed conditions present a scenario with a noticeable decrease in %IoU across all classes and a resultant NDS of 35.3.

Table 4 shows the mean average precision (mAP) of pedestrian, cyclist, and motorcyclist detection under various conditions. Under normal conditions, pedestrian detection is high at 86.3%, while cyclist detection lags at 54.7%. Motorcyclist detection is reasonably high at 77.8%, leading to an overall mAP of 72.93%. However, model performance drops for the night and adverse weather conditions, with the lowest cyclist mAP at 7.5% and 8.3%, respectively. furthermore, rotating images in normal conditions slightly

TABLE 5. Mean average precision (mAP) of different classes with SqueezeNet under various conditions.

Car	Pedestrian	Cyclist	Motorcyclist	Total	Condition
76.1	67.3	8.7	27.8	44.9	Normal
74.7	58.1	5.5	22.3	39.9	Night
73.2	49.0	4.3	19.7	35.7	Weather
75.8	65.6	7.9	21.2	42.1	Rotated
74.9	61.8	4.6	24.1	40.8	Mixed

TABLE 6. Class-wise performance on the CNN models.

Class	Model	%IoU	NDS	Sens.	Sel.
Pedestrian	ResNet18	65.3	64.1%	0.74	0.78
Cyclist	ResNet18	68.4	79.5%	0.65	0.67
Motorcyclist	ResNet18	70.2	80.3%	0.70	0.72
Car	ResNet18	75.1	78.1%	0.78	0.83
Pedestrian	SqueezeNet	78.1	85.4%	0.81	0.83
Cyclist	SqueezeNet	72.6	82.7%	0.73	0.75
Motorcyclist	SqueezeNet	74.0	83.2%	0.76	0.77
Car	SqueezeNet	78.1	85.4%	0.81	0.83

Sens. = Sensitivity Score, Sel = Selectivity Score

reduces pedestrian and motorcyclist detection by about 1%, but impacts cyclist detection, dropping to 52.3%. Combining all adverse conditions reduces overall performance to a mAP of 36.7%, highlighting the challenges the models face in less-than-ideal scenarios.

As shown in Table 3 and Table 5, the inclusion of the vehicle class in our model evaluations using ResNet18 and SqueezeNet shows the tradeoff in model performance metrics under various data conditions. Specifically, the presence of a well-represented vehicle class impacts the learning process and representation of vulnerable road users (VRUs) such as cyclists and motorcyclists, particularly under challenging conditions like night and adverse weather. Normal and Rotated data conditions maintain relatively high performance across all classes, there is a decrease in %IoU and NDS values when vehicles are included, suggesting a shift in model selectivity towards the more frequently represented class. Under Night, Weather, and Mixed conditions, the performance drop for cyclists and motorcyclists is higher, which provides information on potential model overfitting or bias towards vehicles. This adjustment in class weights to include a dominant vehicle category requires a recalibration to prevent overlooking of minority classes. Statistical parity and optimization could benefit from experimenting with different class weighting strategies, enhancing data augmentation for underrepresented classes, and possibly adjusting model architecture to balance detection accuracy across diverse operational scenarios.

Class-wise Performance Metrics: Table 6 shows class-wise performance metrics for two models, ResNet18 and SqueezeNet. For the pedestrian class, SqueezeNet shows better performance than ResNet18, with a higher %IoU and NDS. Cyclists and motorcyclists also see better %IoU and NDS scores with SqueezeNet. Sensitivity and selectivity scores, which measure the models' ability to identify and differentiate classes, are consistently higher for SqueezeNet across all classes, indicating a more refined recognition capability. These metrics show overall better performance from SqueezeNet models in distinguishing and correctly identifying the classes.

TABLE 7. Baseline vs post-mitigation performance.

Model	Metric (Avg)	Baseline	Post Mitigation
ResNet18	%IoU	71.3%	75.6%
ViT	%IoU	74.9%	79.2%
ResNet18	NDS	80.6%	83.7%
ViT	NDS	83.8%	87.1%

TABLE 8. Performance analysis on Argoverse2.

Models baseline performance				
Method	**APped**	**APcyc**	**APmotor-cyc**	**mAP**
CenterPoint	48.6	26.5	37.1	37.4
FS3D	61.4	34.5	51.7	49.2
Post mitigation performance				
CenterPoint-1	54.8	38.8	42.0	45.2
FS3D-1	62.8	36.4	56.7	51.9

Baseline vs Post-Mitigation Performance: Table 7 shows average Intersection over Union (IoU) and NuScenes Detection Score (NDS) before and after applying bias mitigation techniques for ResNet18 and Vision Transformer (ViT). Both models show an improvement in %IoU and NDS after mitigation, with ViT having a high NDS score. ResNet18's IoU improved by 4.3%, and its NDS by 3.1%, while ViT's IoU and NDS improved by 4.3% and 3.3%, respectively. This shows the effectiveness of the bias mitigation strategies.

Table 8 compares the performance of object detection models CenterPoint and FS3D on the Argoverse2 dataset before and after mitigation efforts. Initially, CenterPoint had a mAP of 37.4%, which improved to 45.2% post-mitigation, with gains across all categories–pedestrians up to 54.8%, cyclists to 38.8%, and motorcyclists to 42.0%. FS3D started with a higher baseline mAP of 49.4% and rose to 51.9% after mitigation, showing slight improvement in pedestrian detection to 62.8% and a significant increase for motorcyclists to 56.7%. These results indicate the effectiveness of the mitigation strategies in enhancing detection accuracies.

Table 9 shows performance metrics (average precision) for pedestrian, cyclist, and motorcyclist detection using models CenterPoint and FS3D across the baseline dataset and their results CenterPoint-1 and FS3D-1 with class representation modification. With modified class sample representation for training, the model shows improvements in sensitivity and selectivity, particularly for pedestrians and motorcyclists. For example, selectivity for pedestrians in FS3D-1 increased from 0.82 to 0.89, indicating precise detection with fewer false positives. Motorcyclist detection also improved, with sensitivity rising from 0.70 to 0.76 in FS3D-1, enhancing the detection rate. Cyclists, however, showed minimum gains,

TABLE 9. Class performance on the Argoverse2.

Class	Model	AP	Sens.	Sel.
Pedestrian	CenterPoint	48.6	0.73	0.77
Cyclist	CenterPoint	26.5	0.63	0.68
Motorcyclist	CenterPoint	37.1	0.69	0.74
Pedestrian	FS3D	61.4	0.77	0.82
Cyclist	FS3D	34.5	0.64	0.69
Motorcyclist	FS3D	52.5	0.70	0.78
Pedestrian	CenterPoint-1	54.8	0.81	0.82
Cyclist	CenterPoint-1	38.8	0.62	0.69
Motorcyclist	CenterPoint-1	42.0	0.73	0.75
Pedestrian	FS3D-1	62.8	0.85	0.89
Cyclist	FS3D-1	36.4	0.65	0.70
Motorcyclist	FS3D-1	56.7	0.76	0.80

Sens. = Sensitivity Score, Sel = Selectivity Score

TABLE 10. Performance analysis on Waymo level 1.

AP/APH	Vehicle	Ped	Cyc	mAP/mAPH
CenterPoint	75.1/ 77.6	78.2/ 74.9	71.8/ 70.4	75.0/ 74.3
FS3D	77.5/ 77.2	80.9/ 74.2	76.1/ 75.3	78.1/ 75.5
Post mitigation performance				
CenterPoint-1	77.6/ 78.1	80.5/ 76.2	73.1/ 70.6	77.0/ 74.9
FS3D-1	78.1/ 77.5	81.6/ 75.2	77.0/ 76.1	78.9/ 76.2

TABLE 11. ML performance metrics comparison.

Class	Model	Precision	Recall	mAP
Car	CenterPoint	0.88	0.76	0.80
Pedestrian	CenterPoint	0.83	0.75	0.69
Cyclist	CenterPoint	0.58	0.42	0.47
Motorcyclist	CenterPoint	0.55	0.39	0.41
Post mitigation performance				
Car	CenterPoint	0.85	0.70	0.77
Pedestrian	CenterPoint	0.81	0.68	0.63
Cyclist	CenterPoint	0.57	0.34	0.45
Motorcyclist	CenterPoint	0.52	0.31	0.37

mAP = Mean Average Precision

highlighting the ongoing challenge of accurately detecting class which has common features. These results show the benefits of bias mitigation in enhancing detection accuracy and model fairness.

Table 10 shows a pre and post-mitigation performance analysis on the Waymo dataset. In this test case, we used vehicle, pedestrian and cyclist classes to evaluate performance metrics. As shown in the table, post-mitigation models (CenterPoint-1 and FS3D-1) show improvement in metrics. For e.g., the APH for vehicles increased from 77.6 to 78.1 in CenterPoint-1 and from 77.2 to 77.5 in FS3D-1. The pedestrian class also has improvements in the AP and APH, with FS3D-1 showing an increase from 74.2 to 75.2 in APH. However, for the cyclists class, the metrics improvements are low compared to the other classes, as also observed in the Argoverse2 dataset.

Table 11 shows a detailed results from the ML performance metrics Precision, Recall, and Mean Average Precision (mAP) for classes including Cars, Pedestrians, Cyclists, and Motorcyclists as evaluated using the CenterPoint model. Before mitigation efforts, the Car class has the highest Precision (0.88) and mAP (0.80), showing the model's ability to identify and predict car class and subclass accurately. Pedestrian class, which has the second highest presence in the data subset, also showed robust model performance with Precision at 0.83 and mAP at 0.69. However, Cyclists and Motorcyclists shows significantly lower metrics, with Cyclists showing a particularly low mAP of 0.47, reflecting the challenges the model faces in handling less represented classes. After implementing mitigation strategies, all classes has a decrease in Precision and Recall, suggesting that the mitigation process affects the model's sensitivity and specificity. For e.g., the Precision for Cars decreased to 0.85 and mAP to 0.77, while Pedestrians has a reduction in Precision to 0.81 and mAP to 0.63. Cyclists and Motorcyclists shows lower performance metrics post-mitigation, with Cyclists' mAP further reducing to 0.45. These changes show the model's challenge in accurately detecting less represented classes even post-mitigation, highlighting the need for more refined strategies that can enhance performance without compromising detection accuracy across all classes.

Visualizations: Figure 6 shows attention heatmap for the model before sampling and mitigation strategies. Here, high density can be seen in the pedestrian ('person') class and then on a bicycle, which shows the potential impact of high representations present during model training. Figure 7 and Figure 8 show mean and attention maps for the motorcyclist class by still capturing features from the cyclist present in the input image. Even after using post-mitigation strategies, the attention heatmap is not uniform across the cyclist class. The model used for testing the images has dominant weights from the motorcyclist class, thus failing to capture the features of the cyclists. Figure 9 show the results from the test after the sampling and mitigation strategies, which have an equal number of classes. The overall attention response of the model across the entire scene can be visualised here, as the model processes complex scenes involving multiple objects and VRUS. The attention and density are concentrated on areas with dynamic objects, which are critical, but as can be seen from the image, the model also assigns weights to features represented by cyclists and pedestrians to other classes. The overall key takeaways can be described as:

- For CNN models, error analysis shows ResNet18 has higher false positive rates than SqueezeNet.
- Post-mitigation performance, which involves resampling approaches, shows improvements in both CNN (ResNet18) and ViT.
- Under all conditions, the pedestrian class had higher detection in VRUs. However, overall post-mitigation performance shows a decrease, requiring robust model development.

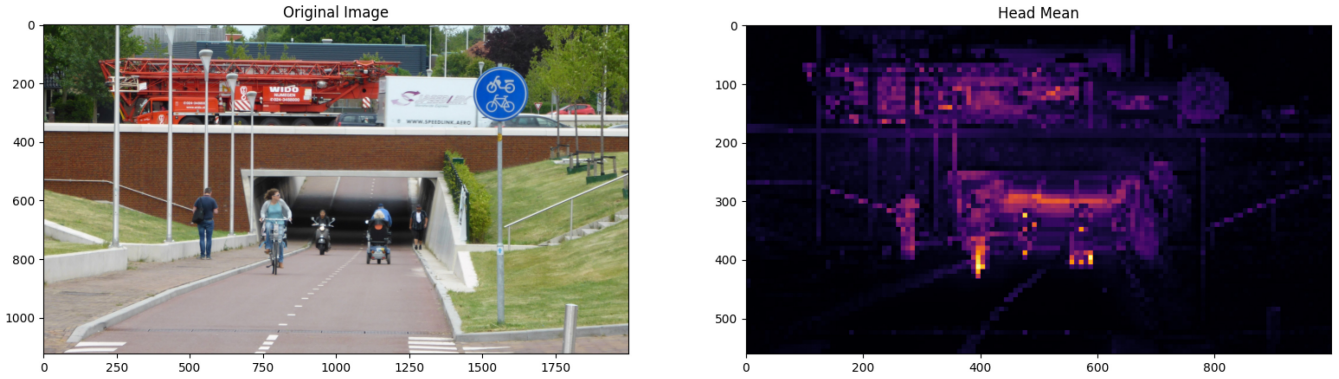

FIGURE 6. Figure showing mean for all three classes.

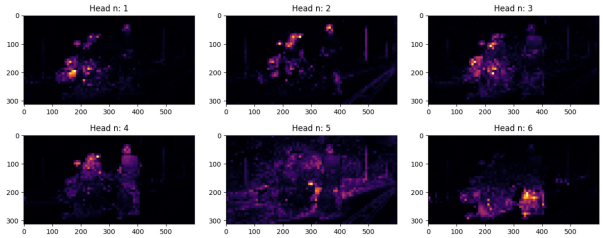

FIGURE 7. Head-wise attention maps for motorcyclist class.

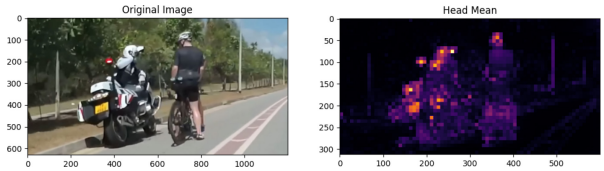

FIGURE 8. Figure showing mean for motorcyclist classes.

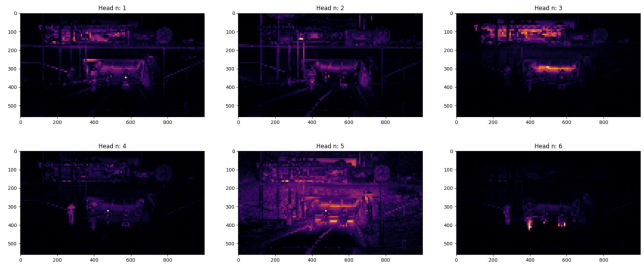

FIGURE 9. Head-wise attention weights for all classes.

- Equal number of samples from classes in a subset, does not necessarily ensure an equal representation of behaviour metrics during model learning.

G. METHODOLOGY ADAPTATION TO OTHER DATASETS

To adapt the proposed methodology to other visual datasets (such as Lyft, A2D2, etc.), the first step is to identify class imbalances or over-representation in the input data using class distribution analysis, which involves statistically evaluating class frequency to ensure fair representation, as biases might arise due to more frequent appearance of particular objects in specific environments. The next step will be to detect these biases using behavioural metrics on the trained model: sensitivity and selectivity. For example, in a dataset like CIFAR-10, where all ten classes have equal representation, the behavioural metrics help to identify potential biases arising from the class's local features, texture, and colour, while in other datasets such as Lyft or A2D2, the focus will be on the models' ability to detect different object sizes and their spatial relationships accurately. The third step is implementing mitigation strategies, including resampling techniques, using oversampling to increase the presence of underrepresented classes and undersampling to decrease over-representation. While CIFAR-10 may need simple adjustments due to its uniform image sizes, Lyft and A2D2 datasets could benefit from more complex data augmentation using rotation, flipping, and adding image corruptions to existing samples of CIFAR-10 and using generative adversarial networks to represent and add complex scenarios to Lyft and A2D2. The last step in adapting the methodology is implementing cost-sensitive learning to modify the training loss functions using weighted calculation (of class in the dataset) and adding it to the proposed loss function.

V. CONCLUSION AND FUTURE WORK

This study explored the issue of class imbalance in driving datasets and its impact on the performance of AI models, specifically focusing on the under-represented classes in a dataset, which are also Vulnerable Road Users (VRUs) in the vehicle ecosystem. Using our detailed methodology, which includes dataset analysis, model evaluation, and bias impact assessment, we detect disparities in the representation and detection of vulnerable road users. Our tests with popular CNN models and vision transformers have shown how dataset biases can lead to alternate learning outcomes, adversely affecting the accuracy and correctness of perception systems. We have also implemented and evaluated bias mitigation techniques, which include cost-sensitive learning and targeted data augmentation. These methods have shown promising results in improving model performance, especially in IoU and NDS metrics. Therefore, introducing a dynamic framework for ongoing bias assessment and model recalibration is an essential step towards developing more equitable AI systems in autonomous driving.

REFERENCES

[1] S. K. Ahmed et al., "Road traffic accidental injuries and deaths: A neglected global health issue," *Health Sci. Rep.*, vol. 6, no. 5, 2023, Art. no. e1240.

[2] R. M. Silva et al., "Vulnerable road user detection and safety enhancement: A comprehensive survey," 2024, *arXiv:2405.19202*.

[3] S. S. Rajan, E. Soremekun, Y. Le Traon, and S. Chattopadhyay, "Distribution-aware fairness test generation," *J. Syst. Softw.*, vol. 215, Sep. 2024, Art. no. 112090.

[4] C. Zhang and A. Eskandarian, "A quality index metric and method for online self-assessment of autonomous vehicles sensory perception," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 13801–13812, Dec. 2023.

[5] M. Yu, K. Gong, W. Zhao, and R. Liu, "LiDAR and IMU tightly coupled Localization system based on ground constraint in flat scenario," *IEEE Open J. Intell. Transp. Syst.*, vol. 5, pp. 296–306, 2024.

[6] Z. He, H. Pei, Y. Guo, D. Yao, and L. Li, "Theoretical trade-off between fairness and efficiency in the cooperative driving problem for CAVs at on-ramps," *IEEE Open J. Intell. Transp. Syst.*, vol. 5, pp. 41–54, 2024.

[7] M. Emu, F. B. Kamal, S. Choudhury, and Q. A. Rahman, "Fatality prediction for motor vehicle collisions: Mining big data using deep learning and ensemble methods," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 199–209, 2022.

[8] Y. Zheng, Q. Wang, D. Zhuang, S. Wang, and J. Zhao, "Fairness-enhancing deep learning for ride-hailing demand prediction," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 551–569, 2023.

[9] S. Rezvani and X. Wang, "A broad review on class imbalance learning techniques," *Appl. Soft Comput.*, vol. 143, Aug. 2023, Art. no. 110415.

[10] T.-Y. Ross and G. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2017, pp. 2980–2988.

[11] J. Siegel and G. Pappas, "Morals, ethics, and the technology capabilities and limitations of automated and self-driving vehicles," *AI Soc.*, vol. 38, pp. 213–226, Feb. 2023.

[12] M. Saini and S. Susan, "Tackling class imbalance in computer vision: A contemporary review," *Artif. Intell. Rev.*, vol. 56, no. 1, pp. 1279–1335, 2023.

[13] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," *IEEE Trans. pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3388–3415, Oct. 2021.

[14] M. Saini and S. Susan, "Bag-of-visual-words codebook generation using deep features for effective classification of imbalanced multi-class image datasets," *Multimedia Tools Appl.*, vol. 80, pp. 20821–20847, Jun. 2021.

[15] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. Conf. Fairness, Account. Transp.*, 2018, pp. 77–91.

[16] D. Lee and J. Kim, "Resolving class imbalance for lidar-based object detector by dynamic weight average and contextual ground truth sampling," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 682–691.

[17] M. Mazumder et al., "DataPerf: Benchmarks for data-centric AI development," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, 2024, pp. 1–20.

[18] H. Alibrahim and S. A. Ludwig, "Hyperparameter optimization: Comparing genetic algorithm against grid search and Bayesian optimization," in *Proc. IEEE Congr. Evol. Comput. (CEC)*, 2021, pp. 1551–1559.

[19] A. Wang et al., "REVISE: A tool for measuring and mitigating bias in visual datasets," *Int. J. Comput. Vis.*, vol. 130, no. 7, pp. 1790–1810, 2022.

[20] N. A. Stanton, P. M. Salmon, G. H. Walker, and M. Stanton, "Models and methods for collision analysis: A comparison study based on the Uber collision with a pedestrian," *Saf. Sci.*, vol. 120, pp. 117–128, Dec. 2019.

[21] A. Shariff, J.-F. Bonnefon, and I. Rahwan, "How safe is safe enough? Psychological mechanisms underlying extreme safety demands for self-driving cars," *Transp. Res. Part C, Emerg. Technol.*, vol. 126, May 2021, Art. no. 103069.

[22] K. Chasalow and K. Levy, "Representativeness in statistics, politics, and machine learning," in *Proc. ACM Conf. Fairness, Account., Transp.*, 2021, pp. 77–89.

[23] H. Caesar et al., "nuscenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11621–11631.

[24] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1–9.

[25] S. Arbabi, D. Tavernini, S. Fallah, and R. Bowden, "Learning an interpretable model for driver behavior prediction with inductive biases," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2022, pp. 3940–3947.

[26] D. Katare, N. Kourtellis, S. Park, D. Perino, M. Janssen, and A. Y. Ding, "Bias detection and generalization in AI algorithms on edge for autonomous driving," in *Proc. IEEE/ACM 7th Symp. Edge Comput. (SEC)*, 2022, pp. 342–348.

[27] H. Itsuji, T. Uezono, T. Toba, and S. Kumar Kundu, "Real-time diagnostic technique for AI-enabled system," *IEEE Open J. Intell. Transp. Syst.*, vol. 5, pp. 483–494, 2024.

[28] D. Katare, D. Perino, J. Nurmi, M. Warnier, M. Janssen, and A. Y. Ding, "A survey on approximate edge AI for energy efficient autonomous driving services," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 4, pp. 2714–2754, 4th Quart., 2023

[29] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*,

[30] L. G. Jaimes, H. Chintakunta, and P. Abedin, "SenseNow: A time-dependent incentive approach for vehicular crowdsensing," *IEEE Open J. Intell. Transp. Syst.*, vol. 5, pp. 307–321, 2024.

[31] R. K. E. Bellamy et al., "AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias," *IBM J. Res. Develop.*, vol. 63, nos. 4–5, pp. 4, pp. 1–4:15, 2019.

[32] S. Bird et al., "Fairlearn: A toolkit for assessing and improving fairness in AI," Microsoft, Redmond, WA, USA, Rep. MSR-TR-2020-32, 2020.

[33] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[34] M. Cordts et al., "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3213–3223.

[35] L. Wang, L. Zhang, X. Qi, and Z. Yi, "Deep attention-based imbalanced image classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 8, pp. 3320–3330, Aug. 2021.

[36] A. S. Tejani, Y. S. Ng, Y. Xi, and J. C. Rayan, "Understanding and mitigating bias in imaging artificial intelligence," *RadioGraphics*, vol. 44, no. 5, 2024, Art. no. e230067.

[37] K. Ghosh, C. Bellinger, R. Corizzo, P. Branco, B. Krawczyk, and N. Japkowicz, "The class imbalance problem in deep learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4845–4901, 2024.

[38] Z. Chen, J. Duan, L. Kang, and G. Qiu, "Class-imbalanced deep learning via a class-balanced ensemble," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5626–5640, Oct. 2022.

[39] D. Elreedy, A. F. Atiya, and F. Kamalov, "A theoretical distribution analysis of synthetic minority oversampling technique (SMOTE) for imbalanced learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4903–4923, 2024.

[40] I. Sarridis, C. Koutlis, S. Papadopoulos, and C. Diou, "BAdd: Bias mitigation through bias addition," 2024, *arXiv:2408.11439*.

[41] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Syst. Appl.*, vol. 73, pp. 220–239, May 2017.

[42] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. CVPR*, 2011, pp. 1521–1528.

[43] A. Figueiredo, P. Rito, M. Luís, and S. Sargento, "Enhancing vehicular network efficiency: The impact of object data inclusion in the collective perception service," *IEEE Open J. Intell. Transp. Syst.*, vol. 5, pp. 454–468, 2024.

[44] M. Hahner, C. Sakaridis, D. Dai, and L. Van Gool, "Fog simulation on real LiDAR point clouds for 3D object detection in adverse weather," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15283–15292.

[45] S. Park, J. Lim, Y. Jeon, and J. Y. Choi, "Influence-balanced loss for imbalanced visual classification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 735–744.

[46] Z. Chen, J. M. Zhang, F. Sarro, and M. Harman, "A comprehensive empirical study of bias mitigation methods for machine learning classifiers," *ACM Trans. Softw. Eng. Methodol.*, vol. 32, no. 4, pp. 1–30, May 2023. [Online]. Available: https://doi.org/10.1145/3583561

[47] Y. Nie, A. S. Zamzam, and A. Brandt, "Resampling and data augmentation for short-term PV output prediction based on an imbalanced sky images dataset using convolutional neural networks," *Solar Energy*, vol. 224, pp. 341–354, Aug. 2021.

[48] Z. S. Rubaidi, B. B. Ammar, and M. B. Aouicha, "Fraud detection using large-scale imbalance dataset," *Int. J. Artif. Intell. Tools*, vol. 31, no. 8, 2022, Art. no. 2250037.

[49] R. Li, Y. Wang, F. Liang, H. Qin, J. Yan, and R. Fan, "Fully quantized network for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2810–2819.

[50] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and< 0.5 MB model size," 2016, *arXiv:1602.07360*.

[51] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11784–11793.

[52] L. Fan, F. Wang, N. Wang, and Z.-X. Zhang, "Fully sparse 3D object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 351–363.

[53] A. Ali, T. Schnake, O. Eberle, G. Montavon, K.-R. Müller, and L. Wolf, "XAI for transformers: Better explanations through conservative propagation," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 435–451.

[54] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-wise relevance propagation: An overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., Cham, Switzerland: Springer Int. Publ., 2019, pp. 193–209, doi: 10.1007/978-3-030-28954-6_10.

[55] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowl. Inf. Syst.*, vol. 33, no. 1, pp. 1–33, 2012.

[56] R. Keshari, M. Vatsa, R. Singh, and A. Noore, "Learning structure and strength of CNN filters for small sample size training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9349–9358.