Research Statement

Bishwamittra Ghosh

PhD Candidate
School of Computing
National University of Singapore (NUS)
https://bishwamittra.github.io

The last decade has witnessed significant progress in machine learning with a host of applications of algorithmic decision-making in different safety-critical domains, such as college admission, recidivism prediction, employee hiring, and clinical processes. In these domains, the decisions by machine learning classifiers have far-reaching consequences, potentially influencing human lives in various ways. Consequently, fairness and interpretability of the deployed classifiers have paramount importance to the end-users, who expect to have a fair decision from an algorithm and an interpretable decision that they can understand. My research vision is to improve the safety-critical properties of machine learning, such as fairness and interpretability, and enhance the applicability of machine learning for social good.

Towards **fair and interpretable machine learning**, my research so far has focused on fairness verification, identification of source of unfairness of classifiers and designing interpretable classifiers with higher scalability. Fairness verification gives a formal certificate of whether a classifier has achieved the desired level of fairness on specified data distribution. In identifying the source of unfairness, we compute fairness influence functions of input features as their contribution towards the bias/unfairness of the classifier. In interpretable machine learning, we focus on designing an interpretable rule-based classifier alternative to black-box classifiers. In particular, we propose a scalable learning framework to generate accurate and small interpretable rules while enabling learning on large datasets. In the technical aspects of my research, I have closely integrated formal methods with machine learning and designed frameworks that can withstand with improved scalability and performance guarantees.

My first research direction is on **fairness in machine learning**. Classifiers relying on merely accuracy-centric learning objectives may become unfair towards certain demographic groups in the data. To this end, my research addresses the fairness verification problem of group and causal fairness metrics, where the goal is to compute the unfairness of a classifier given the distribution of features. We propose a formal approach for fairness verification of linear classifiers [1] and classifiers represented as logical formulas [2]. Furthermore, we verify fairness by considering feature correlations represented as a Bayesian network. We have compared state-of-the-art fairness verifiers and demonstrate higher accuracy and scalability of our approach than the others. Next, we focus on computing fairness influence functions to quantify the contribution of input features on the incurred bias of the classifier [3]. In particular, we compute individual and inter-sectional influences to capture the dynamics of correlated features in the resultant bias. Based on global sensitivity analysis, our approach computes influences more accurately than the state-of-the-art. My second research direction is on **interpretable machine learning**, where the goal is to learn interpretable rule-based classifiers efficiently. Rule-based classifiers, such as decision trees, decision sets, and decision lists, are popular interpretable classifiers, and they can explain the inner working of black-box classifiers, such as neural networks. The challenge in interpretable rule-learning is to trade-off between prediction accuracy and generating the smallest classification rule to favor interpretability; thus, the underlying combinatorial optimization problem cannot scale to large datasets. Therefore, we propose an incremental learning framework based on optimization frame-

works such as MaxSAT [4, 5, 6] and MILP [7] solving to learn small and accurate classification rules in datasets containing a million samples. Our work has been published at premier conferences in artificial intelligence and machine learning (AAAI×3, AIES, ECAI).

References

[1] **Bishwamittra Ghosh**, Debabrota Basu, and Kuldeep S. Meel. Algorithmic fairness verification with graphical models. In *Proceedings of AAAI*, 2022.

[2] **Bishwamittra Ghosh**, Debabrota Basu, and Kuldeep S. Meel. Justicia: A stochastic SAT approach to formally verify fairness. In *Proceedings of AAAI*, 2021.

[3] **Bishwamittra Ghosh**, Debabrota Basu, and Kuldeep S. Meel. "How biased is your feature?": Computing fairness influence functions with global sensitivity analysis (under review). 2022.

[4] **Bishwamittra Ghosh** and Kuldeep S. Meel. IMLI: An incremental framework for MaxSAT-based learning of interpretable classification rules. In *Proceedings AIES*, 2019.

[5] **Bishwamittra Ghosh**, Dmitry Malioutov, and Kuldeep S. Meel. Efficient learning of interpretable classification rules (under review). 2022.

[6] Lorenzo Ciampiconi, **Bishwamittra Ghosh**, Jonathan Scarlett, and Kuldeep S. Meel. A MaxSAT-based framework for group testing. In *Proceedings of AAAI*, 2020.

[7] **Bishwamittra Ghosh**, Dmitry Malioutov, and Kuldeep S. Meel. Classification rules in relaxed logical form. In *Proceedings of ECAI*, 2020.