Research Statement

Bishwamittra Ghosh
Scientist
Institute of High Performance Computing (IHPC), A*STAR, Singapore
https://bishwamittra.github.io

My research is on fairness and explainability in machine learning applied in safety-critical domains. Traditional machine learning, particularly deep learning, is known for unfair predictions towards marginalized sensitive groups and for generating black-box predictions. In my research, I design algorithmic frameworks to *formally quantify fairness in machine learning* [1, 2], *explain the sources of unfairness* [3], and *learn explainable rule-based classifiers* [4, 5, 6]. Prior approaches to these problems are often limited by scalability, accuracy, or both. To address the limitations, I closely integrate automated reasoning, formal methods, and statistics with fairness and explainability to develop scalable and accurate solutions.

During my PhD, I have research collaborations and internships in academia and industry. In addition to fairness and explainability, I have collaborative research on group testing [7], social-spatial group queries [8, 9], and hypergraph core decomposition [10]. We publish our research in leading conferences and journals in artificial intelligence and machine learning: AAAI (2022, 2021, 2020), JAIR (2022), FAccT (2023), ECAI (2020), and AIES (2019); and databases: VLDB (2023, 2018) and TSAS (2022). I present a tutorial on *Auditing Bias of Machine Learning Algorithms: Tools and Overview* in IJCAI 2023. I was awarded the NUS Research Scholarship, Singapore and Moblilex Scholarship at Université de Lille, France.

Research Thrust 1: Fairness in Machine Learning

Fairness in machine learning focuses on quantifying and mitigating the bias or unfairness of the prediction of the classifier towards different sensitive groups in the data. To quantify bias in algorithmic decision-making, multiple fairness metrics have been proposed based on societal norms and beliefs. However, there has been insignificant progress in *formally quantifying existing fairness metrics.* In addition, fairness metrics measure the overall bias of a classifier, but they cannot *explain the sources of bias.* Therefore, our research focuses on two key aspects: formally quantifying bias of a classifier and explaining its sources.

Probabilistic Fairness Quantification

In probabilistic fairness quantification, we formally quantify the bias of a classifier given the distribution of input features—essentially beyond a finite dataset. We propose two approaches to the problem: a general approach for finite classifiers encoded as Boolean formulas [1] and a specific approach for linear classifiers [2].

Fairness Quantification via SSAT. The key idea in quantifying group fairness metrics is to compute the maximum (resp. minimum) probability of predictions of the classifier across all sensitive groups—the probability of selecting White-male vs. Black-female candidates in job applications. We propose a stochastic satisfiability (SSAT) based framework, called Justicia [1], for computing such probabilities. More specifically, the maximum probability becomes the solution of an existential-random (ER)-SSAT formula—we encode the classifier as a Boolean formula, the feature distribution via random Boolean variables, and compute the maximum conditional probability of the satisfaction of the formula for existentially quantified sensitive features. In the presence of multiple sensitive features resulting in exponentially many sensitive groups, SSAT efficiently finds the most (resp. least) favored group by the classifier, thanks to the progress in satisfiability (SAT) solving, and particularly in weighted model counting problem. In experiments, Justicia is more scalable in the fairness quantification of tree-based classifiers than existing SMT or sampling methods.

Tractable Fairness Quantification with Feature Correlation. We extend Justicia to consider feature correlations for an accurate fairness quantification [2]. We consider a Bayesian network to represent the conditional distribution of features—the SSAT formula grows with the complexity of the Bayesian network, calling for a more scalable solution. Therefore, we demonstrate a tractable fairness quantification for linear classifiers by proposing a stochastic subset sum problem, which admits an efficient dynamic programming solution with pseudo-polynomial complexity. Experimentally, Justicia becomes more accurate and scalable than existing fairness verifiers for linear classifiers.

Explaining Fairness: Identifying Sources of Bias

We combine both explainability and fairness in machine learning and propose a framework for explaining fairness. We formalize *fairness influence functions* (FIFs) to quantify the contribution of an individual feature and the intersection of multiple features to the resulting bias of the classifier [3]. Based on global sensitivity analysis, we propose a model-agnostic framework, called FairXplainer, to estimate FIFs. The key idea is to represent fairness metrics using the variance of predictions and apply variance decomposition to compute FIFs. In experiments, FIFs are highly correlated with fairness interventions and demonstrate a higher granular explanation of unfairness through intersectional influences, unlike existing local explainability method SHAP. In addition, FairXplainer approximates bias via FIFs with lower error than prior methods across classifiers such as neural networks and SVMs.

Research Thrust 2: Explainable Rule-based Machine Learning

We learn classifiers explainable by design, such as rule-based classifiers. In rule-based classifiers, for example decision lists and decision sets, the decision boundary is explained using a set of rules relating input features to class prediction. The explainability of such classifiers often depends on the size of the rules—smaller rules with higher accuracy are preferred in practice, particularly by practitioners in the medical domain. Our contributions in rule-based classification are two-folds: a scalable learning framework for classification rules by incremental learning and an improvement of the expressiveness of rules via logical relaxation.

Scalability via Incremental Learning

We introduce an incremental learning framework based on MaxSAT, called IMLI [4, 5], to learn explainable classification rules in propositional logic, particularly in CNF. The CNF learning framework can potentially learn other explainable representations: decision sets, decision lists etc. We design a MaxSAT formulation to jointly optimize the accuracy and explainability of CNF classifiers, and leverage the progress in MaxSAT solving to efficiently learn an optimal classifier. However, the MaxSAT formula grows with dataset dimension and classifier size. To improve scalability, IMLI integrates both mini-batch learning and iterative rule-learning: IMLI learns a CNF classifier by iteratively covering the training data, where in each iteration IMLI solves a sequence of smaller MaxSAT queries respective to mini-batches. In experiments, IMLI achieves the best balance among prediction accuracy, explainability, and scalability, for example, a competitive accuracy and explainability compared to existing rule-based classifiers, and a higher scalability on large datasets with a million samples where explainable and non-explainable classifiers may fail.

Expressiveness via Logical Relaxation

Rule-based classifiers are explainable by design, but they are less expressive. We propose a more expressible yet explainable rule-based classifier, called relaxed-CNF [6], based on a relaxed definition of the standard OR/AND operators in logic. Motivated by checklists in the medical domain such as CHADS$_2$ score, in relaxed-CNF, both the minimum number of literals satisfied in a clause and the maximum number of clauses satisfied in a formula are flexible. As a result, relaxed-CNF generalizes widely used rule representations: CNF, DNF, decision lists, and decision sets. While the combinatorial structure of relaxed-CNF results in exponential succinctness, the direct learning technique is computationally expensive. Therefore, we extend IMLI and propose an incremental mini-batch learning procedure for relaxd-CNF classifiers, called CRR, by leveraging advances in MILP solving. In experiments, CRR generates more accurate yet smaller relaxed-CNF rules compared to alternative rule-based classifiers.

Future Research Plans

My future research is dedicated to developing practical and scalable algorithms for trustworthy machine learning. Machine learning and artificial intelligence have been compared to the new electricity, with the potential to transform various aspects of human life, evident from the overwhelming response to generative AI. Ensuring fairness and explainability in deployed machine learning is now more necessary than ever. To accomplish this, I aim to work in a collaborative environment, gaining insights into real-world challenges and leveraging advances in the field, alongside formal methods, to make significant progress. I have identified key research themes that will guide my work towards this vision.

Fairness and Explainability As a Service. The goal of modern machine learning extends beyond learning patterns from large-scale historical data to ensuring responsible decision-making through careful regulation to establish trustworthiness. For instance, in a job application scenario, a machine learning algorithm must be fair across different demographic groups, resilient to non-actionable changes in candidate profiles, and explainable to allow candidates to understand the decision-making process. My long-term research goal is to offer fairness and explainability as a service with machine learning-based decision-making. Below, I outline several research ideas concerning fairness and explainability:

- **Fairness Auditing.** Our objective is to develop a comprehensive fairness auditing framework for machine learning, focusing on three key questions. (i) *Which fairness metrics to choose?* We aim to identify the most appropriate fairness metrics for specific application contexts, as choosing the right metric is crucial among various notions of fairness. (ii) *How to quantify bias?* We extend formal fairness quantification to encompass broader fairness metrics, including individual fairness, causal fairness, and counterfactual fairness. We apply this extension to unstructured data such as images and texts, and classifiers such as random forests, neural networks, and language models. (iii) *How to explain bias?* We strive to advance our fairness explaining framework FairXplainer to explain bias for both texts and images. For instance, in conversational AI, we intend to highlight the input prompts that trigger biased statements generation by the model. By addressing these three questions, our vision is to design improved bias mitigating algorithms with significant practical impacts.

- **Explainability with Guarantees.** Our research in explainable machine learning spans two main directions. (i) *Explainability by design:* There is a growing interest for explainable machine learning in safety-critical domains, for example, clinical predictions, financial decisions, and self-driving vehicles. Building upon our explainable rule-based classifier IMLI, we aim to enhance learning algorithms for explainable models in large-scale datasets across supervised, semi-supervised, and unsupervised settings. (ii) *Post-hoc explainability:* To explain black-box predictions, we focus on explanations with formal guarantees. For example, an explanation model must be robust, learned in a privacy-preserving manner, and provide the confidence level of explanations to increase transparency and trust in the decision-making process.

Verifiable Machine Learning with Formal Methods. In safety-critical and high-stake domains, the verification of machine learning before deployment is crucial. To this end, formal methods provide a template to verify different checkable properties of machine learning. In particular, SAT, SMT and their variants allow to concentrate on synthesizing constraints from real-world use cases and delegate the solution finding to respective solvers, thanks to the dedicated community in formal methods. Building upon my ongoing research, which involves applying SSAT and MaxSAT to address fairness and explainability challenges, I plan to explore additional formulations in formal methods, such as functional analysis, abstract interpretation, and solvers with expressive theory, to further enhance the verification of machine learning models.

Benchmarking Machine Learning: Improvement in Formal Methods. Interdisciplinary research is an ongoing process that yields valuable contributions to different disciplines. Our work on fairness and explainability in machine learning has had a positive impact on the improvement of solvers in formal methods. For instance, when tackling explainable classification problems, we realized that MaxSAT alone was insufficient for large-scale rule-based classification, prompting the need for an incremental MaxSAT solver. As a result, we contribute to the MaxSAT evaluation competition in 2019 with our MaxSAT benchmarks of explainable classification. Subsequently, the competition introduced an incremental MaxSAT solving track. This observation serves as strong motivation for me to continue contributing to the formal methods community by creating additional machine learning verification benchmarks.

References

[1] B. Ghosh, D. Basu, and K. S. Meel, "Justicia: A stochastic SAT approach to formally verify fairness," in *Proc. of AAAI*, 2021.

[2] B. Ghosh, D. Basu, and K. S. Meel, "Algorithmic fairness verification with graphical models," in *Proc. of AAAI*, 2022.

[3] B. Ghosh, D. Basu, and K. S. Meel, "How biased are your features?: Computing fairness influence functions with global sensitivity analysis," in *Proc. of FAccT*, 2023.

[4] B. Ghosh, D. Malioutov, and K. S. Meel, "Efficient learning of interpretable classification rules," in *Proc. of JAIR*, 2022.

[5] B. Ghosh and K. S. Meel, "IMLI: An incremental framework for MaxSAT-based learning of interpretable classification rules," in *Proc. of AIES*, 2019.

[6] B. Ghosh, D. Malioutov, and K. S. Meel, "Classification rules in relaxed logical form," in *Proc. of ECAI*, 2020.

[7] L. Ciampiconi, B. Ghosh, J. Scarlett, and K. S. Meel, "A MaxSAT-based framework for group testing," in *Proc. of AAAI*, 2020.

[8] B. Ghosh, M. E. Ali, F. M. Choudhury, S. Hasan, T. Sellis, and J. Li, "The flexible socio spatial group queries," in *Proc. of VLDB*, 2018.

[9] S. H. Apon, M. E. Ali, B. Ghosh, and T. Sellis, "Social-spatial group queries with keywords," *ACM Transactions on Spatial Algorithms and Systems (TSAS), 2021*, 2021.

[10] N. A. Arafat, A. Khan, A. K. Rai, and B. Ghosh, "Neighborhood-based hypergraph core decomposition," *Proc. of VLDB*, 2023.