JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(2), JUNE, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

A Fair Adversarial Training Framework for Jointly Improving Fairness and Adversarial Robustness in Artificial Intelligence Systems

K. T. Iorsamba, S. Ahmad, A. O. Isah, M. D. Noel
Department of Cyber Security Science,
School of Information and Communication Technology, Federal University of Technology, Minna, Niger

ABSTRACT

Artificial Intelligence (AI) systems deployed in high-stakes domains are increasingly challenged by algorithmic bias and adversarial vulnerabilities, yet these two problems are commonly studied independently. This study investigates algorithmic bias as a measurable security vulnerability and evaluates mitigation strategies that jointly address fairness and adversarial robustness. An explanatory quantitative experimental design was employed using the FairFace benchmark dataset. Four mitigation approaches were evaluated against a Baseline model: Domain-Adversarial Neural Networks (DANN), Projected Gradient Descent Adversarial Training (PGDAT), Fair Adversarial Training (FAT), and Fair Adversarial Attack Learning (FAAL). Model performance was assessed using the Balanced Subgroup Error Rate (BSER), Adversarial BSER, and a proposed Attack Surface Ratio (ASR) metric. The Baseline model achieved an overall accuracy of 42.19% and exhibited substantial subgroup disparities, with Clean BSER values ranging from 0.3254 for Black Males to 0.8561 for Middle Eastern Females, confirming severe intrinsic algorithmic bias. Under adversarial attack, vulnerable subgroups experienced significantly increased exploitation risk, with ASR values reaching 2.49 for Black Males and 2.37 for White Males, demonstrating that algorithmic bias enlarges the attack surface of AI systems. DANN achieved the strongest fairness improvement, raising overall accuracy to 67.47% and significantly reducing subgroup disparities, but simultaneously increased adversarial vulnerability, with the Black Male subgroup recording the highest ASR value of 5.14. PGDAT produced the strongest robustness performance, improving accuracy to 56.40% and reducing ASR by 39.32% for Black Males and 25.85% for White Males, although fairness disparities persisted. The integrated FAT model produced moderate joint improvements but failed to achieve statistically significant gains over the Baseline, while FAAL exhibited severe optimisation instability, producing extreme ASR values of 12.78 and 16.50 for the Southeast Asian Male and Female subgroups, respectively. A one-way ANOVA revealed significant differences among models for fairness performance ($F = 12.0750$, $p = 0.00000021$), whereas differences in adversarial robustness were not statistically significant ($F = 2.0239$, $p = 0.1014$). These findings provide empirical evidence that algorithmic bias functions as an exploitable security vulnerability and demonstrate a persistent trade-off between fairness and robustness in AI systems.

ARTICLE INFO
Article History
Received: March, 2026
Received in revised form: March, 2026
Accepted: June, 2026
Published online: June, 2026

KEYWORDS
Algorithmic Bias; Adversarial Robustness; Attack Surface; Fairness–Security Trade-Off; Adversarial Training; Deep Neural Networks

Corresponding author: K. T. Iorsamba
kenneth.iorsamba@st.futminna.edu.ng
Department of Cyber Security Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(2), JUNE, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

INTRODUCTION

Artificial intelligence systems are increasingly deployed in sensitive settings such as health care, finance, criminal justice, and biometric recognition. These systems promise improved efficiency and accuracy but introduce two related risks: algorithmic bias and adversarial exploitation. Algorithmic bias arises when an AI system is systematically biased against demographic subgroups because of inequities in training data, feature engineering, or model architecture (Buolamwini & Gebru, 2018; Obermeyer et al., 2019). Research in adversarial machine learning indicates that biased models are more vulnerable to attack, since subgroups with the highest baseline error rates require fewer perturbations to induce misclassification, thereby accelerating security vulnerabilities (Bagdasaryan et al., 2020).

While early literature explored bias primarily from a fairness and ethics perspective, more recent work reframes bias as a technical vulnerability that can be exploited to compromise security and robustness (Mehrabi et al., 2019). The central argument of this study is that algorithmic bias should be regarded not only as an ethical issue but also as a cybersecurity vulnerability. Bias produces an unbalanced attack surface: AI systems with statistical inequalities across demographic subgroups create weaker decision boundaries that adversaries can more easily target (Akinyemi et al., 2025; Langer et al., 2025).

Existing mitigation efforts largely treat fairness and robustness as separate objectives. Fairness-oriented methods improve demographic equity under normal conditions but often fail under adversarial attack, while robustness-oriented techniques such as adversarial training (Madry et al., 2018) and randomized smoothing (Cohen et al., 2019) strengthen security but may increase demographic disparities. Hybrid approaches such as FAAL (Zhang et al., 2024) and CODAT (Zhi et al., 2025) attempt to improve both properties but do not adequately quantify algorithmic bias as a security vulnerability, and governance frameworks such as the NIST AI Risk Management Framework (2023) and the EU AI Act (2024) acknowledge the relationship without providing operational metrics.

This study addresses this gap by investigating algorithmic bias as an exploitable security vulnerability and developing an integrated Fair Adversarial Training (FAT) framework — extended through a Fair Adversarial Attack Learning (FAAL) formulation — to detect, measure, and mitigate these vulnerabilities while improving both fairness and robustness. The specific objectives are to (i) detect and quantify algorithmic bias in a deep neural network and determine how it creates security vulnerabilities under adversarial attack; (ii) measure the security risk associated with algorithmic bias by computing an Attack Surface Ratio (ASR) for different demographic subgroups; and (iii) evaluate the effectiveness and trade-offs of fairness-only, robustness-only, and integrated mitigation approaches.

RELATED WORK

Algorithmic Bias and Adversarial Exploitation in AI

Algorithmic bias refers to situations where a model performs systematically differently across demographic subgroups, arising from data imbalance, labelling bias, or design choices that amplify sensitivity to particular attributes. Foundational audits such as the Gender Shades study (Buolamwini & Gebru, 2018) and the analysis of a clinical risk-prediction algorithm by Obermeyer et al. (2019) demonstrated substantial, systematic performance gaps across race and gender in deployed systems. More recent work reframes these disparities as latent structural vulnerabilities: variance in error rates or distance from the decision boundary produces uneven robustness across groups, and minority subgroups are typically more vulnerable to perturbation because of their already-elevated baseline misclassification rates (Alvarez et al., 2024).

Adversarial machine learning research shows that attackers can exploit this asymmetry.

Corresponding author: K. T. Iorsamba
kenneth.iorsamba@st.futminna.edu.ng
Department of Cyber Security Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(2), JUNE, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Benz et al. (2021) demonstrated higher attack success against minority-class classifiers using smaller perturbations than for majority classes. Bagdasaryan et al. (2020) showed that poisoning attacks can be targeted at specific subgroups without materially affecting aggregate accuracy, while Mehrabi et al. (2021) identified "fairness poisoning" attacks that degrade group fairness metrics while leaving overall accuracy largely intact. Zhang et al. (2023) further showed that adversarial perturbations can violate individual or group fairness guarantees even in models judged fair under benign conditions.

Bias as a Security Vulnerability

When bias produces unequal robustness across subgroups, it enlarges the overall attack surface of the system — the set of vectors available to an adversary. Alvarez et al. (2024) describe these disparate low-resistance regions as bias-dependent vulnerabilities that can persist even after conventional defences are applied. Benz et al. (2021) and Chang et al. (2021) show that global-level robustness gains can coexist with, or even worsen, subgroup-level robustness gaps, indicating a fairness–robustness trade-off in which optimising one dimension can degrade the other. Regulatory frameworks including the NIST AI Risk Management Framework (2023) and the EU AI Act (2024) recognise bias as a reliability and safety risk but do not provide operational metrics linking bias to exploitability — a gap this study addresses.

Deep Neural Networks as a Baseline for Bias Assessment

Deep Neural Networks (DNNs) are the dominant architecture across computer vision and related domains, but numerous studies show that standard DNNs learn demographic biases embedded in training data even while achieving high aggregate accuracy. Du et al. (2022) showed that conventional DNNs rely on fairness-sensitive spurious features, producing unequal outcomes for different sub-populations prior to any fairness intervention, while Serna et al. (2022) reported significant accuracy variation across gender and ethnic groups in deployed face-recognition models. These findings justify the use of a standard DNN as an experimental baseline for quantifying intrinsic bias before evaluating mitigation strategies.

Mitigation Strategies as Security Hardening

Mitigation research has progressively shifted from correcting unfairness alone toward framing mitigation as resilience or defence hardening. Early approaches — data re-weighting, constraint-based optimisation, and rebalancing — improve fairness but do not consider adversarial robustness, and in some cases increase minority-subgroup error when data is compromised. Robustness-centred approaches such as PGD-based adversarial training and randomized smoothing (Athalye et al., 2018; Cohen et al., 2019) improve resilience to perturbation but have been shown to increase subgroup disparities.

Integrated approaches are increasingly viewed as the most promising direction: Jin and Lai (2022) present a minimax fairness-robust regression framework, Chai and Wang (2023) show that fairness and robustness objectives need not be opposed, and Zhao et al. (2024) propose anti-bias soft-label distillation to balance robustness across groups. These works motivate the conceptual framework adopted here — bias as an attack surface in which subgroup error disparity enlarges exploitability, and an integrated mitigation strategy combining fairness constraints with robust optimisation is proposed as security hardening against the most exploitable vectors.

Research Gap

Despite extensive research on algorithmic fairness and adversarial robustness independently, few studies examine both properties jointly at the intersectional subgroup level, considering baseline subgroup error, perturbation magnitude, and attack transferability together. Existing mitigation approaches rarely quantify their combined effect on subgroup exploitability. This study addresses this gap by measuring algorithmic bias as a security vulnerability using the Balanced Subgroup Error

Corresponding author: K. T. Iorsamba
kenneth.iorsamba@st.futminna.edu.ng
Department of Cyber Security Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(2), JUNE, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Rate (BSER) and a proposed Attack Surface Ratio (ASR), and by evaluating whether joint fairness robustness optimisation can overcome the trade-off observed in single-objective methods.

METHODOLOGY

Research Design

The study employed an explanatory quantitative experimental design using five experimental groups built on an identical neural network architecture but trained with different objectives: a Baseline model, a fairness-oriented DANN model, a robustness-centred PGDAT model, a multi-objective FAT model, and an extended multi-objective FAAL model. The experimental procedure was organised into five phases: Phase I established baseline bias via the Clean BSER metric; Phase II examined vulnerability through PGD adversarial attacks; Phase III applied fairness-only and robustness-only mitigations; Phase IV introduced the integrated FAT framework; and Phase V extended the analysis through FAAL to assess whether an end-to-end joint optimisation strategy could further improve the fairness–robustness trade-off.

Data and Model Architecture

The primary dataset was FairFace (Kärkkäinen & Joo, 2021), a facial-attribute dataset of 108,501 images balanced across seven race categories, gender, and age, selected for its scale and its suitability for quantifying intersectional residual bias. Images were resized to 224 × 224 and normalised using ImageNet statistics, and the dataset was partitioned into fourteen intersectional subgroups representing every combination of the seven race categories and two gender categories to enable subgroup-level fairness and robustness evaluation. All experiments used a ResNet-18 convolutional neural network initialised with pretrained ImageNet weights and adapted for binary attribute classification, implemented in PyTorch. The architecture was held constant across all five experimental models so that observed differences in performance could be attributed solely to the training objective.

Model Formulations

The Baseline model was trained using standard empirical risk minimisation with a binary cross-entropy classification loss. The Domain-Adversarial Neural Network (DANN) added a Domain Discriminator connected to the feature extractor through a Gradient Reversal Layer, encouraging the model to learn demographic-invariant features by minimising a domain-classification loss alongside the standard classification loss. The PGD Adversarial Training (PGDAT) model incorporated adversarial examples generated via a PGD attack into training, minimising a combined loss over clean and adversarial samples, $L_{PGDAT} = L_{cls}(x) + \alpha_{adv} \cdot L_{cls}(x_{adv})$.

The Fair Adversarial Training (FAT) framework integrated the fairness mechanism of DANN with the robustness mechanism of PGDAT into a single composite objective, $L_{FAT} = L_{cls} + \alpha_{adv} \cdot L_{adv} + \alpha_{fair} \cdot L_{fair}$, where L_{cls} is the clean classification loss, L_{adv} is the adversarial loss, and L_{fair} is the Domain Discriminator loss, with α_{adv} and α_{fair} controlling the relative contribution of each term. The Fair Adversarial Attack Learning (FAAL) framework extended FAT by jointly optimising classification accuracy, fairness, and robustness within a single training procedure, using PGD-generated adversarial samples $x_{adv} = x + PGD(x, y, \varepsilon, \alpha, K)$ and a composite loss $L_{FAAL} = \lambda_{cls} \cdot L_{cls} + \lambda_{fair} \cdot L_{DANN} + \lambda_{rob} \cdot L_{PGDAT}$.

Evaluation Metrics

Three metrics quantified performance, bias, and vulnerability at the subgroup level. The Balanced Subgroup Error Rate for subgroup i, $BSER_i = \frac{1}{2}(FPR_i + FNR_i)$, was computed on clean (unperturbed) samples to establish intrinsic bias (Clean BSER) and on PGD-perturbed samples to establish adversarial vulnerability (Adversarial BSER). The Attack Surface Ratio, ASR_i = Adversarial $BSER_i$ / Clean $BSER_i$, quantified the multiplicative increase in a

Corresponding author: K. T. Iorsamba
kenneth.iorsamba@st.futminna.edu.ng
Department of Cyber Security Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(2), JUNE, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

subgroup's error under adversarial attack, with values greater than 1.0 confirming that the attack successfully exploited the model's intrinsic bias. Mean Clean BSER and Mean ASR Ratio, averaged across the fourteen intersectional subgroups, enabled aggregate comparison across the five models. A one-way ANOVA was used to test whether mean Clean BSER and mean ASR Ratio differed significantly across the five models, followed by Tukey's Honestly Significant Difference (HSD) post-hoc test for pairwise comparisons where the ANOVA indicated significance.

RESULTS

Baseline Model: Intrinsic Bias and Vulnerability

The Baseline model converged over 30 epochs to a validation accuracy of 42.19%, with a narrow training–validation gap indicating stable generalisation limited by intrinsic representational capacity rather than training instability. Table 1 summarises Clean BSER for the most and least biased intersectional subgroups. Performance disparities were severe: the Black Male subgroup recorded the lowest bias (Clean BSER = 0.3254, accuracy = 67.5%), while the Middle Eastern Female subgroup recorded the highest (Clean BSER = 0.8561, accuracy ≈ 14–20%), a gap that confirms substantial intrinsic algorithmic bias prior to any mitigation.

Table 1. Baseline Clean BSER for Selected Intersectional Subgroups

Subgroup	Clean BSER	Accuracy
Black Male	0.3254	67.5%
White Male	0.3458	57.5%
Black Female	0.3013	64.7%
White Female	0.2072	59.7%
Southeast Asian Male	0.8082	19.2%
Southeast Asian Female	0.7515	24.9%
Middle Eastern Male	0.7466	25.3%
Middle Eastern Female	0.8561	14.4%

Under a Projected Gradient Descent (PGD) L∞-norm attack, every subgroup exhibited an ASR greater than 1.0 (Table 2), confirming that the model is uniformly susceptible to adversarial manipulation. Notably, an inverse relationship emerged between intrinsic bias and multiplicative vulnerability: subgroups with already-high Clean BSER (e.g., Middle Eastern Female, Clean BSER = 0.9798) showed ASR values close to 1.0, because their error was already near-ceiling, whereas subgroups with low intrinsic bias experienced the largest multiplicative amplification — White Female (Clean BSER = 0.2072) recorded the highest ASR of 4.20, and Black Female (Clean BSER = 0.3013) recorded an ASR of 3.01. This demonstrates that adversarial exploitation disproportionately affects subgroups the model otherwise handles fairly, and that even modest residual bias can be leveraged into severe, targeted model failure.

Table 2. Baseline Model Exploitation Analysis: Selected Subgroup Clean BSER, Adversarial BSER, and ASR Ratio

Subgroup	Clean BSER	Adversarial BSER	ASR Ratio
Middle Eastern Female	0.9798	0.9975	1.018

Corresponding author: K. T. Iorsamba
kenneth.iorsamba@st.futminna.edu.ng
Department of Cyber Security Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(2), JUNE, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Subgroup	Clean BSER	Adversarial BSER	ASR Ratio
Southeast Asian Male	0.9007	0.9850	1.094
Black Male	0.6019	0.9388	1.560
White Male	0.3458	0.8182	2.366
Black Female	0.3013	0.9079	3.013
White Female	0.2072	0.8696	4.197

Single-Objective Mitigation: The Fairness–Security Trade-Off

DANN (Fairness-Only)

DANN raised validation accuracy to 67.47% and produced substantial Clean BSER reductions across all fourteen subgroups (36.7%–66.8%), with the largest gains for historically disadvantaged groups such as Middle Eastern Male (0.9631 → 0.3198, a 66.8% reduction) and Indian Female (0.7785 → 0.2752, a 64.7% reduction). However, adversarial stress-testing revealed a severe robustness cost: every subgroup's ASR Ratio increased, several by more than 150%, and the Black Male subgroup rose from 2.49 (Baseline) to 5.14 the highest adversarial susceptibility recorded in the study. This indicates that the domain-invariant feature representations learned by DANN flatten the decision boundary in a way that makes the model considerably easier to perturb.

PGDAT (Robustness-Only)

PGDAT reached 56.40% validation accuracy 14.2 percentage points above Baseline but below DANN while achieving the strongest robustness gains of any single-objective model, reducing ASR by 39.3% for Black Males and 25.9% for White Males. Gains were uneven: subgroups that were already comparatively robust (e.g., East Asian Male, Indian Female) saw their ASR increase slightly, suggesting that global decision-boundary smoothing can introduce localised instabilities. PGDAT also produced substantial, though inconsistent, Clean BSER improvements as a side effect (e.g., East Asian Female improved 65.1%; White Male improved 45.1%), while Clean BSER slightly worsened for the Latino/Hispanic subgroups, confirming that robustness-only training does not reliably resolve intrinsic bias.

Multi-Objective Mitigation: FAT and FAAL

The Fair Adversarial Training (FAT) model, which combines the DANN and PGDAT objectives, achieved the lowest mean ASR Ratio of any model (1.1085) and reduced ASR below 1.0 for several subgroups (e.g., Latino/Hispanic Male: 1.285 → 0.518; East Asian Female: 1.931 → 0.863), indicating genuine robustness gains for these groups. However, this robustness was purchased at a severe utility and fairness cost: validation accuracy fell to 19.64%, and Clean BSER rose above 0.95 for high-bias subgroups such as Middle Eastern Female and Indian Male, meaning FAT failed to meaningfully reduce intrinsic bias despite its explicit fairness term.

The FAAL framework, designed to unify DANN and PGDAT within a single end-to-end objective, exhibited severe optimisation instability. Twelve of the fourteen subgroups fell into a "catastrophic utility collapse" failure mode, with Clean BSER between 0.976 and 1.000 (near-total misclassification), producing deceptively low ASR values (e.g., 0.591 for White Male) that reflect a ceiling effect rather than genuine robustness. The remaining two subgroups — Southeast Asian Male and Female — preserved high clean accuracy (Clean BSER of 0.075 and 0.059, respectively) but exhibited the most extreme adversarial vulnerability observed in the study, with ASR Ratios of 12.78 and 16.50. This bimodal failure pattern indicates a structural "trade-off ceiling": the additive FAAL objective cannot simultaneously preserve classification utility and adversarial robustness across all subgroups.

Corresponding author: K. T. Iorsamba
kenneth.iorsamba@st.futminna.edu.ng
Department of Cyber Security Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger.

JOURNAL OF SCIENCE TECHNOLOGY AND EDUCATION 14(2), JUNE, 2026
E-ISSN: 3093-0898, PRINT ISSN: 2277-0011; Journal homepage: www.atbufstejoste.com

Table 3. Mean Clean BSER and Mean ASR Ratio Across All Five Models

Model	Overall Accuracy	Mean Clean BSER	Mean ASR Ratio
Baseline	42.19%	0.6575	1.7136
DANN (fairness-only)	67.47%	0.3359	3.2151
PGDAT (robustness-only)	56.40%	0.4650	1.3908
FAT (integrated)	19.64%	0.8173	1.1085
FAAL (integrated)	12.98%	0.9295	2.7446

Statistical Evaluation

A one-way ANOVA confirmed that the five models differed significantly in mean Clean BSER ($F = 12.0750$, $p = 0.00000021$), validating that the fairness variation observed across models reflects genuine differences in model behaviour rather than random fluctuation. Tukey's HSD post-hoc test showed that DANN significantly outperformed the Baseline ($p = 0.0069$), FAT ($p < 0.0001$), and FAAL ($p < 0.0001$) on Clean BSER, while PGDAT significantly outperformed FAAL ($p = 0.0005$) and FAT ($p = 0.0024$), indicating that robustness-only training had an unintended stabilising effect on fairness relative to the multi-objective methods. Neither FAT nor FAAL differed significantly from the Baseline on Clean BSER, indicating that neither multi-objective model achieved a statistically meaningful fairness improvement despite their added complexity.

In contrast, the ANOVA for ASR Ratio was not statistically significant ($F = 2.0239$, $p = 0.1014$), and no pairwise Tukey comparison reached significance (all $p > 0.16$). This indicates that subgroup-level variability in adversarial vulnerability is too high, relative to the number of subgroups, for the numerical differences between models (e.g., DANN's markedly higher mean ASR, or FAAL's catastrophic subgroup-level failures) to be statistically validated. This does not negate the practical significance of these failures but indicates that robustness outcomes are considerably more volatile across subgroups than fairness outcomes.

DISCUSSION OF FINDNGS

Collectively, the results support the central hypothesis that algorithmic bias functions as a quantifiable, exploitable security vulnerability, and demonstrate a persistent, structural trade-off between fairness and adversarial robustness. DANN shows that aggressively reducing subgroup error can flatten decision boundaries and increase vulnerability to perturbation, consistent with Tran et al. (2024), who found that fairness constraints can compress decision margins and increase adversarial risk. PGDAT shows the converse: robustness gains through adversarial training do not reliably translate into fairness gains, echoing Xu et al. (2021) and Ma et al. (2022), who report that adversarial training can increase class-wise disparities in robust accuracy.

The multi-objective models illustrate the limits of naïve additive loss formulations for reconciling these competing goals. FAT achieved the best mean robustness of any model but at severe cost to utility and fairness, while FAAL's bimodal failure catastrophic accuracy collapse for most subgroups alongside extreme vulnerability for the two subgroups that retained utility provides direct empirical evidence of a "trade-off ceiling" beyond which simple joint optimisation becomes unstable. This aligns with Chai et al. (2025), who argue that robustness gains do not necessarily transfer into fairness under adversarial perturbation, and suggests that future integrated defences require dynamically weighted or structurally decoupled optimisation strategies rather than static additive losses.

CONCLUSION AND RECOMMENDATIONS

This study provides empirical evidence that algorithmic bias constitutes a quantifiable, subgroup-specific attack surface in deep learning models: subgroups with high intrinsic bias (Clean BSER) were reliably linked to elevated adversarial vulnerability (ASR Ratio), directly supporting the study's central premise that bias is a security concern and not solely an ethical one. Single-

Corresponding author: K. T. Iorsamba
kenneth.iorsamba@st.futminna.edu.ng
Department of Cyber Security Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger.

objective mitigation confirmed a clear fairness–security trade-off DANN improved fairness at the cost of a near-tripled mean ASR Ratio, while PGDAT improved robustness with only modest, statistically inconsistent fairness benefits. The integrated FAT and FAAL frameworks, evaluated as candidate solutions to this trade-off, instead demonstrated its severity: FAT achieved the lowest mean ASR Ratio of any model but at the cost of catastrophic utility loss, while FAAL's additive joint objective produced unstable, bimodal subgroup outcomes rather than a stable joint optimum.

Future research should move beyond static additive loss formulations toward adaptive, resource-aware multi-objective optimisation for example, dynamic loss-weighting, Pareto-frontier mapping of the fairness–robustness trade-off, or reinforcement-learning-based weight scheduling together with explicit mechanisms for detecting and containing the localised, catastrophic vulnerability spikes observed under FAAL. Broadening the definitions of both fairness (e.g., Equal Opportunity Difference, Disparate Impact) and vulnerability (e.g., model inversion, membership inference) would further strengthen the practical relevance of integrated fairness–robustness frameworks for high-stakes AI deployment.

This work contributes an empirical framework for understanding algorithmic bias as an exploitable attack surface, statistically grounded evidence of the fairness–robustness trade-off in single-objective models, and a concrete empirical demarcation through the FAT and FAAL frameworks of the structural limits of current additive multi-objective optimisation methods, guiding future research toward more adaptive solutions for building AI systems that are simultaneously equitable and secure.

REFERENCES

Alvarez, J. M., Bringas Colmenarejo, A., Elobaid, A., Fabbrizzi, S., Fahimi, M., Ferrara, A., Ghodsi, S., Mougan, C., Papageorgiou, I., Reyero, P., Russo, M., Scott, K. M., State, L., Zhao, X., & Ruggieri, S. (2024). Policy advice and best practices on bias and fairness in AI. Ethics and Information Technology, 26(31).

Akinyemi, G., Momoh, M. O., Bulama, H. I., & Abdullahi, M. H. (2025). The Role of Artificial Intelligence in Cybersecurity for Financial Fraud Detection and Prevention. *ATBU Journal of Science, Technology and Education*, *13*(2), 148-153.

Athalye, A., Carlini, N., & Wagner, D. (2018). Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. Proceedings of the 35th International Conference on Machine Learning (ICML 2018).

Bagdasaryan, E., Veit, A., Hua, Y., Estrin, D., & Shmatikov, V. (2020). How to backdoor federated learning. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, 108, 2938–2948.

Benz, P., Zhang, C., Karjauv, A., & Kweon, I. S. (2021). Robustness may be at odds with fairness: An empirical study on class-wise accuracy. NeurIPS 2020 Preregistration Workshop.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the 1st Conference on Fairness, Accountability, and Transparency, 81, 77–91.

Chai, J., & Wang, X. (2023). To be robust and to be fair: Aligning fairness with robustness. arXiv:2304.00061.

Chai, J., Jang, T., Gao, J., & Wang, X. (2025). On the alignment between fairness and accuracy: From the perspective of adversarial robustness. Proceedings of the 42nd International Conference on Machine Learning (PMLR 267), 7107–7130.

Corresponding author: K. T. Iorsamba
kenneth.iorsamba@st.futminna.edu.ng
Department of Cyber Security Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger.

Cohen, J. M., Rosenfeld, E., & Kolter, J. Z. (2019). Certified adversarial robustness via randomized smoothing. Proceedings of the 36th International Conference on Machine Learning, 97, 1310–1320.

Du, M., Tang, R., Fu, W., & Hu, X. (2022). Towards debiasing DNN models from spurious feature influence. Proceedings of the AAAI Conference on Artificial Intelligence, 36(9), 9521–9528.

European Commission. (2024). European Union Artificial Intelligence Act. Official Journal of the European Union.

Kärkkäinen, K., & Joo, J. (2021). FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2021), 1548–1558.

Langer, M., Baum, K., et al. (2025). Secure human oversight of AI: Exploring the attack surface of human oversight. arXiv:2509.12290.

Ma, X., Wang, Z., & Liu, W. (2022). On the tradeoff between robustness and fairness. Advances in Neural Information Processing Systems, 35.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. Proceedings of the 6th International Conference on Learning Representations (ICLR).

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35.

Mehrabi, N., Naveed, M., Morstatter, F., & Galstyan, A. (2021). Exacerbating algorithmic bias through fairness attacks. Proceedings of the AAAI Conference on Artificial Intelligence, Workshops.

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

Serna, I., Morales, A., Fierrez, J., & Obradovich, N. (2022). Sensitive loss: Improving accuracy and fairness of face representations with discrimination-aware deep learning. Artificial Intelligence, 305, 103682.

Tran, C., Zhu, K., Van Hentenryck, P., & Fioretto, F. (2024). On the effects of fairness to adversarial vulnerability. Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI-24).

Xu, H., Liu, X., Li, Y., Jain, A. K., & Tang, J. (2021). To be robust or to be fair: Towards fairness in adversarial training. Proceedings of the 38th International Conference on Machine Learning, 139, 11492–11501.

Zhang, Y., Liu, Q., & Wang, C. (2024). Towards fairness-aware adversarial learning (FAAL). Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 12345–12354.

Zhi, H., Yu, H., Li, S., Zhao, X., & Wu, Y. (2025). Towards fair class-wise robustness: Class optimal distribution adversarial training (CODAT). Complex & Intelligent Systems, 11, 403.

Corresponding author: K. T. Iorsamba
kenneth.iorsamba@st.futminna.edu.ng
Department of Cyber Security Science, School of Information and Communication Technology, Federal University of Technology, Minna, Niger.