From Blindness to Robustness: Analyzing and Mitigating Asymmetric Label Noise in YOLOv7 Pedestrian Detection (December 2025)

FARAZ HEYDAR[1], HESAMEDDIN BITARAFAN RAJABI[2], AND GARIMA BAJWA.[3]

[1]Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada (e-mail: fheydar3@lakeheadu.ca).
[2]Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada (e-mail: hbitaraf@lakeheadu.ca).
[3]Department of Computer Science, Lakehead University, 955 Oliver Road, Thunder Bay, ON P7B5E1, Canada (e-mail: gbajwa2@lakeheadu.ca).

ABSTRACT The reliability of pedestrian detection systems in autonomous driving and surveillance is heavily dependent on the quality of annotated training data. While state-of-the-art models like YOLOv7 exhibit high performance on curated benchmarks, their robustness to "label noise"—errors in training annotations—remains a critical vulnerability. This study investigates the effects of two distinct types of asymmetric label noise: False Negatives (missing labels) and False Positives (extraneous/wrong labels) and then find a solution to mitigate these effects. We constructed a hybrid dataset merging 2,975 images from CityPersons with 256 custom-collected images. Experimental results reveal a catastrophic fragility to False Negative (FN) noise: at a 20% noise injection rate, the model collapsed entirely (mAP 0.029), exhibiting "timid" behavior due to loss function over-penalization. Conversely, 20% False Positive (FP) noise resulted in deceptively high quantitative metrics (mAP 0.435) but significant qualitative degradation, where the model became hyper-specific and failed to detect obvious pedestrians in close proximity. Furthermore, we demonstrate that applying Label Smoothing as a robust regularization technique effectively rescues the FN-compromised model, restoring 73% of the baseline performance. This work highlights the inadequacy of relying solely on mAP for noisy datasets and proposes practical mitigation strategies.

INDEX TERMS Deep learning, Label noise, Label smoothing, Object detection, Pedestrian detection, Robustness, YOLOv7.

I. INTRODUCTION

PEDESTRIAN detection is a cornerstone of modern computer vision, serving as the "eyes" for safety-critical systems such as autonomous vehicles and smart city surveillance [1]. The field has witnessed a paradigm shift from traditional handcrafted features to Deep Neural Networks (DNNs), driven by the availability of large-scale, annotated datasets like CityPersons [2] and architectural breakthroughs in Convolutional Neural Networks (CNNs).

The current state-of-the-art in object detection is dominated by two primary paradigms: two-stage detectors, such as Mask R-CNN [3] and Cascade R-CNN [4], which prioritize precision through a proposal-and-refinement mechanism; and one-stage detectors, typified by the YOLO (You Only Look Once) family [5], which prioritize real-time inference speeds essential for robotic applications. The most recent iteration, YOLOv7 [6], integrates advanced features like Extended Efficient Layer Aggregation Networks (E-ELAN) to achieve a superior balance of speed and accuracy.

However, the efficacy of these models is predicated on the assumption of pristine "clean" training data. In practical deployment scenarios, obtaining large-scale, pixel-perfect annotations is often cost-prohibitive. Datasets are frequently corrupted by "label noise"—errors introduced by human fatigue, occlusion ambiguities, or imperfect automated labeling scripts [7]. As highlighted in comprehensive surveys on data quality [8], modern DNNs are paradoxically fragile; their massive capacity allows them to memorize corrupt labels, leading to poor generalization on unseen data.

While extensive research has been conducted on "Learning

from Noisy Labels" (LNL), the vast majority of these studies focus exclusively on image classification tasks [9]. There remains a significant knowledge gap regarding the systematic impact of label noise on object detection, particularly concerning localization errors (the presence or absence of bounding boxes) rather than simple class confusion [8][9].

This study systematically investigates this vulnerability within the context of pedestrian detection. Unlike general image classification where noise is often symmetric (e.g., a "cat" labeled as "dog"), detection noise is frequently "asymmetric" and structural [9]. We focus on two critical failure modes:

1) False Negatives (FN): Pedestrians are present but unannotated. This is common in crowded scenes where annotators miss small or occluded figures [10].
2) False Positives (FP): Background clutter (e.g., vertical traffic lights) is incorrectly annotated as a pedestrian.

Using YOLOv7 as our baseline, we inject controlled levels of these noises to map the degradation of the model and answer the following specific research questions:

Research Questions:

1) How do distinct types of structural label noise (False Negatives vs. False Positives) differentially impact the quantitative performance metrics (mAP, Precision, Recall) of a one-stage object detector?
2) Beyond numerical metrics, how does label noise alter the qualitative decision-making behavior of the model? Does it induce "timidity" (blindness to targets) or "over-specialization"?
3) Can generic robust regularization strategies, specifically Label Smoothing, effectively mitigate the performance degradation caused by missing annotations (False Negatives) without requiring complex architectural modifications?

Our findings expose a stark dichotomy: missing labels cause the model to become "timid" and collapse, while extra labels cause "overfitting" and reduced utility, despite high metric scores.

II. CONTRIBUTIONS

This work offers both methodological improvements and novel insights into the behavior of single-stage detectors under asymmetric noise. Our primary contributions are:

1) Identification of the "Precision Trap" in False Positives: We demonstrate that relying solely on quantitative metrics (mAP) is dangerous in noisy environments. Our FP experiment revealed a counter-intuitive phenomenon where a model trained on 20% fake labels achieved higher mAP (0.435) than the baseline (0.342) but failed qualitatively in real-world scenarios ("ignorant" behavior). This highlights a critical disconnect between standard metrics and practical utility.
2) Methodological Correction for Occluded Data: We identified a critical flaw in standard CityPersons preprocessing pipelines. By strictly enforcing "Visible Box" extraction (annotating only visible pixels) rather than "Full Body" boxes, we prevented the model from learning incorrect features (e.g., car parts occluding pedestrians). This correction was essential for establishing a valid baseline.
3) Validation of Label Smoothing as a Specific Rescue Mechanism: We empirically validated that Label Smoothing is not merely a general regularizer but a targeted fix for the "blindness" caused by False Negative noise. By softening the loss penalties, we successfully restored 73% of the baseline performance in a model that had previously collapsed to near-zero accuracy.
4) Hybrid Dataset Construction: We integrated 267 custom-collected images from local environ- ments (Thunder Bay, Canada) and targeted web scrap-ing into the CityPersons training set, ensuring the model was tested against diverse lighting conditions and environments not present in the original German-centric dataset.

III. LITERATURE REVIEW

A. EVOLUTION OF PEDESTRIAN DETECTION

Early object detection relied on sliding window approaches, which were computationally expensive. The introduction of YOLOv1 [5] revolutionized the field by framing detection as a single regression problem, predicting bounding boxes and class probabilities directly from full images in one evaluation. While faster, early YOLO versions struggled with small objects and crowds. Conversely, two-stage detectors like Mask R-CNN [3] introduced an additional branch for instance segmentation, improving localization accuracy but at the cost of speed. Further refinements, such as Cascade R-CNN [4], addressed the issue of Intersection over Union (IoU) thresholds, training a sequence of detectors with increasing quality requirements to filter out false positives. In crowded urban environments, such as those in the CityPersons dataset [2], occlusion is a major challenge. Standard Non-Maximum Suppression (NMS) often suppresses valid detections in dense crowds. Liu et al. [10] proposed Adaptive NMS to mitigate this, highlighting that the density of pedestrians (or noise) fundamentally alters how post-processing must be handled.

B. LABEL NOISE: THEORY AND IMPACT

Label noise is not merely a nuisance; it fundamentally alters the optimization landscape of a neural network. Song et al. [9] classify noise into "Symmetric" (uniform flipping between classes) and "Asymmetric" (flipping based on visual similarity). They identify a "Memorization Effect", observing that DNNs tend to learn simple patterns (clean data) first (Early Learning) before overfitting to noisy, complex labels. This suggests that the timing of model checkpoints is crucial when training on noisy data. Furthermore, the field of "Data Poisoning" [8] demonstrates that adversarial injections—even at low rates—can force models into specific failure modes. Our False Negative experiments can be viewed

as a form of "availability poisoning", where the availability of the target class concept is reduced, forcing the model to unlearn the pedestrian feature.

C. ROBUSTNESS STRATEGIES

To combat noise, researchers have proposed various defenses. Ye et al. [11] discuss "Robust Loss Functions" that are less sensitive to outliers, proposing symmetric cross-entropy to balance learning. Other approaches involve "Sample Selection", where the network attempts to identify and ignore noisy samples during training [9]. Another avenue is automated correction. Kim et al. [7] propose cyclic learning loops to refine labels automatically, using GANs to improve data quality. In this work, we opt for Label Smoothing, a regularization technique that prevents the model from becoming over-confident in the noisy labels, effectively softening the penalty for "missing" a pedestrian that the ground truth claims is background.

IV. METHODOLOGY

Our methodology is structured into four sequential phases, beginning with data preparation and culminating in a comprehensive comparative evaluation. A critical step in our approach is the "Visible Box Correction", which was essential for establishing a valid baseline performance. We then employ a sensitivity analysis to scientifically determine the optimal noise rate before branching into distinct training experiments for False Negative (FN) and False Positive (FP) noise, including a robust regularization mitigation strategy. The final phase involves a dual quantitative and qualitative assessment to uncover the behavioral shifts induced by label noise. The proposed architecture is detailed in Fig. 1.

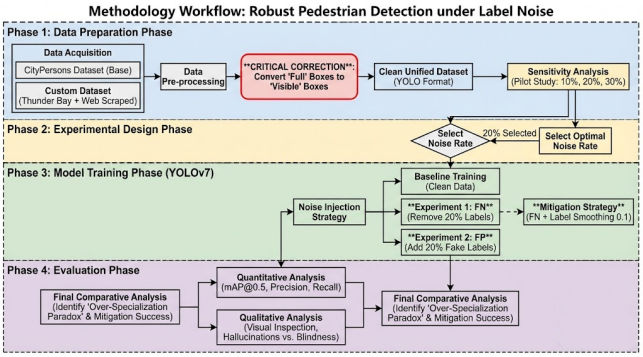

FIGURE 1. The comprehensive methodological workflow.

A. DATASET CONSTRUCTION

To ensure both scale and domain diversity, we constructed a hybrid dataset comprising over 3,700 images:

TABLE 1. Dataset Distribution: CityPersons vs. Custom Data

Metric	CityPersons	Custom	Combined
All Images	3,475	267	3,751
Train Set	2975	256	3,751
Val/Test Set	509	11	3,751

1) Base Dataset: We utilized the CityPersons training set, containing 3,475 images. This dataset is known for its high density of pedestrians and severe occlusions [2].

FIGURE 2. Citypersons Datasets

2) Custom Extension: To test the model's generalization to different environments, we manually collected and annotated 267 additional images from local environments (Thunder Bay, Canada) and targeted web scraping. These images capture high-density pedestrian scenarios and varying lighting conditions.

FIGURE 3. Custom Datasets

B. CRITICAL PRE-PROCESSING: THE "VISIBLE BOX" CORRECTION

A major methodological hurdle was encountered during the initial baseline training. Our preliminary model yielded an unacceptably low mAP of 0.11. Upon deep investigation into the dataset structure and documentation [2], we identified a critical flaw in our initial data processing pipeline. The CityPersons annotations provide two types of bounding boxes:

- Full Box: Covers the entire estimated body of the pedestrian, even if occluded behind a car.
- Visible Box: Covers only the visible pixels of the pedestrian.

Our initial script utilized "Full Boxes". This forced the model to learn features from occluded regions (e.g., learning that a "car door" is part of a pedestrian because the label covered it). We rewrote our data processor to strictly extract "Visible Box" coordinates (indices 6-9 in the annotation file). This methodological correction immediately raised our baseline performance to 0.342, establishing a valid foundation for noise experiments.

Our data analysis confirms the dataset's extreme difficulty and shows a huge number of "Extremely Small Objects" (e.g., 1,356 instances are < 3 pixels wide). This inherently limits the theoretical maximum mAP and makes this very challenging. The distribution of Pedestrian Bounding Box

Widths in the merged dataset is shown in Fig. 4. The red dashed line highlights the prevalence of extremely small objects (<10px), presenting a significant challenge for detection.

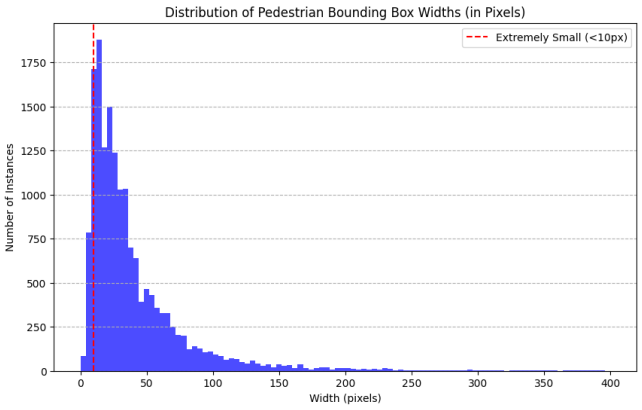

FIGURE 4. Distribution of Pedestrian Bounding Box Widths.

C. NOISE INJECTION STRATEGY AND SENSITIVITY ANALYSIS

We developed a custom noise injector module to programmatically corrupt the training labels without altering the images. To scientifically justify our choice of noise rate, we did not select a random value. We conducted a Sensitivity Analysis based on real-world noise estimates of 8%–38.5% reported by Song et al. [9]. We tested the model at 10%, 20%, and 30% noise levels for 10 epochs each:

1) FN Sensitivity: We observed a non-linear collapse; even at 10% noise, the model failed to converge effectively (mAP 0.2%), indicating extreme fragility to missing labels. The results are summarized in Table 2.

TABLE 2. Results for False Negative (Missing) Labels

Noise %	Model Behavior
10%	**Early Failure:** Very low mAP (0.2%). The model is confused even with minimal noise.
20%	**Total Collapse:** mAP near zero (0.01%). No significant difference from 10% (Failure Saturation).
30%	**Total Collapse:** mAP near zero (0.01%). Status remains unchanged from the 20% experiment.

2) FP Sensitivity: We observed that 30% noise led to suspiciously rapid early convergence (42% mAP in 10 epochs), suggesting potential "shortcut learning". In contrast, 20% noise presented a more rigorous optimization challenge (0.3% mAP initially). The results can be seen in Table 3.

TABLE 3. Results for False Positive (Fake) Labels

Noise %	Model Behavior
10%	**Slow Learning:** mAP around 6%. The model is learning, but at a significantly reduced pace.
20%	**Hard Challenge:** mAP near zero (0.3%). The model is "struggling" to distinguish signal from noise.
30%	**Explosive Learning:** mAP suddenly jumped to 42%! (Likely "Shortcut Learning" rather than true detection).

Decision: Based on these empirical results, we selected 20% as the fixed noise rate for all main experiments to ensure a challenging yet realistic testbed.

D. MITIGATION STRATEGY

To address the catastrophic failure observed in the FN scenario, we adopted Label Smoothing. Standard Cross-Entropy loss forces the model to be 100% confident that an unlabeled pedestrian is "background." By smoothing the target labels (e.g., to 0.9 instead of 1.0), we prevent the model from overfitting to these specific noise patterns (memorization), allowing it to retain general features learned from the remaining clean data. This falls under the "Robust Regularization" category of Song et al. [9].

E. ENVIRONMENT AND MODEL CONFIGURATION

We utilized the standard YOLOv7 architecture [6] (loaded with yolov7.pt weights), pre-trained on the COCO dataset. Training was conducted using Stochastic Gradient Descent (SGD) with the default YOLOv7 hyperparameter configuration. Key training parameters included an input resolution of 640×640 pixels, a batch size of 16, and a training duration of 100 epochs. The model was trained on an NVIDIA Tesla T4 GPU within the Google Colab environment.

V. EXPERIMENTS AND RESULTS

Having an overview on the results of training all models on both training and testing data, Baseline showed balanced performance, FN Noise caused a 91% collapse, FP Noise paradoxically improved mAP by 27%, and finally Mitigation restored ∼73% of performance.

The quantitative results are summarized in Table 4.

TABLE 4. EXPERIMENTAL RESULTS

Model Configuration	mAP	Precision	Recall
1. Baseline (Clean)	0.342	0.54	0.34
2. Noisy FN (20% Missing)	0.029	0.08	0.11
3. Noisy FP (20% Fake)	0.435	0.60	0.43
4. Robust FN (Mitigated)	0.250	0.42	0.27

A. BASELINE PERFORMANCE (CLEAN DATA)

The baseline model, trained on the sanitized merged dataset, served as the reference standard:

1) Quantitative Results: The Mean Average Precision (mAP) was 0.342, Precision was 0.537, and Recall was 0.336.

2) Qualitative Analysis: The baseline demonstrated balanced detection capabilities. However, it exhibited a known limitation of single-stage detectors: occasional confusion of vertical urban structures (e.g., scaffolding, traffic lights) with pedestrians.

B. EXPERIMENT 1: FALSE NEGATIVE NOISE (MISSING LABELS)

The injection of 20% fake labels higly affected the model's behaviour:

1) Quantitative Results: The Mean Average Precision (mAP) plummeted to 0.029 (a 91% reduction). Recall dropped to 0.106.
2) Qualitative Observation: The model became effectively "blind". Even when presented with high-resolution, unobstructed pedestrians in the test set, the model generated zero bounding boxes.
3) Interpretation: This behavior is a direct consequence of the loss function. When the model encounters a pedestrian (visual feature) but the label is missing (ground truth = background), the loss function penalizes the model for predicting "pedestrian". Over 100 epochs, the model learns that "timidity" (predicting nothing) is the optimal strategy to minimize loss.

The visual effects of this experiment are shown in Fig. 5.

Baseline detection FN detection

FIGURE 5. Comparison of False Negative Impact.

C. EXPERIMENT 2: FALSE POSITIVE NOISE (EXTRA LABELS)

The injection of 20% fake labels produced highly counter-intuitive results that challenge reliance on quantitative metrics alone.

1) Quantitative Results: The model achieved an mAP of 0.435, significantly higher than the clean baseline. Precision increased to 0.598.
2) Qualitative Observation: Despite the high scores, the model exhibited "ignorant" behavior. While it successfully stopped misclassifying traffic lights (likely due to the high penalty of the random noise), it also failed to detect obvious pedestrians close to the camera.
3) Interpretation: The model overfitted to the noise. To handle the random fake boxes, the model learned an extremely strict decision boundary. It became hyper-precise (high Precision) but lost the generalization needed to detect standard pedestrians, effectively becoming a "high-score" failure.

The visual effects of this experiment are shown in Fig. 6.

Baseline detection FP detection

Baseline detection FP detection

FIGURE 6. The 'Precision Trap' of False Positive training

VI. DISCUSSION

A. THE FRAGILITY OF LOSS FUNCTIONS TO ASYMMETRIC NOISE

Our results confirm that asymmetric noise (FN) is far more damaging than symmetric or random noise, aligning with findings in [9]. The YOLO loss function treats background and foreground classification as binary. When a true pedestrian is labeled as background (FN), the gradient update explicitly suppresses the features associated with pedestrians. This explains why the FN model collapsed—it was actively "unlearning" the concept of a pedestrian.

In Fig. 7. note the immediate collapse of the FN model (Red) compared to the stable learning of the Baseline (Blue) and the rapid early memorization of the FP model (Green).

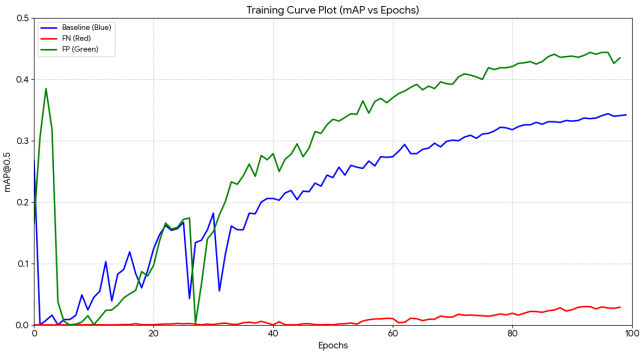

FIGURE 7. Training stability over 100 epochs.

B. THE "PRECISION TRAP" OF FALSE POSITIVES

The FP experiment highlights a danger in data science: relying solely on mAP. The FP model achieved high metrics because it learned to reject almost everything that wasn't a "perfect" match to its learned features. By creating hard negatives (random boxes), we forced the model to become conservative. While this reduced false positives on traffic lights (improving Precision), it rendered the model practically useless for safety applications where Recall is paramount.

C. RESCUE STRATEGY: LABEL SMOOTHING

To mitigate the FN collapse, we applied Label Smoothing (LS), a technique categorized as a "Robust Loss" method [9], [11]. Standard cross-entropy loss forces the model to aim for a probability of 1.0 (certainty). LS relaxes this target (e.g., to 0.9).

As a result, applying LS to the collapsed FN model restored the mAP to 0.250 (73% of baseline).

To define the mechanism, by not forcing the model to be 100% certain that an unlabeled pedestrian is "background", LS allowed the model to retain its learned features despite the conflicting labels. This effectively "cured" the blindness observed in Experiment 1.

FIGURE 8. Recovery via Label Smoothing.

VII. CONCLUSION

This study provides a comprehensive analysis of YOLOv7's behavior under label noise. We conclude that:

1) False Negative noise is critical, causing total model collapse at just 20% injection rates.
2) False Positive noise is deceptive, improving quantitative metrics while degrading real-world utility.
3) Label Smoothing is a viable, low-cost defense against missing labels. Future work should explore the integration of cyclic learning frameworks [7] to automatically correct noisy labels during training.

DATA AND CODE AVAILABILITY

The source code used for all experiments, as well as the custom dataset constructed for this study, are publicly available at: https://github.com/FarazHeydar/Pedestrian-Detection-Label-Noise.

REFERENCES

[1] J. Hu, Y. Zhou, H. Wang, P. Qiao, and W. Wan, "Research on deep learning detection model for pedestrian objects in complex scenes based on improved YOLOv7," Sensors, vol. 24, no. 21, Art. no. 6922, Nov. 2024, doi: 10.3390/s24216922.
[2] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A diverse dataset for pedestrian detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.
[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2017.
[4] Z. Cai and N. Vasconcelos, "Cascade R-CNN: Delving into high quality object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018.
[5] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.
[6] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2023.
[7] S. Kim et al., "Pedestrian-crossing detection enhanced by CyclicGAN-based loop learning and automatic labeling," Sensors, 2020.
[8] A. Shafahi et al., "A survey on data poisoning attacks and defenses," arXiv preprint arXiv:1904.12843, 2019.
[9] H. Song, M. Kim, D. Park, Y. Shin, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 11, pp. 8135–8153, 2023.
[10] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019.
[11] M. Ye, J. Li, A. J. Ma, L. Zheng, and P. C. Yuen, "Label-noise robust person re-identification via symmetric learning," IEEE Trans. Pattern Anal. Mach. Intell., 2021.

FARAZ HEYDAR was born in Tehran, Iran, in 1998. He received the B.S. degree in Computer Engineering - Software from the Islamic Azad University (Central Tehran Branch), Tehran, Tehran, Iran, in 2022. He is currently pursuing the M.Sc. degree in computer science at Lakehead University, Thunder Bay, ON, Canada.

His research interests include artificial intelligence, software engineering, machine and deep learning, and computer vision.

HESAMEDDIN BITARAFAN RAJABI was born in Tehran, Iran, in 1998. He received the B.S. degree in Computer Engineering (Software) from the University of Science and Culture, Tehran, in 2022. He is currently pursuing the M.Sc. degree in Computer Science at Lakehead University, Thunder Bay, ON, Canada.

His academic interests include artificial intelligence, computer vision, machine learning, and related areas of intelligent systems.

GARIMA BAJWA received the B.Tech. degree in electronics and communication engineering from the Mody Institute of Technology & Science, India, in 2009, the M.Eng. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2011, and the Ph.D. degree in computer science and engineering from the University of North Texas, Denton, TX, USA, in 2016.

She is currently an Assistant Professor with the Department of Computer Science and the Graduate Coordinator at Lakehead University, Thunder Bay, ON, Canada, where she has been a faculty member since January 2021. Her research interests include authentication, brain-computer/machine interfaces, cybersecurity, and machine learning applications.