Dipkamal Bhusal

✉ db1702@rit.edu Google Scholar Github Webpage in LinkedIn

Research Overview

I am a Ph.D. candidate at Rochester Institute of Technology, highly interested in building reliable and trustworthy AI systems. My research focuses on understanding deep learning models through explainable and adversarial machine learning. I have worked in computer vision, cybersecurity, large language models (LLMs) and vision language models (VLMs). My goal is to develop principled tools and techniques that enhance the transparency, robustness, and real-world deployment of modern AI systems..

Research Interests

Explainable AI, Adversarial machine learning, Concept-based explanations, Feature attribution methods, Reliable deep learning, Vision and language models

Education

Ph.D. **Rochester Institute of Technology,** Rochester, NY, USA,
2021–Present *Concentration – Computing and Information Science*
Advisor – Dr. Nidhi Rastogi
Thesis – Towards reliable and trustworthy deep learning
Anticipated graduation – April, 2026 .

M.Sc. **Tribhuvan University, Institute of Engineering, Pulchowk Campus** Lalitpur, Nepal,
2019–2021 *Concentration – Information and Communication Engineering*
Advisor – Dr. Sanjeeb Prasad Panday
Thesis – Multi-label classification of thoracic diseases using DenseNet on chest radiographs .

B.E. **Tribhuvan University, Institute of Engineering, Pulchowk Campus** Lalitpur, Nepal,
2012–2016 *Electronics and Communication Engineering*
Advisor – Dr. Nanda Bikram Adhikari
Thesis – Prototyping of a voice command based object recognizing robot using speech and image feature extraction .

Industry Experience

Oct'25– Jan'26 **Applied Scientist Intern**, *Amazon Science*, Tempe, AZ.

1. Designed a multi-stage inference solution to address VLM limitations in document layout intelligence, employing anchor classes and blind self-verification to fix prompt bias and false positives; achieved >90% precision and recall on internal documents.

2. Fine-tuned Qwen3-VL using LoRA to create a unified, single-inference model for layout classification, significantly reducing latency while maintaining >75% precision and recall.

3. Conducted comprehensive failure mode analysis and submitted findings to an internal conference; earned an inclined offer for full-time employment.

July'25– Oct'25 **Research Engineer Intern**, *Athena Software Group*, Rochester, NY.

1. Developed *AthenaBench*, an enhanced dynamic benchmark for evaluating LLM reliability in Cyber Threat Intelligence (CTI), extending prior work on SECURE and CTIBench.

2. Engineered an improved dataset creation pipeline with automated duplicate removal and defined new tasks for risk mitigation strategies.

3. Implemented refined evaluation metrics and contributed to research that was accepted at the ACSAC CTI-Workshop 2025.

Dec'16–Jun'21 **Co-founder/Software Engineer**, *Paaila Technology*, Kathmandu, Nepal.

1. **Dec'16-Nov'17 (ML engineer):** Contributed as a ML engineer in design and development of robotics and AI solutions, primarily worked in face recognition and speech synthesis projects; robots deployed in real-world restaurant and hospital environments.
2. **Dec'17-Nov'18 (Project manager):** As the team grew in size, I took the position of project manager to manage the team, and product development.
3. **Dec'19-Jun'21 (Managing director):** I took the role of director and was involved in planning and managerial activities, and was responsible for the overall growth of the startup. Due to internal conflicts, I resigned in July 2021 to pursue PhD in the USA.

Skills

Languages	Python
Frameworks	PyTorch
Libraries	NumPy, Pandas, Scikit-learn, OpenCV, Matplotlib
Utilities	Jupyter Notebook, Visual Studio, Git, Latex
Electronics	PCB Design, Circuit Simulation, 8 bit microcontroller programming
Industry	Teaching, Project Management, Public Speaking, Business Strategy and Development

Publications

2025

[1] **D. Bhusal**, M. Clifford, S. Rampazzi, N. Rastogi. *FACE: Faithful Automatic Concept Extraction*. NeurIPS 2025.

[2] R. Yang, **D. Bhusal***, N Rastogi, *”Learning to Look: Cognitive Attention Alignment with Vision-Language Models”*. First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models, Neurips 2025. ***Mentorship and Equal contribution.**

[3] S. Suwal, **D. Bhusal***, M. Clifford, N. Rastogi, *”Do Sparse Subnetworks Exhibit Cognitively Aligned Attention? Effects of Pruning on Saliency Map Fidelity, Sparsity, and Concept Coherence”*. First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models, Neurips 2025.***Mentorship and Equal contribution.**

[4] A. Mehrotra, D. Peng, **D. Bhusal***, N. Rastogi, *”Concept-Based Masking: A Patch-Agnostic Defense Against Adversarial Patch Attacks”*.. Reliable ML from Unreliable Data Workshop, NeurIPS 2025. ***Mentorship.**

[5] MT. Alam, **D. Bhusal**, N. Rastogi, *”R+R: Revisiting Static Feature-Based Android Malware Detection using Machine Learning”*. ACSAC 2025

[6] **D. Bhusal**, N. Rastogi, *”Adversarial Patterns: Building Robust Android Malware Classifiers”*. ACM Computing Surveys 2025

2024

[1] A. Mehrotra, **D. Bhusal***, N. Rastogi, *”Hessian Sets: Uncovering Feature Interactions in Image Classification”*. Attributing Model Behavior at Scale (ATTRIB), NeurIPS 2024. ***Mentorship**.

[2] **D. Bhusal***, MT. Alam*, L. Nguyen, A. Mahara, Z. Lightcap, R. Frazier, R. Fieblinger, GL. Torales, N. Rastogi *”SECURE: Benchmarking Generative Large Language Models for Cybersecurity Advisory”* at 40th Annual Computer Security Applications Conference (ACSAC). *Equal Contribution.

[3] MT. Alam*, **D. Bhusal***, L. Nguyen, N. Rastogi, *”CTIBench: A Benchmark for Evaluating LLMs in Cyber Threat Intelligence*”. NeurIPS 2024 (Spotlight paper. *Equal Contribution).

[4] **D. Bhusal**, MT. Alam, MK. Veerabhadran, M. Clifford, S. Rampazzi, N. Rastogi. *"PASA: Attack Agnostic Unsupervised Adversarial Detection using Prediction & Attribution Sensitivity Analysis"*. 9th IEEE European Symposium on Security and Privacy (2024)

[5] S. Kasarapu, **D. Bhusal**, N. Rastogi, SM. Pudukotai Dinakarrao, *"Comprehensive Analysis of Consistency and Robustness of Machine Learning Models in Malware Detection"*. Great Lakes Symposium on VLSI 2024.

2023

[1] **D. Bhusal**, R. Shin, AA. Shewale, MK. Veerabhadran, M. Clifford, S. Rampazzi, N. Rastogi. *"SoK: Modeling Explainability in Security Analytics for Interpretability, Trustworthiness, and Usability"*. 18th International Conference on Availability, Reliability and Security (2023)

[2] MT. Alam, **D. Bhusal**, Y. Park, N. Rastogi, *"Looking Beyond IoCs: Automatically Extracting Attack Patterns from CTI"*. 26th International Symposium on Research in Attacks, Intrusions and Defenses (2023)

arXiv

[1] **Dipkamal Bhusal**, SP Panday, "Multi-Label Classification of Thoracic Diseases using Dense Convolutional Network on Chest Radiographs"

[2] MT Alam, **Dipkamal Bhusal**, Y Park, N Rastogi, "CyNER: A Python Library for Cybersecurity Named Entity Recognition".

Under-review

[1] **D. Bhusal**, MT. Alam, M. Clifford, S. Rampazzi, N. Rastogi, *"Training for Trustworthy Saliency Maps: Adversarial Training Meets Feature-Map Smoothing"*.

Research Projects

- **"Fixing spurious features learned by a model":** I am investigating whether we can utilize post-hoc concept explanations to inspect spurious features learned by image classifiers and guide the model in unlearning them using few-shot examples.

- **"Concept-Driven Adversarial Defense":** I proposed a new defense against adversarial patch attack that combines concept based interpretability to suppress the most influential concept activation vectors during inference. The method successfully neutralizes adversarial patches regardless of size or location, setting a new state-of-the-art on Imagenette. I mentored two students on this project.

- **"Geometry of concepts in LLM:"** I extended my work on faithful concept explanations for image classifiers to LLMs by studying the *geometric relationships* between different concept vectors in order to map the structure of a LLM's "concept space". This can tell what LLMs learn, and either the statistical concept representations align or diverge from human intuition.

- **"Right for Right Reason:"** CNNs frequently "cheat" by exploiting superficial correlations. In this project, I mentored an undergraduate student to develop a scalable framework that leverages vision-language models to automatically generate semantic attention maps using natural language prompts, and train a model for classification and attention alignment to improve generalization, and reduce shortcut reliance.

- **"Faithful concept extraction technique":** Concept explanations provide human comprehensible form of explanations to image classifiers. However, existing methods lack faithfulness to the underlying model. In this work, I proposed a faithfulness aware concept extraction framework.

- **"Interpretability price of pruning":** While pruning creates efficient models, its effect on interpretability is unclear. In this project, I mentored a graduate student at RIT to evaluate interpretability of ResNet-18 ImageNette model across pruning levels. We found that moderate pruning improves saliency map faithfulness and sharpens human-aligned attention, while excessive pruning merges concepts and undermines interpretability.

- **"Feature-map smoothing for saliency map interpretability":** In this work, I investigated the trade-off between stability and sparsity of saliency maps and proposed the use of a smoothing layer during adversarial training to obtain high quality saliency maps.

- **"Uncovering feature interactions in image classification":** In this work-in-progress paper, published at NeurIPS'24 ATTRB Workshop, we developed a technique that leverages the Hessian matrix to detect and attribute pairwise feature interactions in image classifiers. This was my first paper as a mentor. We have extended the work with attribution axioms.
- **"Benchmarking large language models":** We introduced a knowledge-intensive framework called 'SECURE (Security Extraction, Understanding & Reasoning Evaluation)', a benchmark designed to assess LLMs performance in realistic cybersecurity scenarios. We extended the knowledge intensive LLM evaluation framework proposed in SECURE and designed LLM benchmarks for CTI-specific tasks, called CTIBench. Currently used by Google, Cisco, Trend Micro, and other companies.
- **"Unsupervised adversarial detection":** In this work, I developed a practical method for utilizing sensitivity of model prediction and feature attribution to detect adversarial attack on various deep learning models.
- **"Modeling explainability":** In my first deep dive into explainable machine learning, I provided a comprehensive analysis of feature attribution based explanation methods and demonstrated their efficacy in different security applications.

Teaching

Instructor **Explainable AI (DSCI 789) at RIT**, *Fall 2024*, My advisor took a sick leave and I took her classes., Designed & delivered lectures; assignments; capstone mini-projects.

Lecturer **IIMS College (Kathmandu)**, *Sept. 2020 - Aug. 2021*, As a lecturer in the Computer Science department, I taught two BSc. IT undergraduate courses: Introduction to Python and Machine Learning.

Advising & Mentorship

Summer'25 **Ayushi Mehrotra & Derek Peng**, *California Institute of Technology & University of California, Berkeley*, Mentored Ayushi and Derek on building interpretability based defense against adversarial patch attack, Got paper accepted at NeurIPS 2025 Reliable ML from Unreliable Data Workshop.

Summer'25 **Ryan Yang**, *Brown University*, Mentored Ryan who visited RIT as a undergrad researcher on XAI research specifically, designing CNN models which are right for right reasons, Got paper accepted at NeurIPS 2025 CogInterp Workshop.

Fall'24-
Spring'24 **Sayali Rajesh Kale and Achyut Sridhar Kulkarni**, *RIT Masters in Data Science*, Supervised their data science capstone project on concept-based explanations..

Fall, 2024 **Sanish Suwal**, *RIT Masters in Computer Science*, Supervised his final year capstone project on study of post-hoc explanations under neural network pruning, Got paper accepted at NeurIPS 2025 CogInterp Workshop.

Spring'23-
Fall'24 **Ayushi Mehrotra**, *Troy High School*, Research on interactive feature attribution method for image classifiers, Got paper accepted at NeurIPS 2024 Attributing Model Behavior at Scale Workshop.

Service

Program Committee

[1] IEEE Secure and Trustworthy ML (SATML) 2025

Reviewer (Journals)

[1] Journal of Artificial Intelligence Research (2024-2025)

[2] IEEE Transactions on Artificial Intelligence (2025)

[3] Computational linguistics (2025)

Reviewer (Conferences)

[1] IEEE Secure and Trustworthy ML (SATML) 2025

[2] NeurIPS (2025)

[3] CVPR (2025)

Reviewer (Workshops)

[1] First Workshop on CogInterp: Interpreting Cognition in Deep Learning Models, NeurIPS 2025

[2] Attributing Model Behavior at Scale (ATTRIB), NeurIPS 2024

Honors and achievements

Scholarship	Financial Support for Ph.D. in Computer Science at RIT, 2021-Present.
Training	Conducted two-week training in Python and Data Science at IIMS College, Kathmandu, 2021.
Award	National ICT Innovation Award by Ministry of Communication and Information Technology (Nepal Government) for Paaila Technology, 2019 .
Scholarship	Full scholarship for Masters in Information Engineering at Pulchowk Campus, Tribhuvan University, 2019.
Award	Most Creative Business of Nepal by Antarprena. Represented Team Nepal at global finals in Copenhagen, Denmark, 2018.
Award	Best Startup of Nepal by ICT Magazine, 2017.
Contest	Winner of Object Oriented Programming Competition by FlipKarma at Pulchowk Campus, 2014.
Award	Secured third position at national level quiz competition Quizmania broadcasted on national television (2012).
Scholarship	Full scholarship for bachelor in engineering at Pulchowk Campus, Tribhuvan University, 2012.
Scholarship	Full Scholarship at Balkumari College, Chitwan (10+2 college equivalent to junior and senior high school level in US) + College Topper + First Position at final examinations in the whole district.

Selected news media

Republica Network: Using AI for better customer experience.

Digital Trend: On waiter robots of Nepal.

AFP News Agency: Nepal's first robot waiter ready for orders.

Kantipur (Nepali): Paaila Technology.

NDTV: In A First, Nepal's Restaurant Uses Robots As Waiters.

India Today: On waiter robots of Nepal.

South Chine Morning Post: Meet Ginger: Nepal's first robot waiter is ready to take your order .

New Business Age Magazine: Embracing the Age of AI, Robotics and ML in Nepal.

Republica Network: Ginger Robot of Nepal.

Kathmandu Post: Emergency Ventilators.