Dipkamal Bhusal

✉ db1702@rit.edu Google Scholar Github Webpage LinkedIn

Overview

I am dedicated to advancing the field of reliable AI that improves trust and understanding of deep learning models. My research focuses on understanding deep learning models with tools and techniques from explainable machine learning and adversarial machine learning. I have extensively worked in computer vision and cybersecurity models, with recent exploration of reliability of large language models (LLMs) .

Education

Ph.D. 2021–Present **Rochester Institute of Technology,** Rochester, NY, USA,
Concentration – Computing and Information Science
Advisor – Dr. Nidhi Rastogi .

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

B.E. 2012–2016 **Tribhuvan University, Institute of Engineering, Pulchowk Campus** Lalitpur, Nepal,
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 .

Experience

August 2021–Present **Graduate Research Assistant**, *Rochester Institute of Technology*, Ai4SecLab, Primary research lies at the intersection of machine learning and security.

1 **“Towards improving saliency map interpretability using feature map smoothing”**, *Under-review*, In this work, we investigate the trade-off between stability and sparsity of saliency maps in naturally and adversarially trained models and propose the use of a smoothing layer during adversarial training to obtain high quality saliency maps.

2 **“Hessian Sets: Uncovering Feature Interactions in Image Classification”**, *NeurIPS’24 ATTRB Workshop*, 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.

3 **“Ctibench: A benchmark for evaluating llms in cyber threat intelligence”**, *NeurIPS’24*, We extend the knowledge intensive LLM evaluation framework proposed in SECURE and design LLM benchmarks for CTI-specific tasks, Currently used by Google, Cisco, Trend Micro.

4 **“SECURE: Benchmarking Large Language Models for Cybersecurity”**, *ACSAC’24*, We introduce a knowledge-intensive framework called ‘SECURE (Security Extraction, Understanding & Reasoning Evaluation)’, a benchmark designed to assess LLMs performance in realistic cybersecurity scenarios.

5 **“PASA: Attack Agnostic Unsupervised Adversarial Detection using Prediction & Attribution Sensitivity Analysis”**, *EuroS&P’24*, We develop a practical method for utilizing sensitivity of model prediction and feature attribution to detect adversarial attack on deep learning models.

6 **“SoK: Modeling Explainability in Security Analytics for Interpretability, Trustworthiness, and Usability”**, *ARES’23*, My first work in explainable machine learning where we provide a comprehensive analysis of feature attribution based explanation methods and demonstrate their efficacy in different security applications.

Dec’16-Jun’21 **Co-founder/Software Engineer**, *Paaila Technology*, Co-founder of an AI startup in Kathmandu.

1 **Dec’16-Nov’17**, *ML engineer*, Contributed as a ML engineer in design and development of robotic and AI solutions, Worked in face recognition and speech synthesis projects.

2 **Dec’17-Nov’18**, *Project manager*, As the team grew in size, I took the position of project manager to manage the project on our robotics projects..

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..

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

Publications

[1] A Mehrotra, **Dipkamal Bhusal***, N Rastogi, *”Hessian Sets: Uncovering Feature Interactions in Image Classification”* at Attributing Model Behavior at Scale (ATTRIB), NeurIPS 2024. ***Mentorship**.
[2] MT Alam*, **Dipkamal Bhusal***, L Nguyen, N Rastogi, *”CTIBench: A Benchmark for Evaluating LLMs in Cyber Threat Intelligence*" at NeurIPS 2024 (Spotlight paper. Top 2-3% of accepted papers. *Equal Contribution).
[3] **Dipkamal 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.
[4] **Dipkamal Bhusal**, MT Alam, MK Veerabhadran, M Clifford, S Rampazzi, N Rastogi. *”PASA: Attack Agnostic Unsupervised Adversarial Detection using Prediction & Attribution Sensitivity Analysis”* at 9th IEEE European Symposium on Security and Privacy (2024)
[5] **Dipkamal Bhusal**, R Shin, AA Shewale, MK Veerabhadran, M Clifford, S Rampazzi, N Rastogi. *”SoK: Modeling Explainability in Security Analytics for Interpretability, Trustworthiness, and Usability.”* at 18th International Conference on Availability, Reliability and Security (2023)
[6] MT Alam, **Dipkamal 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)
[7] S Kasarapu, **Dipkamal 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.
[8] **Dipkamal Bhusal**, N Rastogi, *”Adversarial Patterns: Building Robust Android Malware Classifiers”*, ACM Computing Surveys 2025

Under Review [1] **Dipkamal Bhusal**, MK Veerabhadran, M Clifford, S Rampazzi and N Rastogi, ”On the connection between model robustness and saliency map interpretability”.
[2] MT Alam, **Dipkamal Bhusal**, N Rastogi, ”Revisiting Static Feature-Based Android Malware Detection”.

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”. on arXiv

Peer-Reviewer

2024 2nd Workshop on Attributing Model Behavior at Scale, NeurIPS 2024.

2024 IEEE Transactions on Artificial Intelligence.

2024 Journal of Artificial Intelligence Research.

2025 Computational linguistics.

2025 Journal of Artificial Intelligence Research.

2025 Pattern Recognition.

Mentorship and Supervision

2023-2025 *Ayushi Mehrotra, Troy High School.* Research on interactive feature attribution method for image classifiers..

2024-2025	*Sanish Suwal, RIT Masters in Computer Science.* Supervised his final year capstone project on study of explanation methods on different model training strategies..
2024-	*Sayali Rajesh Kale and Achyut Sridhar Kulkarni, RIT Masters in Data Science.* Supervising their data science capstone project on concept-based explanations..

Skills

Languages	Python
Frameworks	PyTorch, Keras
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

Teaching

Oct 2024- Dec 2024	Lecturer of Explainable Artificial Intelligence (DSCI 789) at RIT.

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.
Volunteer	Worked with the youth team of Needleweave society to install prefabricated houses for victims of Gorkha earthquake.
Arts	Established a music group Silver Strings for writing and publishing songs on different themes.
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.
Award	Secured second position at inter-college oratory competition organized at the 25th anniversary of Balkumari College, 2011.
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.

Graduate Courses

Quantitative Foundations, Deep Learning, Statistical Machine Learning, Foundation of Algorithms, Software Engineering, Neural Network, Image Processing, Big Data.

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.

Selected Certifications

1. AI for Medical Treatment, Medical Diagnosis, and Medical Prognosis by deeplearning.ai on Coursera. Certificate earned at June, 2020.
2. Deep Learning Specialization by deeplearning.ai on Coursera. Certificate earned in June 2020. Courses include Neural Networks and Deep Learning, Improving Deep Neural Networks: Hyperparameter tuning, Regularization and Optimization, Structuring Machine Learning Projects, Convolutional Neural Networks and Sequence Models..
3. Project Management Principles and Practices Specialization by University of California, Irvine-The Paul Merage School of Business on Coursera. Certificate earned at May 2020.
4. Mathematics for Machine Learning: Linear Algebra and Multivariate Calculus by Imperial College London on Coursera. Certificate earned at April 2020..