KAVITA KUMARI

kavitak128@gmail.com +49-15156311398 Darmstadt, Germany
in Kavita Kumari Last update: 01/06/2025

EXPERIENCE

Postdoctoral Researcher - AI & Security

Technical University of Darmstadt - System Security Lab

December 2022 – Now Darmstadt, Germany

- Main research on the integration of physics concepts into the modeling of AI/ML to solve complex problems.
- Concrete experience with LLMs and NLP, Federated learning, Explainable AI (XAI), Bayesian statistics, Generative AI, Diffusion models, Probability distributions, Gaussian mixture modeling, Reinforcement learning, Speech processing, Energy-based modeling, Physics-augmented ML, GANs.
- **A PC member in NeurIPS & AISTATS & also served in the committee of IEEE S&P and USENIX.**

Keywords: Physics, Mathematics, Frequency, AI, Federated Learning, XAI, Non-parametric Bayesian modeling, Split Learning, Watermarking, Distributed Computing, Privacy-Preserving schemes.

Teaching/Research Assistant - Computer Science

The University of Texas at San Antonio - SPriTELab

August 2016 - August 2022 San Antonio, Texas, US

- Main research on applications and theoretical modeling of complex problems in AI/ML.
- Game theoretic analysis to analyze the privacy attacks in mobile applications and Explainable AI (XAI) based Membership Inference Attacks (MIA) to infer input membership in AI/ML models by forming an analogy to economic models.
- Used probabilistic Machine learning to design a general backdoor defense in Federated Learning. Also, thesis in uncertainty modeling in repeated interaction scenarios.
- Teaching assistant experience.

Keywords: Research, Machine Learning, Pytorch, Game Theory, Stochastic Control Theory, Dynamical Systems, ML/AI Security, Market volatility, Bayesian modeling, Federated Learning.

Exchange student

Technische Universität Darmstadt

June 2021 -July 2022 (7m) Darmstadt (Remote), Germany

- Completed a project in which I designed a backdoor defense in Federated Learning. I used concepts from non-parametric Bayesian Modeling. The paper was accepted at IEEE S&P.

Keywords: Mathematics, Bayesian Modeling, Non-parametric Bayesian modeling, Federated Learning, Machine Learning.

EDUCATION

PhD Computer Science

The University of Texas At San Antonio, GPA:3.8/4.0

August 2016 – August 2022 San Antonio, Texas, US

BTech Computer Science

Aligarh Muslim University, GPA:8.9/10.0

August 2011 – May 2015 Aligarh, India

PUBLICATIONS

- Kavita *et al.*, *"VoiceRadar: Voice Deepfake Detection using Micro-Frequency and Compositional Analysis"*, NDSS 2025.
  Contribution: Designed a novel audio/speech deepfake detector named VoiceRadar to identify fake audio/speech content accurately. It draws inspiration from wave formulation using the Micro-doppler effect. I wrote the paper and conducted the experiments to demonstrate the efficacy of VoiceRadar. Got **Distinguished Paper Award**.
- Rieger *et al.*, *"SafeSplit: A Novel Defense Against Client-Side Backdoor Attacks in Split Learning"*, NDSS 2025.
  Contribution: Designed a novel Rotational distance metric that assesses the orientation shifts of the server's layer parameters during training and contributed to the writing of the paper. Got **Distinguished Paper Award**.
- Kavita *et al.*, *"DEMASQ: Unmasking the ChatGPT Wordsmith"*, NDSS 2024.
  Contribution: Designed a novel ChatGPT detector named DEMASQ, which accurately identifies ChatGPT-generated content that draws inspiration from the multifaceted nature of human communication. I wrote the paper and conducted the experiments to demonstrate that DEMASQ obtained an accuracy of 96.5% compared to the maximum of 47% of existing detectors.
- Kavita *et al.*, *"Xplain: Analyzing Invisible Correlations in Model Explanation"*, Usenix security symposium, 2024.
  Contribution: Designed a new explanation technique that improves existing path techniques to uncover the hidden relationship between different features of the input sample. I wrote the paper and also conducted the experiments.
- Pegoraro *et al.*, *"DeepEclipse: How to Break White-Box DNN-Watermarking Schemes"*, Usenix security symposium, 2024.
  Contribution: Supported in the design of DeepEclipse, a novel unified framework designed to remove white-box watermarks, and helped write the paper.
- Kavita *et al.*, *"Towards a Game-theoretic Understanding of Explanation-based Membership Inference Attacks"*, GameSec 2024.
  Contribution: Modeled and analyzed the interactions between a system comprising of target ML model and its corresponding XAI method using continuous-time stochastic signaling game framework to study explanation-based Membership Inference attacks.
- Pegoraro *et al. "To ChatGPT, or not to ChatGPT: That is the question!"*, arXiv 2023.
  Contribution: Supported in the comprehensive and contemporary assessment of the most recent tools in ChatGPT detection and helped to write the paper.
- Kavita *et al.*, *"BayBFed: Bayesian Backdoor Defense for Federated Learning"*, IEEE Symposium on Security and Privacy, 2023.
  Contribution: Designed and implemented a generic backdoor defense framework called BayBFed, which utilized probability distributions over client updates to detect malicious updates in Federated Learning. I wrote the paper and also conducted the experiments.
- Kavita *et al.*, *"Analyzing Defense Strategies Against Mobile Information Leakages: A Game-Theoretic Approach"*, GameSec 2019.
  Contribution: Modeled and analyzed the interactions between the defense mechanism and mobile applications using Game Theory to prevent malicious applications from stealing private user data embedded in zero-permission sensors. I wrote the paper and also conducted the experiments.

TECHNICAL SKILLS

- **Programming Languages**: Python, C, Java
- **Libraries/Frameworks**: Pytorch, Tensorflow, pandas, numpy