

SUDHARSHAN BALAJI

<https://sudharshanbalaji.com>

GitHub: <https://github.com/bsudharshan2001>

Email : sudharshanbalaji@usf.edu

Mobile : +1 (656) 216-5583

LinkedIn: <https://www.linkedin.com/in/sudharshanbalaji>

EDUCATION

Doctor of Philosophy, Computer Science Present

University of South Florida, Tampa, FL

Master of Science (Thesis Track), Computer Science May 2025

University of South Florida, Tampa, FL

Bachelor of Engineering, Computer Science and Engineering June 2023

Amrita University, Coimbatore, India

Relevant Courses: Machine Learning, Neural Networks and Deep Learning, Security & Privacy in ML, Software Engineering, Mining of Massive Datasets, Intro to LLMs, Topics in NLP, Introduction to AI, Privacy Preserving Trustworthy-Cyberinfrastructures

SKILLS

Programming Languages: Python, C, C++, Java, Dart

Database Management: MySQL, MongoDB

Productivity Tools: Microsoft Office 365, Jira, Git

AI & ML Domains: AI Safety & Alignment, Machine Unlearning, LLMs/VLMs

Multi-Agent Systems, Trustworthy Federated Learning

Frameworks & Libraries: PyTorch, TensorFlow, Keras, OpenCV, HuggingFace Transformers, Flask

Web Development: HTML, CSS, JavaScript, NodeJS

EXPERIENCE

Graduate Research Assistant - SPRAI Lab, USF, Tampa, FL, USA Aug 2024 - Present

- Continuing doctoral research under Prof. Ning Wang focused on safety and alignment problems in AI and Byzantine-resilient Federated Learning.
- Conducting research on AI safety and alignment, machine unlearning, multimodal unlearning in vision-language models, multi-agent systems and trustworthy federated learning.

Graduate Instructional Assistant - SPRAI Lab, USF, Tampa, FL, USA Sep 2024 - May 2025

- Conducted thesis research on security applications of large language models titled "LLMs for Network Intrusion Detection: A Comprehensive Analysis" working with Graduate Advisor Prof. Ning Wang to comprehensively analyze methodologies and exploring the use of LLMs, MLMs, and MLM embeddings for network intrusion detection across scales (Successfully defended the thesis - [Link to thesis](#)).
- Built coursework, supporting instruction and grading for two courses, "Privacy Preserving Trustworthy Cyber-Infrastructures" under Prof. Attila Yavuz, and "Human Aspects of Cybersecurity" under Graduate Advisor Prof. Ning Wang, during Spring 2025.
- Supported instruction and grading for "Security and Privacy in ML" under Prof. Guangjing Wang, assisting students with questions during office hours and helping ensure clarity in assessments for a class of ~ 100 in Fall 2024.

Machine Learning Engineering Intern - dotkonnekt Innovation Labs, Bangalore March 2023 - July 2023

- Contributed to a cross-functional team of engineers, scientists, and product managers to successfully develop and deploy machine learning models and algorithms for the company's D2C Experiential Platform, enabling personalized shopping experiences
- Implemented a cutting-edge recommender engine, encompassing content-based filtering systems, collaborative filtering systems, and hybrid recommenders, for the company's D2C platform
- Integrated Huggingface models with LangChain, including open-source language models such as BERT and Falcon-40B, into the company's platform, contributing significantly to the platform's language processing capabilities
- Leveraged AWS SageMaker for model training, proficiently experimenting with various instance types to optimize model performance
- Deployed models using Node.js and Flask, ensuring seamless integration with existing infrastructure and an improved user experience.

PROJECTS/RESEARCH

Cross-Modal Unlearning in Vision-Language Models (CRUX)

Jan 2026 - Present

- Investigated bidirectional safety-unlearning transfer across LLaVA, InstructBLIP, and IDEFICS, identifying asymmetric and architecture-dependent gaps between text and visual modalities.
- Developed CrossInf, an influence-guided method that identifies and updates LoRA blocks responsible for cross-modal behavior using only single-modality forget supervision and a small cross-modal probe.
- Matched the cross-modal safety performance of multimodal unlearning on LLaVA and IDEFICS while updating only a few selected blocks, without directly training on second-modality forget data.
- Demonstrated that typographic attacks can recover harmful behavior after text-side unlearning, revealing shallow modality-specific forgetting in VLMs.
- Evaluated CrossInf against random-block, same-modality influence, and multimodal-unlearning controls using PKU-SafeRLHF, JailbreakV, FigStep, VQA-v2, and TruthfulQA.
- *Currently under review at AAAI 2027 under the title: “One Modality to Forget Them All: Enhancing Cross-Modal Unlearning in Vision-Language Models.”*

Quantization-Robust LLM Unlearning (QuantUnlearn): [Report] [Code]

Aug 2025 - Jan 2026

- Investigated how deployment-time model compression can unintentionally restore knowledge removed through machine unlearning.
- Reproduced an ICLR study across multiple unlearning strategies using a 7B language model and the MUSE benchmark.
- Developed influence-guided layer selection and quantization-aware training defenses that matched the original method’s robust forgetting performance and improved it in selected settings.
- Evaluated knowledge removal, privacy leakage, and model utility, identifying practical trade-offs between reliable forgetting and performance preservation.

Development of CADRADs Prediction System:

Jun 2024 - Aug 2025

- Spearheaded the development of CCTA-RADS, the largest publicly available dataset of 940 real-world Coronary CT Angiography (CCTA) reports from Tampa General Hospital, annotated with expert-verified CAD-RADS scores under IRB approval.
- Lead a multi-disciplinary collaboration with Prof. Ning Wang (USF SPRAI Lab), Dr. Zhiyu Liu, Dr. Shone Almeida, and Dr. Mathew Karvelil (USF Morsani College of Medicine) to address linguistic and structural heterogeneity in clinical text
- Designed an end-to-end two-stage Clinical NLP pipeline for automated CAD-RADS scoring, integrating an LLM-based report parser (ChatGPT-4o-mini, Mistral DocAI, GPT-4.1) that normalizes free-text reports with fine-tuned BERT-family classifiers (BioBERT, ClinicalBERT) trained on structured JSON data
- Achieved a 90.12% F1-score with BioBERT, surpassing direct LLM prompting and raw-text fine-tuning methods, while improving classification of clinically critical CAD-RADS classes (3-4) and non-diagnostic cases (N).
- Developed a HIPAA-compliant text-processing and data sanitization pipeline to ensure secure report handling and consistent terminology standardization across all training stages.
- Integrated explainable AI (SHAP, LIME) for model interpretability, enabling visualization of key sections influencing each prediction to enhance clinical transparency and trust.
- Deployed the trained model as a web-based clinical decision-support tool, allowing radiologists to upload de-identified reports for automated scoring with interpretability visualizations
- Work supported by Office of Naval Research under Grant N00014-24-1-2730
- *First-author paper published in the ACL 2026 Industry Track and presented as a poster: “Learning from Textual Radiology Reports: A Benchmark Dataset for Coronary CT Angiography.”* [Paper] [Poster] [Presentation] [Code]

Additional Projects: Varuna — Cyclone Intensity Estimation Dashboard (Smart India Hackathon 2022); Civitas — Community Management App (Honeywell Smart City Hackathon 2021).

ACHIEVEMENTS & SCHOLARSHIPS

- Lead a team to Design and Develop a Community Management App using Flutter for HoneyWell Smart City Hackathon 2021. Placed among Top 25 Teams all over India.
- Amrita Vidhyanidhi Scholarship: Awarded 90% of the tuition fee waiver for the entirety of bachelor’s degree