Avinash Amballa

Webpage: https://amballaavinash.github.io **Mobile:** 6095054919
Linkedin: https://www.linkedin.com/in/avinashamballa **Mail:** amballaavinash@gmail.com

Education

University of Massachusetts Amherst *USA*

MS Computer Science *Aug 2023 - July 2025*

- Courses: Reinforcement Learning, Responsible AI, Methods of applied statistics

Indian Institute of Technology Hyderabad (IIT-H) *India*

Bachelor of Technology in Electrical Engineering with Minor in Computer Science and Engineering *Jul 2017 - June 2021*

- CGPA: **8.8/10.0**
- Courses: Data Structures, Algorithms, DBMS, Computer Architecture, Reinforcement Learning, Pattern Recognition and Machine learning, Image processing, Representation Learning, Calculus, Regression Analysis, Combinatorics and Graph theory, Matrix Analysis, Random Processes, Internet of things, Signal Processing, Modulation Techniques, Information science

Work Experience

Bosch Global Software Technologies *Banglore, India*

SENIOR ENGINEER, Bosch AIShield *Aug 2021 – July 2023*

- Led *responsible AI research*, ensuring the development of AI algorithms aligned with strict adherence to ethical values. Specialized in *AI security*, implementing robust measures to safeguard AI models against a wide range of vulnerabilities.
- Conducted comprehensive vulnerability assessments and defense analyses, prioritizing factors like *explainability*, *robustness*, *fairness*, and *drift* across AI model applications, including Image Classification, Segmentation, Object Detection, Time Series Analysis, and Language models against *adversarial attacks, poisoning, extraction, and inference attacks*.
- Played a pivotal role in pioneering AI security research during the early phases of *Large Language Models (LLMs)* development, with a specific focus on countering Jailbreaking and prompt injection attacks. Proposed a defensive strategy to safeguard LLMs from aforementioned attacks. These contributions laid the groundwork for the creation of *AIShield Guardian*.
- Actively contributed to *product development* at Bosch AIShield, with responsibilities spanning microservices, pipelines, and logging. Demonstrated excellence in managing *customer partnerships* with industry leaders such as Whylabs and ClearML. Fostered productive *internal collaborations* across healthcare, automotive, and BFSI sectors.
- Successfully translated research outcomes into tangible achievements, including significant product enhancements, the publication of high-impact *technical papers*, and the acquisition of *valuable patents*.

GE Digital *Banglore, India*

Software Development intern *May 2020 – July 2020*

- Enhanced web application translation by implementing cutting-edge language models including *encoder-decoder, transformers*, and *BERT*. This innovative approach replaced the conventional method of translation reliant on XML and JSON.
- Orchestrated the deployment of the web interface, seamlessly integrating it with a *Flask API*.

Publications

[1] Targeted attacks on Time Series Forecasting *under review*

Yuvaraj Govindarajulu, **Avinash Amballa**, Pavan Kulkarni, Manojkumar Parmar *in ACML 2023*

[2] Discrete Control in Real-World Driving Environments using Deep Reinforcement Learning *arxiv preprint*

Avinash Amballa, Advaith P, PRADIP SASMAL, Sumohana Channappayya *2211.15920*

Patents

[1] A Method to detect AI poisoning attacks from the Data and/or Model *IN Patent: 202241068482*

Avinash Amballa, Yuvaraj Govindarajulu, Manojkumar Parmar *docket number: 404446*

[2] A Method of Targeted Attack on Timeseries Models to alter the DIRECTION of the Model Output (A method of assessing vulnerability of an AI system and a framework thereof) *IN Patent: 202241065028*

Yuvaraj Govindarajulu, **Avinash Amballa**, Manojkumar Parmar *docket number: 403873*

[3] A Method of Targeted Attack on Timeseries Models to alter the MAGNITUDE of the Model Output (A method of assessing vulnerability of an AI system and a framework thereof) *IN Patent: 202241065034*

Yuvaraj Govindarajulu, **Avinash Amballa**, Manojkumar Parmar *docket number: 403874*

[4] A Method of Sponze attack on Deep Learning Models to increase the inference time (A method of assessing vulnerability of an AI system and a framework thereof) *IN Patent: in progress*

Avinash Amballa, Yuvaraj Govindarajulu, Manojkumar Parmar *docket number: ongoing*

Projects

AlphaConnect-4

Prof Vineeth N Balasubramanian *Jan 2020 - Apr 2020*

- Inspired by deep mind's AlphaGo, implemented competitive *Multi-agent Reinforcement Learning* on Connect-4 game env
- Utilized a combination of *Monte Carlo Tree Search (MCTS)* for opponent modeling and *Policy Gradients* for agent reinforcement (single agent and single opponent)). Designed the game environment as well.
- Achieved impressive results by training the agent on low-dimensional game boards and successfully applied transfer learning techniques to enable the agent's performance in higher-dimensional environments, all with minimal additional training.

Gyro Correction in IMU sensors

Prof. K Sri Rama Murty, DRDO India (Defence Research and Development Organisation) *Apr 2021 - Jul 2021*

- Spearheaded the creation of a gyro correction model for IMU sensors to mitigate noise and axis misalignment issues.
- Leveraged diverse architectural approaches, including *DB-LSTM, LSTM with attention mechanisms*, and *Transformer Encoder* coupled with Huber Loss, while conducting rigorous training on the EUROC dataset.
- Through meticulous *hyperparameter optimization*, achieved superior performance with attention-based models, surpassing the capabilities of existing Dilated CNN methods in this domain.

Explaining Adversarial Examples $\&$ **Robustness**

Prof. Aditya T Siripuram *Jan 2021 - Apr 2021*

- Visual Explanations: Utilized variants of *Grad-CAM* and *GRAD-FAM* techniques to produce insightful visual explanations for adversarial samples. Analyzed the behaviors of Convolutional layers to enhance model interpretability.
- Frequency Domain Analysis: Conducted in-depth research into the frequency domain analysis of adversarial examples employing *Fourier transforms* and *filters*. Evaluated the robustness of models against adversarial attacks.
- Complex-Valued Neural Networks: Currently involved in ongoing research focused on explaining adversarial examples within a frequency and complex space using *complex valued neural networks*.

VICAP: VIdeo Captioning And Prediction

Prof. Aditya T Siripuram *Sep 2020 - Dec 2020*

- Implemented a *vision-language* video captioning method utilizing CNN with a DB-LSTM encoder-decoder architecture.
- Engineered a three-step search algorithm, employing Optical Flow techniques, to predict missing frames within video sequences. Additionally, utilized conditional *Generative Adversarial Networks (GANs)* for further frame prediction accuracy.
- Currently expanding capabilities in predicting missing frames within videos by exploring *self-supervised learning*.

Articles

[1] Medium article : ChatGPT - The future of Conversational AI

Skills

Coding	C, C++, Python, Java, R, Arduino, MATLAB, Latex
AI/ML	Tensorflow, PyTorch, Scikit-learn, OpenCV, openAI gym
Web Dev	HTML, CSS, JavaScript, jQuery, flask, Node.js, Express.js
Misc.	PostgreSQL, Azure, Git, Docker, AKS, Unity, Elasticsearch, Nginx, Open Source, NPTEL Deep Learning, Coursera's Google Data Analytics, Coursera's Deep Learning Specialization

Awards and Achievements

2022 **Promising Startup award** Bosch AIShield at Bosch FitFest

2022 **Global Info Sec award** Bosch AIShield

2018-19 **Showcased my works at Inter IIT** Tech Meet - 2018 at IIT Bombay and Tech Meet - 2019 at IIT Roorkee

2017 **Secured** 12^{th} **rank nationwide** in the KL University exam and received a prize worth 75k INR

Positions of Responsibility

2020 **Research Assistant** under Prof. Channapayya and Siripuram at IIT-H

2019 **Teaching Assistant** for the course Digital Signal Processing under Prof. K Sri Rama Murty at IIT-H

2018-19 **Core member** of Elektronica(Electronics, AI Club) and Cepheid(Astronomy, Astrophysics Club) at IIT-H

2018-19 **Security Coordinator** at IIT-H tech and cultural fest "ElanNvision"

Other Interests

Travel, Cricket, Photography $\&$ editing, Astrophysics, and Quantum Mechanics