

Soumyajyoti Dutta

Email: soumyajyotidutta23@gmail.com | Contact: (979) 326-2896 | GitHub | LinkedIn

Machine Learning for Cybersecurity / Small and Large Language Models / Automated Program Synthesis

Education

Texas A&M University , College Station, TX <i>Ph.D. in Computer Science, Department of CSCE</i> Advisor: Dr. Marcus Botacin	Spring 2024 – Fall 2029
Thesis: LLMs for Automated Rule Generation and Model Interpretability in Cyberdefense.	
Texas A&M University , College Station, TX <i>M.S. in Computer Engineering, Department of ECE</i>	Fall 2022 – Fall 2023
Coursework: ML Theory, Deep Learning, Data Mining, Randomized Algorithms, ML-Based Cyberdefenses.	
SRM University , Chennai, India <i>B.Tech. in Electronics and Communication Engineering</i>	Fall 2017 – Spring 2021

Expertise

Core Competencies:

- **End-to-end ML pipelines:** large-scale datasets, feature engineering, training, evaluation, deployment.
- **Classical ML:** classification, clustering, regression, anomaly detection, forecasting.
- **Distributed training:** trained 50+ models on multi-node GPU clusters (**DDP/NCCL optimization**).
- **Data analysis:** statistical analysis and visualization for actionable insights.
- **Systems optimization:** cache design, tokenization efficiency, memory-optimized loaders.

Technical Stack:

Python (NumPy, Pandas, scikit-learn, Matplotlib, PyTorch, TensorFlow, HuggingFace), C++, Rust, YARA, Django/Flask, Git, AWS, SLURM

Professional Experience

Texas A&M University <i>Research Assistant (Affiliation: Botacin's Lab)</i>	<i>College Station, TX</i> Spring 2024 – Present
<ul style="list-style-type: none">• Built LLM-based YARA rule generation pipelines using reinforcement curriculum learning and AST-embedded tokenization.<ul style="list-style-type: none">– Developed YARA syntax-aware tokenizer; reduced token-space by 10×.– Introduced curriculum-based difficulty scheduling using rule types and complexity tiers.• Designed Brownie & Puff Evaluation Function (BLEU + Parsing + SAT semantics) for training-time inference evaluation.<ul style="list-style-type: none">– Enabled fine-grained interpretability; benchmarked across T5, BART, LLAMA.• Extended Halstead & Cyclomatic metrics with YARA Control Flow Graphs for learning-curve analysis.• Built a scalable dataset framework (translation, patching, IOC detection) producing 100M+ labeled rules.• AutoPYara (open-sourced, IEEE Euro S&P 2026 submission):<ul style="list-style-type: none">– Re-implemented biclustering-based rule generation (AutoYara); evaluated multiple clustering algorithms.– Developed heuristics for optimal K selection and introduced Augmented K-means.– Provided foundational framework for future AutoPYara research expansions.– Achieved 14% performance gain via optimized centroid selection, 10% via sub-clustering, and 8% improved generalization in low-sample threat-hunting scenarios.	

Texas A&M University <i>Graduate Student Researcher</i>	<i>College Station, TX</i> Summer 2023
---	---

- Mathematical modeling of **prostate cancer genomic pathways**; simulated multi-drug interactions.
- Supported biologists through computational simulations for model validation.

Cognizant Technology Solutions <i>Junior Software Engineer</i>	<i>Kolkata, India</i> 2021–2022
<ul style="list-style-type: none">• Developed scalable full-stack web applications for logistics automation (Client: TJX Companies).• Optimized REST endpoints for latency & concurrency and collaborated with QA/DevOps on CI/CD pipelines.	

Projects

- **Machine Learning Malware Detection Model (1st Place)** — End-to-end ML system for Windows PE classification.
 - Integrated **EMBER**, **BODMAS**, **Benign-NET** datasets (2M+ samples).
 - Extracted PE-header and byte-level features using **LightGBM/XGBoost**.
 - Achieved **96.2% accuracy, 0.97 F1**; designed best-performing adversarial attacks.
- **Multiperspective Hawkeye** — Implemented ISCA'16 cache replacement policy in zsim; extended with PC-tracking and multi-policy benchmarking.
- **HelloPentagon** — Explainable ML-based malware defense chatbot integrating ML classification with LLM triage.
- **Carotid Artery USG Analysis** — CNN-based ultrasound classification (Springer Chapter, 2022).
- **WildFire** — Spatiotemporal ML modeling for wildfire spread prediction using climate and satellite data.
- **Fashion-MNIST** — CNN classifier comparing ResNet/VGG architectures with regularization analysis.

Teaching & Outreach

- **Teaching Assistant (Upcoming), CSCE 439 — Data Analytics for Cybersecurity (Spring 2026):** Core data analytics foundations including clustering, supervised ML, anomaly detection, and security-focused data visualization applied to attacks, vulnerabilities, and adversarial behaviors.
- **Partner & Presenter, TAMU CS Day (2024, 2025):** Conducted interactive demos on cybersecurity and AI research for high school students.
- **Teaching Assistant, ECEN 325 – Electronics (Summer 2023):** Supported lectures, grading, and lab sessions.
- **Technical Coordinator, TAMU Student Research Week (Spring 2023):** Automated judging and submission infrastructure.