Bao Gia Doan, Ph.D.

Australia giabaodoan1320@gmail.com baodoan.net

baogiadoan baogiadoan Google Scholar

I am a Postdoctoral Research Fellow at UNSW Sydney specializing in AI systems security, adversarial machine learning, trustworthy AI, and the security evaluation of AI agents. My research develops robust and auditable AI systems for security-critical and defence applications, spanning AI attack surfaces, RAG/LLM security, malware detection, backdoor attacks, and physical-world adversarial threats.

SUMMARY

- Over five years of research experience in AI systems security, adversarial machine learning, cybersecurity, trustworthy AI, computer vision, natural language processing, and data mining.
- Developed defensive strategies against AI attack surfaces including evasion, backdoor, malware, agent/RAG reliability, and physical-world attacks using Bayesian learning, adversarial robustness, and security-focused model evaluation.
- Collaborated with industry partners and defence stakeholders on applied AI security, trustworthy machine learning, and real-world deployment challenges.
- Coordinated and taught university-level machine learning with course administration experience for a cohort of more than 300 students, including tutor hiring, assessment, grading, and student consultation.
- Published in top-tier venues and journals including ICML, AAAI, RAID, ACSAC, IEEE Transactions on Information Forensics and Security, and IEEE Transactions on Dependable and Secure Computing.
- Served as Program Committee member for ICLR, AAAI, NeurIPS, ICML, ECCV, ICCV, and CVPR.
- Skilled in Python, PyTorch, TensorFlow, Keras, Scikit-Learn, Hugging Face, MLflow, Docker, LangChain, NumPy, Pandas, Matplotlib, Seaborn, and SQL.

TEACHING & SUPERVISING EXPERIENCE

Supervisor

University of New South Wales – Sydney
School of Computer Science & Engineering Sydney, NSW

Research Project A/B/C Semester 1, 2026

- Supervising student research projects in AI security, trustworthy machine learning, and applied cyber security.

Main Lecturer - Course Coordinator

University of Adelaide
School of Computer Mathematical Sciences Adelaide, SA

Introduction to Statistical Machine Learning (ISML) Semester 2, 2024

- Coordinated and administered a large undergraduate/postgraduate machine learning course with more than 300 students.
- Managed teaching operations including tutor recruitment, teaching team coordination, assessment design, grading, consultation sessions, and student support.
- Delivered lectures and learning activities on statistical machine learning with emphasis on rigorous experimentation and practical model evaluation.

Guest Lecturer

University of Adelaide
School of Computer Mathematical Sciences Adelaide, SA
Introduction to Statistical Machine Learning (ISML): Deep Learning October, 2022

- Delivered guest lecture material on deep learning concepts and applications.

Teaching Assistant

University of Adelaide Adelaide, SA
School of Computer Mathematical Sciences 2020-2023
Mobile & Wireless System

- Supported tutorials, assessment, marking, consultation, and student learning activities.

RESEARCH AND WORK EXPERIENCE

UNSW Sydney Sydney, NSW
Postdoctoral Research Fellow July 2025 – Present

- Conducting research on AI systems security, including the attack surfaces of AI agents, RAG systems, and large language models for defence and other security-critical applications.
- Leading research on domain-grounded RAG assessment for defence documents, including the DoRA benchmark with 6.5K curated, intent-conditioned QA instances and auditable evidence passages.
- Improved RAG QA task success by up to 26% and reduced hallucination rate by 47% via DoRA-supervised fine-tuning.
- Collaborating with industry partners and defence stakeholders on applied AI security, evaluation, and research translation.

University of Adelaide Adelaide, SA
Postdoctoral Research Fellow August 2022 – March 2025

- Conducted pioneering research in adversarial machine learning, significantly enhancing cybersecurity defenses with highly cited publications.
- Secured competitive grant funding totaling $300K through collaborations with UNSW and DSTG, delivering advanced AI solutions for national security.
- Led and mentored research teams, achieving breakthrough results in adversarial learning algorithms and publishing extensively in prestigious A* conferences (e.g., AAAI, ICML).

Intel Ho Chi Minh City, Vietnam
Senior Process and Equipment Engineer July 2014 – September 2018

- Translated extensive manufacturing data into strategic insights to support critical decision-making processes.
- Directed continuous improvement initiatives, significantly enhancing factory performance and reducing operational expenses.
- Managed technology integration and transfer of new Intel products from global sites to Intel Vietnam.

EDUCATION

University of Adelaide Adelaide, SA
Ph.D. Computer Science 2018 - 2022
Dissertation: *Towards Robust Deep Neural Networks*
Awarded the **Dean's Commendation for Doctoral Thesis Excellence**.

Royal Melbourne Institute of Technology (RMIT) Saigon South Campus, Vietnam
Master of Engineering 2012 - 2014
Major: *Electronic and Computer Engineering*
Graduated with **Distinction** (the highest possible award).

Danang University of Technology Danang, Vietnam
Advanced Program - Bachelor of Electronics and Communication Engineering 2007 - 2012
Major: *Digital Systems*
Graduated with **Excellent**. Top 5% among 1000 students graduating in 2012 at DUT.

SELECTIVE PUBLICATIONS

- **Doan, B.G.**, Joshi, A., Elinas, P., Bodhankar, A., Leslie, O., Marchant, T. and Salim, F., 2026. Domain-oriented RAG Assessment (DoRA): Synthetic Benchmarking for RAG-based Question Answering on Defense Documents. *arXiv preprint arXiv:2604.17943*.
- **Doan, B.G.**, Jokandan, A.S., Guo, X-Y., Mohammadi, A., Rokny, H., Sejdinovic, D., Ranasinghe, D.C. and Abbasnejad, E., 2025. Bayesian Low-Rank LeArning (Bella): A Practical Approach to Bayesian Neural Networks. In the 39th *AAAI Conference on Artificial Intelligence (AAAI)*. (CORE Rank: A*)
- **Doan, B.G.**, Yang, S., Montague, P., De Vel, O., Abraham, T., Camtepe, S., Kanhere, S.S., Abbasnejad, E. and Ranasinghe, D.C., 2023. Improved the Robustness of Malware Detectors with Bayesian Neural Networks. In the 37th *AAAI Conference on Artificial Intelligence (AAAI)*. (CORE Rank: A*)
- **Doan, B.G.**, Abbasnejad, E., Shi, J.Q. and Ranasinghe, D.C., 2022. Bayesian Learning with Information Gain Provably Bounds Risk for a Robust Adversarial Defense. In the 39th *International Conference on Machine Learning (ICML)*. (CORE Rank: A*)
- **Doan, B.G.**, Nguyen, D.Q., Montague, P., De Vel, O., Abraham, T., Camtepe, S., Kanhere, S.S., Abbasnejad, E. and Ranasinghe, D.C., 2024. On the Credibility of Backdoor Attacks Against Object Detectors in the Physical World. In the 40th *Annual Computer Security Applications Conference (ACSAC)*. (CORE Rank: A)
- **Doan, B.G.**, Nguyen, D.Q., Montague, P., De Vel, O., Abraham, T., Camtepe, S., Kanhere, S.S., Abbasnejad, E. and Ranasinghe, D.C., 2024. Bayesian Learned Models Can Detect Adversarial Malware For Free. In the 29th *European Symposium on Research in Computer Security (ESORICS)*. (CORE Rank: A)
- **Doan, B.G.**, Xue, M., Ma, S., Abbasnejad, E. and Ranasinghe, D.C., 2022. TnT Attacks! Universal Naturalistic Adversarial Patches Against Deep Neural Network Systems. *IEEE Transactions on Information Forensics and Security*. (SCImagoJR Rank: Q1, Impact Factor: 7.178)
- Yang, S., **Doan, B.G.**, Montague, P., De Vel, O., Abraham, T., Camtepe, S., Ranasinghe, D.C. and Kanhere, S.S., 2022. Transferable Graph Backdoor Attack. In the 25th *International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*. (CORE Rank: A)
- Gao, Y., Kim, Y., **Doan, B.G.**, Zhang, Z., Zhang, G., Nepal, S., Ranasinghe, D.C. and Kim, H., 2021. Design and Evaluation of a Multi-Domain Trojan Detection Method on Deep Neural Networks. *IEEE Transactions on Dependable and Secure Computing*. (SCImagoJR Rank: Q1, Impact Factor: 7.329)
- **Doan, B.G.**, Abbasnejad, E. and Ranasinghe, D.C., 2020. Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems. In the 36th *Annual Computer Security Applications Conference (ACSAC)*. (CORE Rank: A)

GRANTS AND AWARDS

Chief Investigator, "Enhancing Robustness of Machine Learning Models", collaborative project with DSTG, University of Adelaide, 2024–2025 - Awarded: $300,000.

Dean's Commendation for Thesis Excellence - University of Adelaide, 2022 - Recognized for outstanding academic performance and dedication.

Faculty of Engineering, Computer and Mathematical Sciences (ECMS HDR) Traveling Grant - University of Adelaide, 2020 - Awarded: $3,000.

International Wildcard Ph.D. Scholarship - University of Adelaide, 2018-2022 - Prestigious scholarship covering tuition fees and stipends for the entire Ph.D. program - Awarded: $278,000.

Intel Factory and Department Recognition Awards - Intel Corporation, 2014-2018 - Multiple awards recognizing significant contributions to company success at various levels, from departmental to factory-wide.

Intel Vietnam Master Scholarship - RMIT University, 2012-2014 - Full scholarship recipient, selected from thousands of applicants at top Vietnamese universities - Awarded: $50,000.

Texas Instruments Vietnam University MCU Design Contest - Grand Prize Winner (Central Region), 2011-2012.

Student Scientific Research Conference (Danang University) - Recognized for outstanding results, 2012.

Petro Vietnam Scholarship - Awarded for outstanding academic performance (top 5%) - Full tuition fee coverage for entire university education, 2010-2011.

Intel Vietnam Engineering Scholarship - Recognized for outstanding academic performance and dedication - Full tuition fee coverage for the English language course, 2010.

TECHNICAL SKILLS

Programming and Machine Learning

- Proficient with Python and popular ML frameworks including PyTorch, TensorFlow, Keras, and Scikit-Learn for developing and deploying machine learning models.
- Skilled in using Hugging Face, MLflow, Docker, and LangChain for model optimization and scalable AI deployments.
- Experienced in leveraging Gemini 2.5 Pro and advanced transformer-based models such as ModernBERT.

Data Manipulation and Analysis

- Expert in data processing and analysis using NumPy, Pandas, and SQL.
- Skilled in automated data sanitization using Cleanlab and labeling functions with SnorkelAI.

Data Visualization

- Proficient with OpenCV, Matplotlib, Seaborn, and Plotly for effective data visualization and insights presentation.

Version Control and DevOps

- Skilled with Git, GitHub, and BitBucket for collaborative software development and version control.
- Experienced in building continuous integration and continuous deployment (CI/CD) pipelines utilizing Docker, MLflow, and cloud platforms.

Research Software and Skills

- Advanced proficiency in LaTeX for academic writing and technical documentation.
- Experienced in experimental design and data annotation through Amazon Mechanical Turk.

PROFESSIONAL SERVICE

Journal Reviewer 2021-present

IEEE Transactions on Information Forensics and Security

Journal of Machine Learning Research

Neural Computing & Applications

Program Committee 2022-present

- 2026: ICLR (CORE Rank: A*); AAAI (CORE Rank: A*) and AAAI AI Alignment special track.
- 2025: NeurIPS (CORE Rank: A*); ICML (CORE Rank: A*).
- 2024: NeurIPS, ICML, ICLR, ECCV, and AAAI.
- 2023: NeurIPS, ICCV, and AAAI.
- 2022: ICML, ECCV, and CVPR.

PROFESSIONAL MEMBERSHIPS

Association for the Advancement of Artificial Intelligence
Institute of Electrical and Electronics Engineers (IEEE)

PRESS

Unite AI
“Why Adversarial Image Attacks Are No Joke”
(https://www.unite.ai/why-adversarial-image-attacks-are-no-joke/)

Gigazine
“TnT, a new attack method for image recognition AI that misleads a picture of a flower into the face of former President Obama”
(https://gigazine.net/gsc_news/en/20211130-universal-naturalistic-adversarial-patches/)

REFERENCES

Available upon request