Postgraduate Seminar Series

Venue: Graduate Seminar Room
Date & Time: June 13, 2026 at 3:00 PM

Speaker Information

Abdur Rashid Tushar (Std No. 0422052001) is a part time M.Sc. student in the department of CSE, BUET. He completed his undergraduate studies at Bangladesh University of Engineering and Technology (BUET) in 2022. His research interest lies in the field(s) of Computer Security and Artificial Intelligence. He is currently doing his postgraduate thesis under the supervision of Sukarna Barua and Dr. M Sohel Rahman. He will be speaking about hisongoing research in this talk.

Enhancing Malware Family Classification using Supervised Contrastive Learning: A Grouped-Feature Transformer Approach

This study enhances supervised contrastive learning for malware family classification on the EMBER2024 static-feature corpus. EMBER2024 violates three assumptions of the standard supervised contrastive recipe of Khosla et al. (2020) in measurable ways: 52 % of the family labels have ClarAVy confidence below 0.5, the per-dimension family information differs by approximately 60× across the twelve EMBER feature groups, and the per-week test macro-F1 of the strongest static baseline decreases by 0.0064 per week across the twelve-week test horizon, because malware families evolve over time. We modify the supervised contrastive recipe at the four points where it interacts with the data. For the encoder, we propose the Grouped-Feature Transformer (GFT), which tokenises the 2,568-dimensional EMBER input at the level of the twelve feature groups through a Group-Wise Feature Embedding (GWFE) stage, and which uses a Global Encoder Representation Vector (GERV) as the only contrastive readout. For the loss, we propose TH-SupCon (Temporal–Hard-negative Weighted Supervised Contrastive Loss), a composite extension of Khosla et al. along four structural axes: a Temporal Drift Factor (TDF) that anchors per-instance recency to the per-class latest training week, top-K = 50 Top-k Negative Sampling (TNS) which is directed at the confusable infostealer cluster, a self-normalising per-pair weighting that makes the loss scale-invariant in the weight kernel, and a Label Confidence Factor (LCF) which down-weights low-confidence positive pairs. For the augmentation, we propose TAMA (Type-Aware Malware Augmentation), a dimension-aware regime in which the operators are dispatched per (group, sub-block) data-type, so that every augmented view is a valid on-manifold malware feature vector by construction. We also propose its multi-resolution extension TAMA-MR, which renders each anchor at two perturbation resolutions simultaneously, in order to train invariance across a range of feature fidelities. For inference, we propose MDHF (Multi-Depth Head Fusion), which trains one lightweight Dual-Stream Head per intermediate transformer layer of a single SupCon-trained encoder, and which combines the head softmaxes by Product-of-Experts fusion. The headline architecture, MTF (Multi-Scale Temporal Fusion), is the equal-weight Product-of-Experts fusion of six GFT-MDHF pipelines which differ only in the TH-SupCon recency hyperparameter $\alpha \in \{0.6, 0.7, 0.8, 0.9, 0.99, 1.0\}$. On the EMBER2024 412-class temporal split, MTF obtains macro-F1 = 0.6358, which is +1.53 points above our re-trained XGBoost baseline at 0.6205, and +2.34 points above our re-trained LightGBM baseline at 0.6124; both baselines use 412 classes with ≥ 500 training samples per class.