

Postgraduate Seminar Series

Venue: Graduate Seminar Room

Date & Time: October 17, 2026 at 3:15 PM

Speaker Information

Md. Nazmul Hasan Sakib (Std No. 0422052005) is a part time M.Sc. student in the department of CSE, BUET. He completed his undergraduate studies from Bangladesh University of Engineering & Technology (BUET) in 2021. His research interest lies in the field(s) of Cybersecurity, Machine Learning, and Deep Learning. He is currently doing his postgraduate thesis under the supervision of Dr. A. B. M. Alim Al Islam. He will be speaking about his ongoing research in this talk.



A HYBRID RESCUER DEFENSE MECHANISM FOR PREDICTIVE MODELS AGAINST A NOVEL DATA POISONING ATTACK BASED ON RANDOMIZATION

This study develops a data poisoning attack against predictive models with continuous targets, and a defense that counters it. Such models now set medical doses, forecast electricity demand and value property, so an attacker who reaches their training data can steer decisions that matter. Almost all work on poisoning concerns classification, and the defenses built for continuous targets filter using the largest residuals. A label drawn to imitate the clean data therefore survives, and because the discarded subset is of fixed size, clean data fills the quota instead. We treat the defense itself as the object of study. Running eight randomized attacks against the state of the art iTrim defense, over thirteen datasets and nine predictive models, records what the filter decides about every training point. Labeling each point by whether it is poisoned and whether the filter removed it gives four outcome classes, and a dataset of the defense's own behavior. A Set Transformer learns that behavior by reading the training set as a whole. The trained model then serves twice. As an attacker it searches for labels that damage the model and evade the filter at once. As a defender it becomes the Hybrid Rescuer, overturning the decisions of iTrim it judges wrong with high confidence. The surrogate reproduces the filtering decisions with a weighted F1 of 0.978, so the behavior of a residual-based defense is learnable. The attack built on it survives filtering better than any randomized variant and flip attacks. The defense lowers the attack success rate and the false attack rate together, because it purges poison the filter kept and also restores clean data the filter discarded, which no prior defense attempts.