MSc Project Literature Review

When doing and writing a literature review, it is good practice to:

* summarise and analyse previous research and theories;
* identify areas of controversy and contested claims;
* highlight any gaps that may exist in research to date.

|  |
| --- |
| **Technical Approaches** |
| Initital literature reviews mention SVM & CNN (??) being used to model the SCADA systems. |
|  |
| StatQuest- Support Vector Machines  Stamer explains intuition of Support Vector Classifers when applied to data that can be partitioned using a single ………… . Furthermore he introduces the concept of ……… which create a soft area which allows for misclassification and outliers. The values for the soft area are claculated using cross validfation to calculate the optimum offset?? .  SVM is introduced as a solution when classifications can’t be split using a single line??? In their original dimension. The SVM moves the data to a higher dimensional reperesentation so that that SVC can be applied.  The choice of dimension can use a kernel function such as the Polynomial |
|  |
| **Adversarial Attacks and Mitigation for Anomaly Detectors of Cyber-Physical Systems**  <https://www.researchgate.net/publication/351891007_Adversarial_Attacks_and_Mitigation_for_Anomaly_Detectors_of_Cyber-Physical_Systems> |
| Chen et al describe the methods used by CPS to identify anomalous behaviour which is indicative of a cyber attack. They describe that typically a CPS has two forms of defence: Firstly, an anomaly detector which is a Machine Learning model (often based on a neural network model) which is trained on the systems physical data. Secondly, rule checkers ( or invariant checkers) are used which check values against the acceptable parameters or known relationships between components in the CPS  .  Chen et al assume a ‘white’ box level of access to the anomaly detector, that is a full understanding of it’s behavious and accesst to the data it was trained on .It is assumed the rule checker is only a black or grey level of access so it’s behaviour must be learnt from the librarian logs etc..  The team ‘crafts noise’ over the signal between actuators and sensors then use a ‘genetic algorithm’ to optimise the noise so that both detection systems are deceived to the degree that their classification accuracy is reduced by over 50%.  The report mentions how attacks on the CPS typically involve spoofing or manipulating the network packets and neural network based detectors are effective at identifying these. This paper seeks to create attack possible when there is ‘insider’ level access- the attacker knows the anomaly model.  The focus of the paper is to create noise which will lead the anomaly detector and rules checkers to misclassify the activity.  For example, if the attack can use noise to shrink the difference between the actual value and the predicted value then the anomaly detector will assert more false positives “when a detector misclassifies a real attack as normal behaviour” (Yifan Jiaa, 2021). Jiaa et al assert that “existing adversarial attacks have limited effectiveness in the presence of rule checkers” but that genetic algorithms based on the white-box gradient based approach can remedy this.  The paper defines a CPS as PLC’s which are connected to actuators and sensors which are the interface to the physical world. The PLC run software for the control logic of these peripherals which it is connected to by a circular/ ring network operating at ‘Layer 0’. Layer 0 is taken to be at a photon/ electron level- i.e. sub-bit level so continuous or discrete signals.  These PLCs are connected to a central SCADA system by a star network operating at layer 1- the physical layer.  It is assumed that rule checkers reside within the PLCs- for example to open a valve using an actuator when a particular sensor value is met. The anomaly detector is assumed to reside on the SCADA system.  The paper describes the two test beds used for the research – the Swat & WADI plants that model a water treatement and water distribution plants respectively.  The SWAT plant is described as having 68 sensors and actuators in total, a number of these are standby in case of failures and were not cond=sidered in the paper. It is noted that the sensors are typically continuous values and the acutators are discrete. This is understandable as the output of the PLC is likely to a motor controlller or relay which handle things like soft start for motors or gradual closing of valves in order to avoid the water hammer effect (me).  **‘Our approach is inspired by a white-box gradient-based approach [33],’ N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, A. Swami, The limitations of deep learning in adversarial settings, in: 2016 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2016, pp. 372–387** |
| (Yifan Jiaa, 2021) |
| The SWat testbed has a historian which records the physical state of the system, this is “a fixed ordering of all the sensor readings and actuator configurations at a particular timepoint” (Yifan Jiaa, 2021). The report uses the following notation to denote the system state (***x***) where subscript ***a*** and ***s*** are used for acutators and sensors respectively. |
| (Yifan Jiaa, 2021) |
| SCADA, Historian and Human Machine Interface workstations sit at higher levels and allow operations such as changing control code/ parameters within the PLCs.  The Threat model used is White Box where attacker has access to physical signals at layer 0, full knowledge of the RNN based anomaly detector but can only judge rule checker from ***Status*** in the historian.  The authors use gradient based methods where by the original attack signals have noise added which is basedon the loss gradient of the RNN. This does not affect the attack but leads to the attack being misclassified. |
|  |
|  |
|  |
|  |
|  |
| **Project Plan** |
| Literature Review- Whats is CPS, How are attacks made, How are they defended against,  Approx System Design- diagrams, types of control sytems, types of comms, network topographies  Define Threat Model Attack will be based on- White box advesarial attack  Methosd for learning the system- SVM/ Neural networks…..THESE ARE EFFECTIVELY ANOMALY DETECTORS???? Compare predicted outcome of model with reality if attacks are successful- would mean successfully spoofing sensor values so one model has ‘real’ values and other has spoofed values.  Methods for preparing attacks- Gradient-based approach?? Genetic Algorithms???  Learn the system (assume all done at application layre??) Find min & max parameters values from  Dataset.  Define what constitutes a succesfull attack- identify vulnerabl bits of system…. Osmosis? Ph? Learn/ verify system by varying inputs and checking outputs change?  Devlop attacks which are detectable, use genetic??? To sucessfully  Investigate techniques to hide attack  Evaluate against testbed at Uni  Write Up  Present |