SAND2015-8850C

Evolving Decision Trees to Detect Anomalies in Recurrent ICS Networks

Jasenko Hosic, Jereme Lamps, Derek H. Hart
Sandia National Laboratories
Albuquerque, New Mexico 87123
Email: jhosic@sandia.gov, jlamps@sandia.gov, derhart@sandia.gov

***Abstract*—Researchers have previously attempted to apply machine learning techniques to network anomaly detection problems. Due to the staggering amount of variety that can occur in normal networks, as well as the difficulty in capturing realistic data sets for supervised learning or testing, the results have often been underwhelming. These challenges are far less pronounced when considering industrial control system (ICS) networks. The recurrent nature of these networks results in less noise and more consistent patterns for a machine learning algorithm to recognize. We propose a method of evolving decision trees through genetic programming (GP) in order to detect network anomalies, such as device outages. Our approach extracts over a dozen features from network packet captures and netflows, normalizes them, and relates them in decision trees using fuzzy logic operators. We used the trees to detect three specific network events from three different points on the network across a statistically significant number of runs and achieved 100% accuracy on five of the nine experiments. When the trees attempted to detect more challenging events at points of presence further from the occurrence, the accuracy averaged to above 98%. On cases where the trees were many hops away and not enough information was available, the accuracy dipped to roughly 50%, or that of a random search. Using our method, all of the evolutionary cycles of the GP algorithm are computed *a-priori*, allowing the best resultant trees to be deployed as semi-real-time sensors with little overhead. In order for the trees to perform optimally, buffered packets and flows need to be ingested at twenty minute intervals.**

I. Introduction

The ability to accurately identify intrusions or anomalies on a network has long been a goal of the security community [14], [16], [17]. Network administrators and security engineers alike want to know exactly what is occurring on their networks, but the amount of data flowing across all the nodes in large-scale networks is simply too great to manually examine. As a result, solutions have been created that analyze data in a number of automated ways. Network-based intrusion detection systems and host-based intrusion detection systems each have their merits, but they often depend on specifically crafted rules which can be circumvented. Signature-based approaches often suffer from being too specific to capture every kind of attack. Unique approaches involving machine learning or anomaly detection methods often achieve high accuracy rates, but still suffer from some imperfections. In a large-scale network scenario, the sheer volume of data generally renders these techniques unusable. Even with a 5% false positive rate, human experts would need to manually inspect an insurmountable number of false alarms.

Industrial control system (ICS) networks can be massive as well, but generally contain far less noise. The way in which programmable logic controllers (PLCs) and remote terminal units (RTUs) transmit data back to a front end processor (FEP) is cyclical. Usually data is transmitted at recurrent intervals, and this periodic nature is a much simpler pattern to define than the arbitrary chaos of a corporate network with endpoints controlled by users. We propose a machine learning approach that takes advantage of this fact in order to detect unwanted behavior on an ICS network.

Although using machine learning techniques on ICS network data was attempted in the past [10]–[13], the amount of complete solutions is sparse. Furthermore, the goals of prior work were fundamentally different. Rather than determining what type of undesired behavior occurred on the network, most of the research in this area focuses on identifying existing or new attacks. Our approach focuses on three key goals. First, we aim to detect the occurrence of specific effects or events, such as outages of devices on the network, not causes. Second, we want our solution to have a minimal presence on the network. Rather than considering all data flowing across every node in a network, we collect data flowing through one tap in the network, reducing the amount of processing power and space required to execute the solution. By extracting precise data about the traffic, it may be possible to glean key details about devices several hops away from the data collection tap. Third, we want to minimize the amount of information required to accurately determine if the event occurred. To show the validity of our approach, we tested our algorithm at different taps on the network with varying levels of access to the data pertinent to identifying each event.

Our proposed solution for meeting the aforementioned goals employs genetic programming (GP) to evolve decision trees. GP is an iterative, population-based meta-heuristic technique that follows an evolutionary cycle to produce solutions represented by trees or graphs. Multiple guesses at the solution are made, evaluated, and recombined to produce new solutions. After many such cycles emulating natural selection and evolution, a best solution is chosen [7]. In our approach, this reinforcement technique evaluates evolved trees based on their ability to use normalized extracted features to accurately identify specific events in a labeled training set. This type of technique was used to great effect in [6], and avoids some of

the problems with classic decision tree algorithms [15]. The purpose of the resultant decision trees is to act as sensors that take in buffered packet capture (PCAP) and netflow data, compute the feature extractions, and identify whether a specific event occurred. All of the machine learning iterations are performed *a-priori*, allowing the sensor trees to quickly flag buffers of data as soon as the feature extraction is complete.

In this paper, we focus on three experiments, each tested at three different locations in the network. Our test network is a standard 24-bus power network, virtualized on a server with virtual machines simulating ICS traffic such as modbus. We used Minimega, a tool developed by Sandia National Laboratories, to deploy the virtual network [1]. The three events that we tried to detect are: a specific router failing, any router in the network failing, and a specific FEP failing. Our points of presence were at different locations of the 24-bus network, with some being close to the network outages while others were several hops away. The majority of our results achieved 100% accuracy in detecting specific events, and only dipped below 98% when too little data was provided to the machine learning algorithm.

This method can be used to answer a number of different questions about the state of ICS networks, given a rich feature set and the right amount of training data. Is the recurrent pattern of communication broken? Has a large exfiltration of data occurred? Has the process logic or firmware on a PLC been updated across the network? The remainder of this paper explores our methodology and test data in greater detail. The GP parameters are explained, and an analysis of the nine experiment results is shown. We end with a look at potential future work regarding this idea.

II. Related Work

Using GP to develop new heuristics is not a unique concept. It has been explored with great success in prior works [2]–[5]. The novel aspect of our research lies in the application of the algorithm as well as the feature extractions. Others have proposed a similar approach [10], [12], [17], but there are some fundamental differences in the machine learning algorithms used and the goal they are trying to achieve. We used the effective portions of their work to enhance ours and create an accurate method for detecting anomalies.

Sommer et al. [14] provided some warnings about using machine learning to detect network intrusions. Their work focused on less recurrent networks, but the conclusions are important to consider nonetheless. The major claim is that machine learning has had limited success in the intrusion and anomaly detection domain because that problem type is fundamentally different than the kinds of problems with which machine learning algorithms excel. Machine learning is better at determining similarities than it is at finding new, meaningful outliers. This is perhaps why machine learning techniques such as support vector machines (SVM) have the greatest success in classification problems where the possible groupings are known. Sommer et al. go on to suggest that the few successes these techniques have experienced in the security domain, such as spam filtering, are a result of detecting variations of known attacks rather than new attacks altogether. Lastly, there exists a high cost for misclassifying data. False negatives are generally unacceptable, and false positives require a great human effort to resolve.

The implications of these challenges can be seen with the work of Yin et al. [18]. Their work was a great inspiration for our approach, but ultimately suffered from some of the problems described by Sommer et al. They used a GP approach to develop new rules for anomaly detection. The classic DARPA network data set was used to evolve the new rule set. This achieved great results, outperforming prior work by detecting 84 out of 148 events, but it is not practical in a real scenario with live networks. During their testing phase, an unacceptable number of attacks went undetected. Their algorithm required two passes, which resulted in some undesirable inefficiency. Despite these few issues, their findings were a great starting point, and highlighted some of the problems we sought to overcome.

Lu et al. [9] used the same data set as [18], but recognized some of the issues involved with it. The DARPA data has long been the standard in network attack testing data, but it is outdated and incomplete. It also is not intended to represent periodic ICS networks, so we have opted not to use it. Lu et al. propose a slight alternate method as well. The result of their findings is also excellent, as their false positive and false negative rates are slightly above 5%, which is still too high for practical applications, but suggests that a similar approach would perform even better on a recurrent network. The technique attempts to find variations of known attacks, as opposed to new attacks, as is recommended by Sommer et al.

The recent work from [10] is one of two that closely resembles our goals. They use Supervisory Control and Data Acquisition (SCADA) network data and SVM to identify attacks. Over 1500 packets of data are considered in their approach, and their results are impressive. With certain splits of their data, they achieve nearly 100% accuracy. They extracted several features, such as packet rate and packet size. While the effects of an attack are our primary interest, Maglaras et al. attempted to identify the attack itself. Their promising results were nonetheless an inspiration, and we attempted to capitalize on their progress. Our data sets are much larger and consider many more protocols (both ICS communication and otherwise).

The other research that attempts to identify anomalies in ICS or SCADA networks is from Mantere et al. [11]–[13]. Throughout their work, they discussed the challenges of using machine learning on recurrent networks, discovered valuable features to extract, and created a prototype that uses self-organizing maps (SOM) to relate their extracted features. To assist in the feature extraction process, Mantere et al. use Bro, a network security monitoring tool. Bro can aggregate data quickly and reduce some of the overhead associated with custom data extraction. While their results are mostly a proof of concept, initial testing demonstrates they can achieve as little as zero to three false positives per day. However,

they did not perform any attacks throughout their training and testing data, so these results are still theoretical. They used packet captures from a real ICS network running for many days, which was an essential step in representing the problem practically. While we did not have the same kind of access to live data, our emulation techniques provide a realistic alternative. The RTU and FEP communications accurately model live systems, allowing the GP algorithm to evolve sensors that could be placed in real systems after testing.

III. Methodology

Developing decision trees to act as sensors in a live network must first begin with selecting appropriate features to extract from the available data. We deployed a virtual 24-bus power network with over 140 nodes, including routers, RTUs, FEPs, and other components. The RTUs and FEPs communicated with ICS traffic such as modbus, and the routers establish dynamic routing through the OSPF protocol. If FEPs were unable to communicate with their corresponding RTUs due to network or device failure, they continuously tried to reestablish connections, greatly increasing the amount of traffic they generated. Figure 1 shows a sampling of our network structure. The same structure exists in multiple branches of the network. The rest was omitted for clarity. Figure 2 shows how devices are connected to the routers. The network contains multiple human-machine interface (HMI) workstations, an HMI server housing the HMI applications, and a historian that aggregates all RTU status updates and FEP interactions. Not all branches in the network have as many devices connected to the routers.

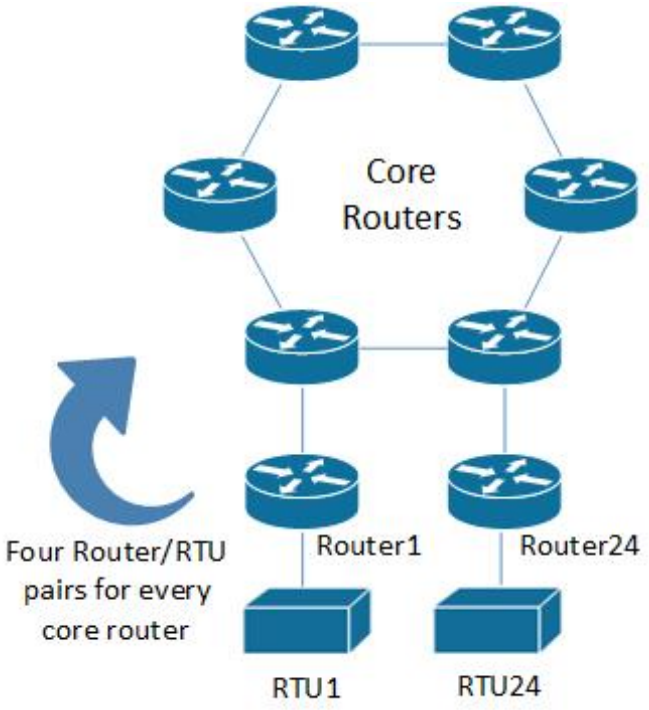

Fig. 1. A subset of our 24-bus network

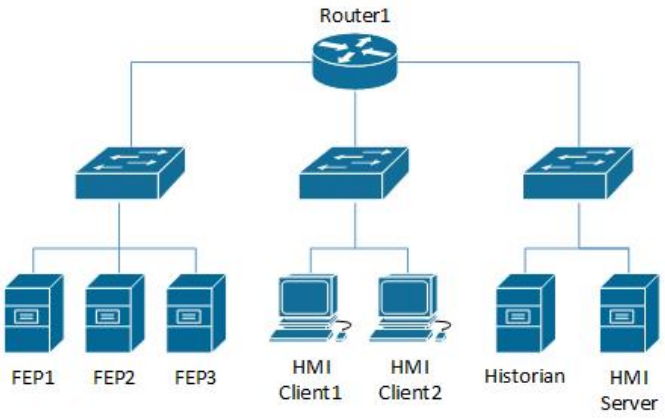

Fig. 2. An example of devices connected to a router in the 24-bus network

The goal was to detect anomalies in semi-real-time. True real-time detection is difficult to achieve because a large amount of data needs to be aggregated before useful features can be extracted. We tested our method with many shorter buffer sizes, but found the best accuracy with twenty minute buffers. Even larger buffers may have produced better results for our more difficult tests, but we wanted to minimize the amount of storage and processing required to implement our solution. Table I describes our nine experiments and the location of the data buffering tap in each experiment.

TABLE I
Experiment Descriptions

Exp. No.	Fail Event	Tap Location
1	Specific router	Next to target router
2	Specific router	Several hops, next to RTU
3	Specific router	Several hops, next to FEP
4	Any router	Next to router
5	Any router	Next to RTU
6	Any router	Next to FEP
7	Specific FEP	Several hops, next to router
8	Specific FEP	Several hops, next to RTU
9	Specific FEP	Several hops, next to diff. FEP

We performed data collection at each of the three tap locations listed in Table I, consisting of ten total hours, or several gigabytes, of PCAP and netflow traffic. Throughout the data gathering, we forced multiple device failures. Our intent was to truly stress the versatility of our feature extraction and machine learning, so many of the data sets we collected captured numerous outages occurring at one time, with many devices coming back up after several minutes. Table II shows a description of the ten data sets that we used to comprise our training and testing data sets. At the start of each data collection, the network was fully operational and in a steady state. In our descriptions, we define t_x to represent the xth minute for the data set. For example, t_5 defines the 5th minute of data collection for the particular data set.

TABLE II
DATA COLLECTION DESCRIPTIONS

Data No.	Data Description
1	No fail events for the 20 minute period
2	At t_{10} a router fails
3	At t_{10} a router fails At t_{17} it comes back online
4	At t_5 a router fails At t_{10} another router fails
5	At t_{10} a different router fails
6	At t_{10} a FEP fails
7	At t_{10} a FEP fails At t_{17} it comes back online
8	At t_7 a FEP and two routers fail At t_{14} one of the routers come back online
9	At t_5 a router fails At t_{10} it comes back online At t_{15} it fails again
10	At t_{10} a FEP and two routers fail

For these data sets to act as indicators in discerning the occurrence of the events, meaningful feature extraction algorithms need to be used. The feature extraction algorithms provide metrics that can be used by the GP to evolve the decision tree sensors. The following is a description of the features we extracted, followed by a comparison of their validity as discriminators of event occurrence.

A. Temporal Data

Five of the features we extracted were low-level temporal features. We computed the average number of packets sent by the target across the duration of the buffer. We also calculated the rate OSPF packets were being sent in cases where the target was a router. Another feature extracted the standard deviation of the number of packets sent at every minute of the buffer. We recorded the longest period of time a device would be silent. Lastly, we captured the average duration of each flow that the target initiated. The temporal data used in [10] and [11] was a great inspiration for these choices. If the device has a high packet rate early in the buffer and the rate drops, several of these features will be impacted. Some of our data sets attempt to circumvent these temporal features with short, periodic outages. If the device comes back up quickly after each outage, the average may not dip enough to be detected by the feature.

The importance of the flow duration feature is that it can detect differences in the types of communications used by the devices. It is possible for this feature to be helpful when discovering other, more nuanced network anomalies.

B. Communication Failure

We also wanted our sensors to capture obvious indicators of communication failure. The trees are given a count of the "ICMP unreachable" packets, incomplete handshakes (SYN, SYN/ACK, without the last ACK), and communication pattern breaks. The communication pattern breaks are calculated by determining the pattern at which the target communicates. For instance, routers have a pattern of sending OSPF Hello packets every 10 seconds, any deviations in this pattern can be flagged as a potential router outage and counted. This is particularly helpful due to the recurrent behavior of these networks. The FEPs generally receive RTU information at fixed intervals with little discrepancy.

C. Communication Profile

Lastly, we included some high-level features to capture a profile of the kinds of communications in the network. We captured the length of each packet sent by the target, as well as the time-to-live (TTL) of every packet coming across the tap. The packet size would differ if the content of the communication changed, and the TTL would hint at a change in path, likely due to an outage or new device on the network.

We also made use of collocations within our features. Collocations are a measure of how two things are usually grouped. The classic example of collocation involves language. The words "running" and "water" are much more likely to occur together than the words "walking" and "water". Likewise, "tall" and "tree" make more sense together than "high" and "tree". These words are therefore more strongly connected. To numerically represent these strong connections, a mutual information equation can be used. The formula for mutual information compares the probability of two objects occurring together if they are independent and the probability of their actual occurrence together [8]. The formula is shown in Equation 1.

$$I = P(XY) \cdot \log \frac{P(XY)}{P(X)P(Y)} \tag{1}$$

Within the netflows, we found strongly connected source-destination pairs occurring adjacent to each other and flagged any weak connections (based on an experimentally calculated threshold of 0.05). We used collocations for two separate PCAP features by first creating a label for each packet type, such as "ACK packet" or "OSPF Hello packet". The collocation feature then used these human-readable labels to determine what packet types often occur together. The feature was split in two, with the target as the source and as the destination. These features should indicate if packets that have previously not been seen, such as ICMP unreachable packets or new SYN packets attempting to establish connections, are captured.

D. Single Feature Tests

We ran these features individually on our data sets to validate the need for our GP approach. We chose thresholds that allowed each feature to perform optimally on the datasets. Figure 3 shows the average accuracy of our best-performing features on one of our easier experiments, Experiment 4 from Table I. The average was calculated by using k-fold cross validation and splitting our data sets into training and testing sets, then using the optimal threshold from training to categorize the remaining test data. We used a k value of five, and the training set size in each of the five folds was 80% of the total data set size. This meant that each data set was in the testing set exactly one time.

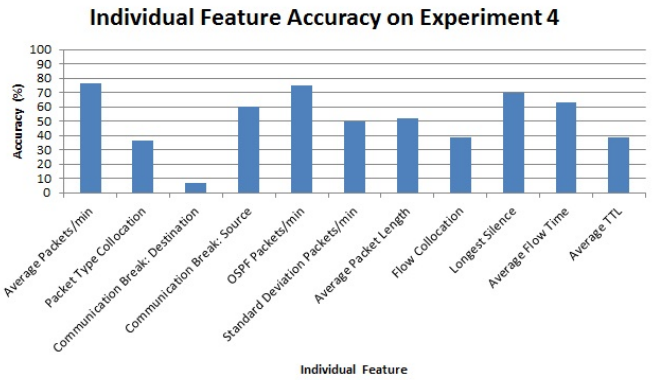

Fig. 3. Individual feature accuracy on Experiment 4

The highest average accuracy that any single feature achieved was 76.67%. Clearly, the features alone cannot produce acceptable results, justifying the need for machine learning to relate the extracted information.

E. Genetic Programming Algorithm

We developed a GP algorithm that evolves decision trees for each type of event. If multiple events need to be detected, a different tree can evolved for each event given the right features and training data. It is possible for one data buffer to contain many different events of interest, and this approach makes it possible to detect all of them individually with separate trees.

The terminals, or leaf nodes, in the GP trees are represented by the feature extraction algorithms presented in subsections III-A through III-C. The features were all normalized to be between zero and one, inclusive, to prevent any one feature type from outweighing the others due to its numerical value. These normalized values within the terminal nodes of the trees are related by fuzzy logic operators AND, OR, NOT, NOR, NAND, and XOR. The rules of fuzzy logic operators dictate that an AND of two floating point values is the minimum of the two, while OR is the maximum of the two values. A NOT is equal to one minus the normalized value. Using these rules, the other operators can be extrapolated. The reason for using fuzzy logic operators over setting a hard threshold (such as evaluating any value over 0.5 as true), is that reinforcement learning algorithms such as GP need to determine a solution's quality relative to other solutions. If every tree only returned a value of one or zero, there would only be two values upon which a solution's quality could be evaluated. By forcing floating point values to be returned by a tree, the spectrum of values between zero and one can be used by the algorithm to determine fitness. This allows for a hierarchy of solutions and makes the search space continuously multimodal.

The fitness function is key to evolving successful solutions. Our fitness function is simple. The training data is split into two sets, the PCAPs and netflows that captured the event occurring, and those that did not capture the event. Algorithm 1 shows how the fitness function uses these two sets to compute a numeric value to represent the quality of each solution.

Algorithm 1 Fitness function pseudocode

```
function get_fitness(tree)
    total ← 0
    for all data ∈ data_sets do
        val ← eval_tree(tree)
        if val ≥ accept_thresh && event then
            total ← total + num_not_event
        end if
        if val ≤ 1 − accept_thresh && not_event then
            total ← total + num_event
        end if
    end for
return total
end function
```

The purpose of adding the number of events in the training set to the total fitness if a true negative is detected (and vice versa for a true positive) is to remove any bias from unequal quantities of events and non-events.

IV. Experimental Setup

For any GP algorithm to appropriately evolve a heuristic, the right parameters must be chosen. Otherwise, the selective pressure of the algorithm could be too elitist and not allow for enough exploration of the search space, or too relaxed and never converge to a global optimum. Most security practitioners are not machine learning or GP experts, and cannot be expected to tweak many parameters. As such, we used one set of default parameters in almost every experiment, and only modified a few inputs if our results were unacceptable. Even without complex parameter tuning or the use of meta-evolutionary techniques, these parameters perform well. Table III lists the parameters used in each experiment.

TABLE III
GP parameters

Parameter	Default	Exp. 5	Exp. 6
μ	100	200	200
λ	20	40	40
max depth	4	4	5
selection	k-tourn.	k-tourn.	k-tourn.
survival	k-tourn.	k-tourn.	k-tourn.
k	7	5	5
crossover	single-point	single-point	single-point
mutation	sub-tree	sub-tree	sub-tree
mutation rate	.15	.8	.85
termination	5000 eval.	5000 eval.	5000 eval.
Acceptance Thresh.	.9	.75	.75

The use of a high acceptance threshold forces the algorithm to evolve trees that clearly separate the data into categories of events and non-events. The greater population and offspring sizes on tougher experiments encouraged exploration of the search space. To reduce the complexity of final solutions, we

introduced parsimony pressure throughout such that smaller solutions were favored when the fitness was identical.

We tested the experiments with thirty runs for each of the cross validations. In total, we completed 150 runs of 5000 evaluations for each experiment. We used 80% of the data as training data and the rest as testing data. Just as with the single feature tests, the five-fold cross validation ensured that each of our ten data sets appeared in the testing set exactly once.

V. Results

Figures 4 through 6 show the average maximum fitness values across generations during training.

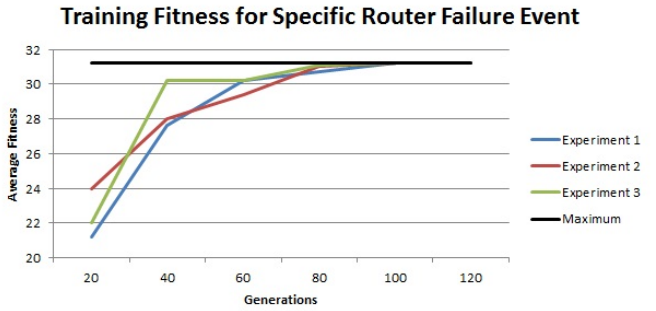

Fig. 4. Average Fitness vs Generations for the target router failure event

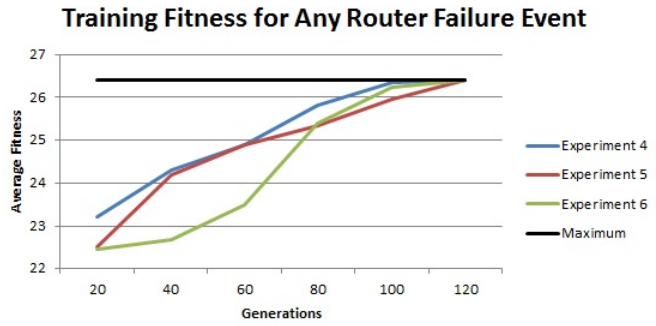

Fig. 5. Average Fitness vs Generations for any router failure event

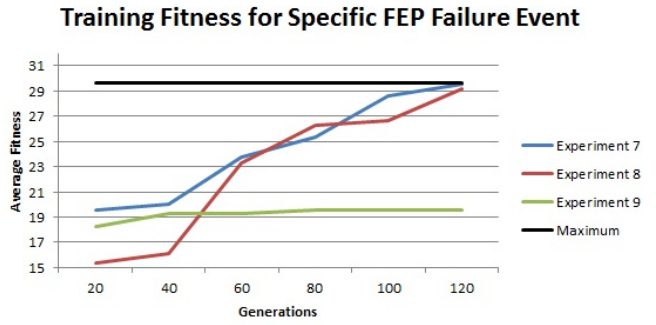

Fig. 6. Average Fitness vs Generations for the target FEP failure event

At 5000 evaluations, the populations converge to maximum values in nearly all of the experiments. Determining the failure of a FEP from a tap near a different FEP was ultimately no better than a random search, and the training fitness does not converge to the possible maximum.

When using the best solution in to identify the testing data, the average accuracy was 100% for the majority of the experiments. Experiments 5 and 6 had an average accuracy of just over 98%. Figure 7 shows a histogram of the average accuracies across all runs in all data splits.

Fig. 7. Accuracy per experiment, averaged over all runs and cross validations

Figure 8 and Figure 9 show sample overfitting plots for two data splits in experiments 5 and 6, respectively.

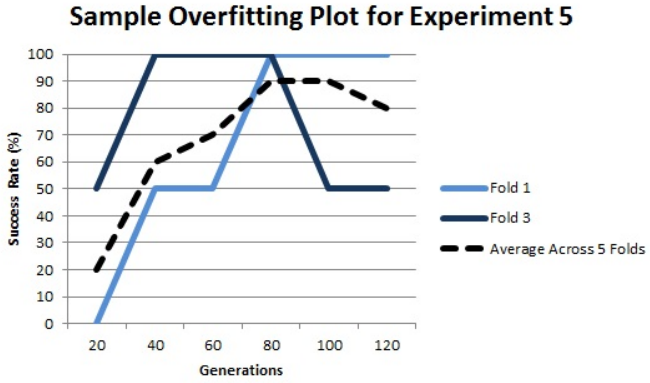

Fig. 8. Sample overfitting plot for experiment 5

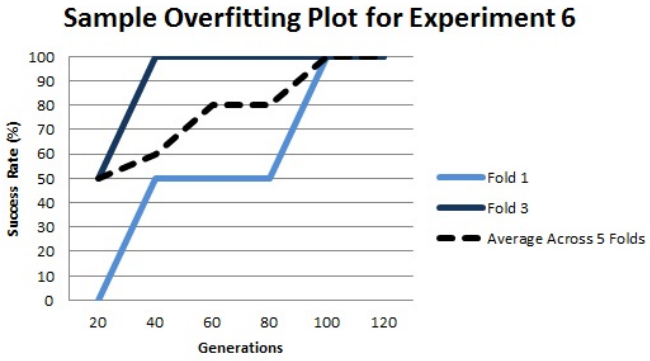

Fig. 9. Sample overfitting plot for experiment 6

In supervised learning, when parameters are not appropriately tuned, there is always a danger of overfitting. In our case, the evolutionary process may bias the results heavily in favor of the training set, such that the resultant solutions are too specialized to correctly identify new data sets.

These plots were generated by verifying the accuracy of the best solution against the testing set at fixed generation intervals. If overfitting occurs, the average accuracy drops in later generations. Figure 8 is an example of this. Note that the average accuracy shown is a representation of the accuracy in identifying the testing set only. Accuracy in identifying the training set was omitted. While overfitting is something that should be avoided, it does not have a large impact on the overall average accuracy of our solutions. The trade-off is avoiding parameter tuning for slightly worse results. We believe not performing parameter tuning to be a more realistic approach, as practitioners likely will not have enough machine learning expertise to adjust the parameters effectively.

VI. Discussion

Most of the accuracies are perfect, highlighting the importance of using machine learning to identify known events over attempting to discover new useful outliers. The recurrent nature of ICS networks was a vital contributing factor in our solution's success. The poor accuracy on experiments 8 and 9 was a result of inadequate information. The taps in those particular locations in conjunction with our feature extraction algorithms did not provide enough information to discern events from non-events. It is obvious that collecting data near the source of the event allows for greater insight into what occurred. While it is still possible to glean some information many hops away, it requires nuanced features. The types of data we collected is intended to be a worst-case scenario, with multiple devices going down, several coming back, or even a single router fluctuating between failing and operating normally. If a failure actually occurred, it would likely be easier to detect than the data that we used. The sensor trees had no problems identifying events in data sets where a lone failure occurred with no subsequent normal operations. The toughest data sets were those in which multiple routers failed because the FEPs increased their communications, and vital paths through the network were severed.

Throughout the testing process, we generated thousands of sensor trees. Figure 10 and Figure 11 show two example trees for experiment 2. The trees illustrate the variety that can be achieved while maintaining a high accuracy. The tree in Figure 10 will indicate that the router has failed if the device's longest period of silence is greater than the number of collocation anomalies, or if the standard deviation of packets increases immensely. This makes logical sense with the type of event the tree is detecting. Similarly, the tree in Figure 11 makes use of a number our features to produce a more complicated, but likely more versatile tree. The great difference in tree structure is due to the solutions belonging to different data splits. The training set used to produce Figure 11 was more difficult to optimize against.

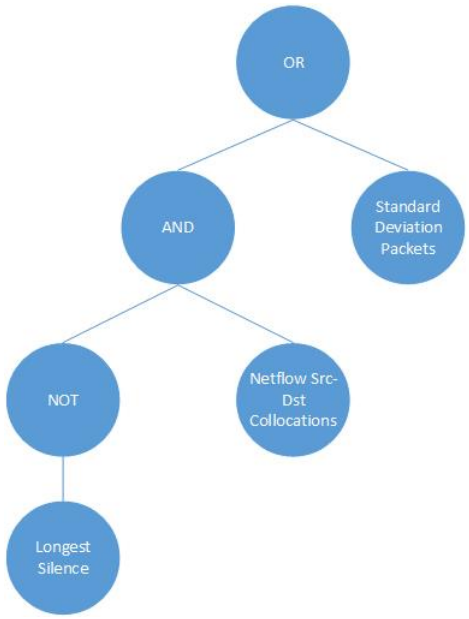

Fig. 10. Sample decision tree for experiment 2

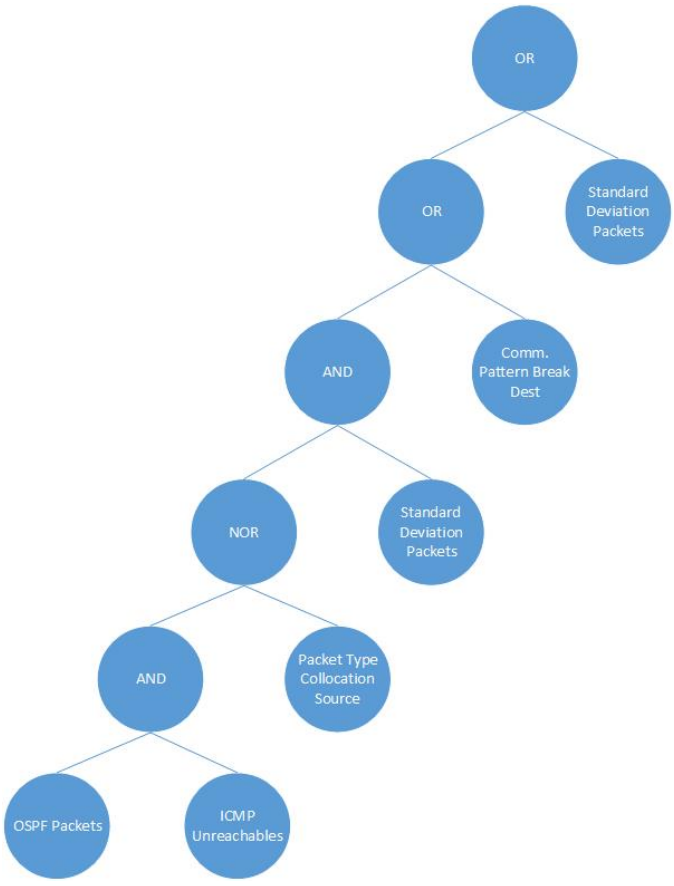

Fig. 11. Sample decision tree for experiment 2

In order to deploy this approach as a viable semi-real-time solution on a live system, it would be best to train on data capturing both the event and non-events, as we did

for our experiments. Providing both types of data prevents over-specialization to one single data type, and forces pattern recognition over outlier detection when deployed [14]. The sensor should also be placed as close to the event as possible, and a buffer size of twenty minutes should be used for optimal results. For more complete coverage and fewer false positives or false negatives, several different evolved decision trees should be placed at every location, and their consensus should be used to determine anomalies. If multiple varied trees agree that an event happened, it is more likely than if one tree returns a positive result. Because the trees are generated ahead of time, evaluation would be nearly instantaneous once the data buffer has been filled.

It is theoretically possible to detect more varied events than those explored in this paper, even if they are much more subtle. Rather than detecting strictly failures, the trees could be trained on data where devices react abnormally. If a firmware or logic update needs to be detected, this approach would likely perform very well given enough training data and the right feature selection algorithms. The results we achieved would not be nearly as positive if the features were not as diverse. The rich feature set is key to capturing the many possibilities that are associated with a network anomaly.

VII. Conclusion

Current capabilities in anomaly detection through the use of machine learning techniques have been limited. While some encouraging research has been done in the area, the fundamental goals of the previous work made the machine learning techniques not as effective as they could be. Rather than attempting to detect new attacks, our aim was to identify the occurrence of particular, known events. Specifically, we wanted to detect network failures in recurrent ICS networks. The periodicity of these networks allows for simple pattern recognition and meaningful feature extraction. We developed a GP approach to evolve decision trees for the purpose of relating the extracted features and identifying the network failures. The trees were tested in multiple locations on a large-scale virtualized network. Five of our nine experiments resulted in 100% average accuracy across a statistically significant number of runs. Two others achieved over 98% average accuracy. The experiments that had poor results suffered from a lack of available data.

Future work will attempt to build on our successes. Rather than virtualizing the network, we intend to include real ICS hardware in the loop. We will also add a diversity of ICS communication protocols. We would also like to detect more varied events, such as data exfiltration or PLC firmware updates. To succeed in these endeavors, we will need to include an updated feature set. The features used in the network failure experiments may not all be meaningful discriminators in future experiments.

References

[1] Minimega.org. http://minimega.org. Accessed: 2015-09-29.

[2] M. Bader-El-Den, R. Poli, and S. Fatima. Evolving timetabling heuristics using a grammar-based genetic programming hyper-heuristic framework. *Memetic Computing*, 1(3):205–219, Oct. 2009.

[3] E. K. Burke, M. R. Hyde, G. Kendall, G. Ochoa, E. Ozcan, and J. R. Woodward. Exploring Hyper-heuristic Methodologies with Genetic Programming. In *Computational Intelligence: Collaboration, Fusion and Emergence*, pages 177–201. Springer, Berlin-Heidelberg, Germany, Mar. 2009.

[4] E. K. Burke, M. R. Hyde, G. Kendall, and J. R. Woodward. Automatic Heuristic Generation with Genetic Programming: Evolving a Jack-of-all-Trades or a Master of One. In *Proceedings of the 9th annual conference on Genetic and Evolutionary Computation*, GECCO '07, pages 1559–1565, 2007.

[5] E. K. Burke, M. R. Hyde, G. Kendall, and J. R. Woodward. A Genetic Programming Hyper-Heuristic Approach for Evolving 2-D Strip Packing Heuristics. *IEEE Transactions on Evolutionary Computation*, 14(6):942–958, Dec. 2010.

[6] J. Hosic, D. R. Tauritz, and S. A. Mulder. Evolving Decision Trees for the Categorization of Software. In *IEEE 38th International Computer Software and Applications Conference Workshops*, COMPSACW, pages 337–442, July 2014.

[7] J. R. Koza. Overview of Genetic Programming. In *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, pages 74–78. MIT PRESS, Cambridge, MA USA, 1992.

[8] J.-F. Lin, S. Li, and Y. Cai. A new collocation extraction method combining multiple association measures. In *Proceedings of the Seventh International Conference on Machine Learning and Cybernetics*, volume 1, pages 12–17, July 2008.

[9] W. Lu and I. Traore. Detecting New Forms of Network Intrusion Using Genetic Programming. In *Computational Intelligence*, pages 475–494, Aug. 2004.

[10] L. A. Maglaras and J. Jiang. Intrusion Detection in SCADA systems using Machine Learning Techniques. In *Science and Information Conference*, SAI 2014, pages 626–631, 2014.

[11] M. Mantere, M. Sailio, and S. Noponen. Network Traffic Features for Anomaly Detection in Specific Industrial Control System Network. In *Future Internet*, pages 460–473, Sept. 2013.

[12] M. Mantere, M. Sailio, and S. Noponen. A Module for Anomaly Detection in ICS Networks. In *Proceedings of the 3rd international conference on High confidence networked systems*, HiCoNS '14, pages 49–56, Apr. 2014.

[13] M. Mantere, I. Uusitalo, M. Sailio, and S. Noponen. Challenges of Machine Learning Based Monitoring for Industrial Control System Networks. In *26th International Conference on Advanced Information Networking and Applications Workshops*, WAINA '12, pages 968–972, Mar. 2012.

[14] R. Sommer and V. Paxson. Outside the Closed World: On Using Machine Learning For Network Intrusion Detection. In *IEEE Symposium on Security and Privacy*, SP 2010, pages 305–316, Mar. 2010.

[15] J. Sun and X.-Z. Wang. An initial comparison on noise resisting between crisp and fuzzy decision trees. In *Proceedings of 2005 International Conference on Machine Learning and Cybernetics*, volume 4, pages 2545–2550, Aug. 2005.

[16] C. Y. Teo. Machine learning and knowledge building for fault diagnosis in distribution network. In *Electrical Power & Energy Systems*, pages 119–122, Apr. 1995.

[17] C.-F. Tsai, Y.-F. Hsu, C.-Y. Lin, and W.-Y. Lin. Intrusion detection by machine learning: A review. In *Expert Systems with Applications*, pages 11994–12000, 2009.

[18] C. Yin, S. Tian, H. Huang, and J. He. Applying Genetic Programming to Evolve Learned Rules for Network Anomaly Detection. In *Advances in Natural Computation*, ICNC 2005, pages 323–331, Aug. 2005.