A Wireless Intrusion Detection for the Next Generation (5G) Networks

A Master's Thesis

Presented to

Department of Network and Computer Security

In Partial Fulfillment

of the Requirements for the

Master of Science Degree

State University of New York

Polytechnic Institute

By

Richard Ferrucci, ferrucr@sunypoly.edu

Under the Supervision of

Dr. Hisham Kholidy, hisham.kholidy@sunypoly.edu

May 2020

A Wireless Intrusion Detection for the Next Generation (5G) Networks

Declaration

I declare that this project is my own work and has not been submitted in any form for another degree or diploma at any university or other institute of tertiary education. Information derived from the published and unpublished work of others has been acknowledged in the text and a list of references is given.

Richard Ferrucci

05/20/2020

Executive Summary

5G data systems are closed to delivery to the public. The question remains how security will impact the release of this cutting edge architecture. 5G data systems will be sending massive amounts of personal data due to the fact that everybody in the world is using mobile phones these days. With everyone using a 5G device, this architecture will have a huge surface area for attackers to compromise. Using machine learning techniques previously applied to 802.11 networks. We will show that improving upon these previous works, we can have a better handle on security when it comes to 5G architecture security. We find that using a machine learning classifier known as LogIT boost, combined with a selected combination of feature selection, we can provide optimal results in identifying three different classes of traffic referred to as normal, flooding, and injection traffic. We drastically decrease the time taken to perform this classification while improving the results. We simulate the Device2Device (D2D) connections involved in the 5G systems using the AWID dataset. The evaluation and validation of the classification approach are discussed in details in this thesis.

Keywords: 5G systems, Cybersecurity, Intrusion detection, Machine Learning, Device2Device, LogIT boost, Wireless Systems.

Contents

Executive Summary 2

Chapter 1: Introduction 6
802.11 Networks and Security 8
802.11 Architecture and Vulnerabilities 9
5G Topology 12
5G Vulnerabilties 14

Chapter 2: State of the Art 15

Chapter 3: Our Proposed Contribution 18

Chapter 4: Experiment Analysis and Evaluation 22
Figure 1: Feature selection using best first and cfs subset evaluator 24
Figure 2: Feature selection using ranker and ClassifierAttributeEval 26
Figure 3: Adaboost Classification 26
Figure 4: LogIT Boost classification 27
Figure 5: J48 Tree Classification 27
Figure 6: Bayes Net Classification 27
Figure 7: ADAboost with CFSsubset and Best First feature selection 28
Figure 8: Logit Boost with CFSsubset and BestFirst feature selection 28
Figure 9: J48 with CFSsubset and BestFirst feature selection 28
Figure 10: BayesNet with CFSsubset and BestFirst feature selection 29
Figure 11: AdaBoost with ClassifierAttribEval and Ranker feature selection 29
Figure 12: LogIT Boost with ClassifierAttribEval and Ranker feature selection 30
Figure 13: J48 with ClassifierAttribEval and Ranker feature selection 30
Figure 14: BayesNet with ClassifierAttribEval and Ranker feature selection 30
Classifier with no Feature Selection 31
Table 1: Results from Classifier without feature selection 31
Classifier with no feature selection 31
Table 2: Results from Classifier without feature selection 31
Classifer with CFS Subset and Bestfirst 31
Table 3: Results from Classifier with CFSsubset and BestFirst feature selection 31
Classifer with CFS Subset and Bestfirst 31
Table 4: Results from Classifier with CFSsubset and BestFirst feature selection 32
Classifier with classifierattribeval and ranker 32
Table 5: Results from Classifier with ClassifierAttribEval and Ranker Feature Selection 32
Classifier with ClassifierAttribEval and Ranker 35
Table 6: Results from Classifier with ClassifierAttribEval and Ranker Feature Selection 32
Chapter 5: Conclusion 32

References 32

Chapter 1: Introduction

The emergence of 5G data is upon us now, and it is more important now more than ever to start turning our heads to the security landscape surrounding 5G data. Surprisingly, there has not been much work put into security risks and avoidance to this date. There are many problems surrounding the research into 5G data. The lack of proper datasets and test beds are acting as signiact blockers into the potential research into 5G systems. This being the case, we must draw parallels between existing wireless technologies to 5G. The reason for this is that there are numerous publicly available datasets for wireless technologies. Also, wireless network testbeds are easily accessible, in fact, most people today have wireless technology set up in their home, even businesses are moving to a wireless backbone infrastructure. This being the case, it will be easier to get meaningful results out of the numerous techniques we will apply to the datasets. We will be using the AWID dataset for our experiments in this research. The AWID dataset was built for this exact type of work. The goal of the AWID dataset was to act as an essential building block for researchers to apply different types of machine learning techniques to this large dataset in an effort to strengthen our posture for intrusion detection when related to wireless technologies. The AWID dataset consists of a number of different CSV files that are converted from PCAP files in wireshark. They created full and reduced datasets to ease the burden of computing power necessary to provide significant research. Each version has a training and testing csv file. This allows for the training and testing data to be directly correlated which allows for much more meaningful results. Each csv file has about 300000 lines of data. Each line of data represents a wireless packet that transmitted over the line. Each line has around 158 attributes which allow for us to take a deep dive into which of these attributes are most meaningful when performing machine learning techniques and for the overall goal of finding what attributes are crucial for intrusion detection. For this research we will use the reduced training. The reduced training set will have less records which means less records of attacks but in order to run the techniques against the full datasets you need extreme computing power which is not easily attainable. Even using the reduced dataset we will get results comparable to that of the full datasets.

802.11 Networks and Security

Now we will look at structure and transmission of data in both 802.11 wireless networks and 5G networks in order to draw parallels between the two so that we can see that the AWID dataset may still apply to certain aspects of 5G networks. Since there is no readily available dataset for 5G research, this is the way that we can apply intrusion detection techniques to 5G networks even though we do not have the dataset or test bed available at the time of writing. First we will introduce 802.11 networks. These are wireless networks that you are accustomed to seeing. You see these in businesses and homes. Your ISP provides you with a coaxial cable and a modem. The coaxial cable plugs into the modem which allows the modem to translate analog waves into digital. Most modems are then plugged into a router which can broadcast data in the form of Wi-Fi. Now, the router must be wi-fi enabled which means it has to be able to broadcast a certain type of packet so that wireless devices can see the presence of a wireless network, otherwise the network will not be broadcasted making it difficult for wireless devices to join the network. Before we get into frames and packets, we will talk about security because after all, that is the most important part of any network. 802.11 networks introduce three types of security techniques. These techniques are WEP which stands for wired equivalent protection, WPA, which stands for WiFI protected access, and finally WPA2, which stands for WiFi protected access version 2. WEP was the first security protocol and it had glaring issues which made it quite vulnerable to availability attacks. An attack against availability means that an attacker tries to take the network down so that is no longer available for the clients to use. The WEP protocol was also vulnerable due to the general use of initialization vectors. WEP used shared keys through the use of initialization the vectors. The issue is that the initialization vectors were static, cleartext, and far too small. WEP also had no cryptographic protection. This means that the integrity of packets were unable to be checked. This poses a serious issue of not knowing if the packet or frame you received actually belongs to the person that you expect it to belong too. WEP was proven to be very insecure so a new protocol was developed known as WPA. WPA added MIC and TKIP. MIC stands for message integrity checks which allowed for the receiver of data to be sure that the data was untouched and was from

the person that was desired. TKIP stands for temporal key integrity protocol. This fixed the issues with static keys. The keys were temporary and were on a per packet basis. WPA was found to be ineffective due to the fact that people using WEP could easily change from WEP to WPA using a firmware update. Even though the firmware was updated, the WEP backbone was still in use which made WPA also ineffective. Finally WPA2 was developed and is still in use today. WPA2 brought AES and CCMP into the fold. CCMP stands for counter mode cipher block chaining message authentication code protocol. The big improvement here was that it brought 128 bit keys to the table which would make it inefficient for an attacker to try to break the key.

802.11 Architecture and Vulnerabilities

Now that we have examined the security protocols with WiFi, we move to the more important information that correlates to the research in this paper. The architecture and attack types used against this architecture will be more focused on throughout this paper due to the sole fact that this is where I will attempt to draw parallels to 5G data networks and prove that existing datasets may suffice for preliminary testing when it comes to security and intrusion detection. It is important to remember that although 5G networks are different from wireless networks, there are parallels to be drawn so that researchers may use these existing data types to further the researchers progress. The architecture in a nutshell will look like numerous access points interconnected to service many stations. The stations in a wireless world refer to the clients attempting to connect to the access points. Data switches are used to connect access points all over a specific building or wireless networks. This allows data to flow between access points seamlessly and allows clients to move positions without seeing a drop in coverage. Similarly, when you use a phone, you are connected to wireless towers that cover a specific area. When you leave one area, and move to the next, you will not see a drop in coverage because the towers will hand off the association to the next so that you do not notice a change in performance even though you are associating to a different tower.

Frames are the way that data is transmitted through a wireless network. There are 3 different frame types when we are considering a wireless network. The three different frames are management, control,

and data. Each frame serves a different purpose in maintaining a wireless network. A management frame is focused on establishing connection between clients and access points. This can be compared to when you are on a laptop and you need to connect to a wireless network. You select the network you would like to join, you are prompted for a password, then you are either granted access or denied access based on the integrity of the password you input. This looks quite simple to the user but there is a ton going on behind the scenes. Management frames have many sub categories. A few of these categories include authentication, deauthentication, beacons, and association requests/ responses. Authentication and deauthentication frames handle inputting the correct password and the access point validating that password as either correct or incorrect. Association requests and responses handle the client actually being associated or connected to a certain access point. This is what would happen if you are moving through different coverage zones. You are associated with one access point, you reach the bounds of that access point and you are disassociated with that specific access point and you are reassociated with an access point that is better fit for your geographic location. The same applies to data networks such as 4G and 5G. Finally, beacon frames are what allows your laptop to see that there is actually a network that is available to connect to. The access point itself will broadcast beacon frames out, your laptop will notice these frames and will then include this wireless network as an available option to connect to. Next we discuss control frames. Control frames are put in place to ensure delivery of data between the client and the access point. They are in charge of the handshake which consists of a request to send, clear to send, then acknowledgement of receival. The final type of frame is the data frame. The data frame is simply the frame in charge of gathering the data that needs to be sent. It is collected from the different layers of the stack.

To wrap up our discussion about wireless networks we will introduce attacks that wireless networks are vulnerable to. Since wireless networks are vulnerable to many types of attacks due to the fact that it has been around for a long time which have made it easier for attacks to be developed and tested. Since there are so many, we will focus on attacks against availability. This will also help us draw the parallel between WiFI and 5G because as of now, the most prevalent attacks against 5G networks is attacks on its availability. We will also include attacks on integrity. Like previously mentioned, there are so many developed attacks

against WiFi that it would be impractical to list them all so we will list a select few to narrow our focus and get true results. First we will talk about a DoS attack. DoS stands for denial of service, which also means this is an attack against availability. A denial of service attack can be applied by using a brute force approach. A brute force approach would require the use of multiple different computers all targeting one victim. This one victim will be overwhelmed with the amount of traffic being sent its way rendering it useless. If the attack is performed against the access point, this can take down the whole network. The router or access point will be flooded with packets, not allowing it to route packets properly and eventually causing the CPU to be overwhelmed which would end up taking the whole system down. This would require a lot of different computers because routers nowadays can handle a tremendous amount of packets. WPA was also vulnerable to a certain DoS attack that could shut the system down without overloading it with traffic. The way it was performed is that an attacker would send unauthorized data to the access point and this would cause the access point to shut down because it would assume it was under attack and this was the way of dealing with it. The next attack we will look at is known as a man in the middle attack. This attack was directed at disturbing integrity. This can be done in a number of ways. The most prominent or well known attack is called the Evil Twin attack. This attack would require an attacker sending a number of disassociation packets to the router, this would essentially shut down the router, the attacker would broadcast its own wifi network with the same name as yours, so your computer would connect to the fake network and allow the attacker to intercept all data being sent because all of the data you would be sending would go to the attackers nic card, then the attackers nic card would send it to the open web. This would not appear fraudulent to an unknowing user because you were still connected to the wifi name that you selected and you would still be able to access the open web. The only issue is that there was a man in the middle capturing all of your traffic and passwords before sending it to the proper channels. These attacks are also prevalent in 5G networks which will be proven in the upcoming sections when we introduce the topology, architecture, and vulnerabilities of 5G networks.

5G Topology

Now that we have covered 802.11 wireless networks enough for reader comprehension we will introduce 5G networks. 5G networks are new to just about everyone that is not involved in the development of the infrastructure or not following the technology industry closely. Now, obviously 5G is becoming a buzz word in the industry but do many people know what it actually consists of or how it is going to be implemented in today's day and age. This section of my research will be dedicated to informing the reader on how the topology will look, and what it will actually consist of so that we can draw parallels to existing WifI networks.

First we will look at the components and topology of a 5G network. A 5G network consists of two major parts. The parts are called the radio access network and the core network. The radio access network is the first aspect of the 5G architecture we will examine. A way a phone connects to a radio access network can be compared to the way that laptops or internet connected devices connect to a access point. The radio access network allows a phone to connect to the network using radio connections. This acts as a middle man to the core network. The same way the access point acts as a middle man to the world wide web. In this case, the phone connects to the radio access network, the radio access network communicates with the core network to transmit data to the open web or wherever its destination may be. A radio access network contains a base station and antennas. Each of these components correspond to a certain geographical region of cover. You may be familiar with radio access networks of the past. Some of these consist of UMTS, LTE (long term evolution), and GSM. The general idea of a RAN base station is to take packets of data, and convert them to radio waves so that they can be transmitted. The antennas serve a general purpose that most would be aware of. The antenna is used to send and receive transmissions from other base stations. There are many different types of antennas and they each have a case where they are most useful. For example, an omnidirectional antenna may be used for shorter distances. The reason for this is, is in the name. It emits signals in all directions which means its range is limited but its radius is not. A parabolic is unidirectional meaning it sends and receives in one direction. The distance here is greatly amplified but if the antenna is

not pointing at the preferred target, the signal may be very small. Next, we talk about the radio access network controller. The controller has two main responsibilities. The first being the host to the nodes that are connected to the specific radio access network. The second being connecting the two different types of core networks. Those networks being the packet switched and the circuit switched. The packet switched looks like a network switch. The idea of both are extremely similar and it is easier to think of it as one in the same. The main thing to remember about the radio access network is that it is what your device connects to. The access network connects to the core network. Carrying almost the same responsibilities as an access point. We think of it as an access point in this case because the whole idea of this work is to draw parallels between a 5G network and wireless networks of today's day and age.

Now let's move onto the next aspect of the 5G architecture. The core network rounds out the architecture and may be the most vital part in the grand scheme of things. The core network handles all voice, data, and internet data the connected devices send and receive. 5G networks are making more use of cloud services to better handle data and interconnectivity. The claim is that they will add multiple servers to better handle latency issues that 4G data systems currently face. 5G is implementing a technique called network slicing. What this really looks like is a local server to a particular sector designated for that sector. This local server will handle all data from a particulare sector and communicate it back to the central server. This is useful for emergency services and businesses that may handle critical data. Let's say a hospital is designated as a sector. That hospital will have its own designated server dedicated to its and only its data transmissions. This allows for much lower latency especially in times of crisis. Another technique 5G will use is known as network function virtualization. This service allows for virtualization of critical network hardware. For instance, a firewall no longer needs dedicated hardware. Using network function virtualization we can spin up a firewall as software on a virtual machine and implement it on the network on the fly. There are many applications for this functionality. Any network hardware available as software can now be cloud based which limits hardware vulnerabilities and hardware costs. The cloud servers already exist so there will not be need for any unnecessary hardware.

5G Vulnerabilities

Now that we have an idea of what 5G actually is, we must talk about security implications. After all this is one of the most important things to consider when developing new techniques or methodologies. Cyber criminals work day and night to find new vulnerabilities in emerging technology. 5G is recognized as being more secure then its predecessor. Surprisingly enough, 5G faces some of the same vulnerabilities as wireless networks. Although this actually may not be that surprising after looking at all of the similarities between the two. For this research we will highlight the attacks that are most similar to that of a wireless network. This will allow us to apply our dataset in a way to get meaningful results. We will look at four different vulnerabilities. These vulnerabilities are replay attacks, man in the middle attacks, denial of service, and brute forcing.

First we will look at the replay attack. This is important because this is a quick way to cause financial damage to someone. This works by running up this cell phone plan by either depleting data or minutes. This works by an attacker eavesdropping on data communication. They will then repeat the message to the device. This is done by stealing the key. Since the attacker has the key, the attacker then can repeat whatever messages they want over and over again. This will result in high data usage and it will not be aware to the user that this is happening.

Next we look at the man in the middle attacks. There are already tools developed to aid attackers in this process. One of these tools is known as MNmap. An attacker can set up a fake base station, this allows the attacker to gain required knowledge such as device type, operating system, and what it is being used for. After this is done, the attacker can use a targeted attack against the device since they know exactly what the device is and what is being used for. Another way to perform this type of attack is to attack the radio access network that a device is connected to. The RAN controls what data speeds the device is permitted to use. By doing this, an attacker can alter the RAN by making it think a different type of device is connected that is allowed less data speed. This means that an attacker can dramatically slow down the speed of a connected device to make it almost unusable. This would also qualify as a denial of service attack if the attacker varies the device so much that it won't allow any data speed whatsoever.

Chapter 2: State of the Art

For previous works we must look a little deeper than just 5G work. The issue here is that since there is no dataset for this type of work, so finding work regarding machine learing is difficullt. For this we will look at a paper titled “Intrusion Detection in 802.11 Networks: Empirical Evaluation of Threats and a Public Dataset”. This is an okay scenario due to the fact that we have spent the better part of this research focusing on the similarities or parallels between 5G systems and 802.11 networks. The paper we will refer to was written by Constantinos Kolias, Georgios Kambourakis, Angelos Stavrou, and Stefanos Gritzalis. We will collectively group these people when referring to their work. Their work focused on 802.11 networks with no referral to 5G networks. Therefore, they were able to put a better focus on the dataset since it was dedicated to 802.11 networks. Their classification was broken into 3 distinct types of attacks. These categories included flooding, injection, and impersonation. Flooding attacks typically produced results of increased traffic, or a temporary increase in management frames which is explained above. Injection attacks would typically produce smaller sized packets that had the appearance of being legitimate. This type of attack would include fragmented packets so that the router would consistently look for the other pieces of the packet when there really isn't anything else to be found. This is compared to a replay attack in a 5G world. This would run customers out of data or increase bills without any knowledge of the client. The final category was classified as impersonation attacks. In an 802.11 network this would look like a fake access point being introduced to the network. Clients would unknowingly connect to the unauthorized access point and that access point would be able to record data being sent and the attacker would gain access to this information. This is the same in a 5G network where a fake station or radio access network could be set up so that devices would unknowingly connect to it. There was also a category labeled normal. This is just traffic that is benign to the network and is normal for day to day activity. Their paper spends a great deal of time explaining the different types of attacks that would fall under these categories. It also spends a great deal of time explaining the architecture of an 802.11 network. The main focus for this research are the

techniques they used to classify data and to really tell what is malicious and what is not. There is also work done by Rongpeng Li, regarding software defined architectures for centralized threat management. Li uses an “intelligent intrusion system”, to detect intrusions. Li states that “it flexibly combines security function modules which are adaptively invoked under centralized management”. This work uses machine learning algorithms to essentially detect zero day threats, or threats that are not yet containing known signatures.

Now, we will look at the author's setup and the techniques they used to gather meaningful results. The authors' setup included 1 desktop, 2 laptops, 2 smartphones, 1, tablet, and 1 television. This is a typically home or small business network setup. The simulated attacker was a laptop running the Kali OS using a seperate nic card to inject packets and to monitor traffic. There was also a seperate laptop running in monitor mode to observe the traffic being sent in a non biased way. The dataset they used was the AWID dataset which we are also using in this work. The AWID dataset is explained above but more specifically the authors utilized the reduced training and test set. The reduced datasets contain 1,795,575 records and 1,633,190 is normal traffic while the remaining records are malicious. The records are simply packets captured over an hour of network traffic. The normal traffic took up about 45 minutes of the full hour that network data was captured. The other 15 minutes of traffic was the malicious data being sent over the network. 54.5 percent of the malicious traffic consisted of injection attacks, 18.5 percent was dedicated to flooding attacks, and 26.8 percent was dedicated to impersonation attacks. Each packet or record included 156 attributes. Some of these attributes included source address, destination address, packet number, initialization vector, and much much more. As I found out, the values are in string and nominal values. In order to run machine learning techniques against this, the authors chose to filter the attributes and convert them all to a nominal type attribute. In the authors research they ran the J48 algorithm, random forest algorithm, and OneR algorithm. The J48 algorithm produced the best results while Random Forest and OneR had the second best results in terms of true positive and false positive rates. The authors go on to speak about attribute selection and how it will cause better results. This is true because some of the attributes in a packet may not help the proper identification of malicious packets. Most of the attributes just add computational overhead. The authors use manual attribute selection which may work in several cases. For

our research we will use Weka for feature selection then apple those features to see if we can gain faster identification while also achieving the same or better results. Other state of the art papers were published by kholidy et. al that study the intrusion detection in different domains such as cloud computing and SCADA system security. These papers use different machine learning approaches, we adjust these approaches in the 5G network domain in this thesis.

Chapter 3: Our Proposed Contribution

Since we now know that wireless intrusion techniques are applicable to 5G data systems we can now apply improved techniques to wireless datasets to show that we can improve upon previous results and detect intrusions in a more time efficient and accurate manner. We will do this by applying different machine learning classifiers while also automating the feature selection process. The feature selection process is important because the AWID dataset is so large and has so many different attributes that the time taken to parse this data can be long and some of the data will actually act as a blocker because it does not contribute to the end goal. The data must be pertinent to our end goal to actually be useful. For our work, we will use weka 3.4 which is an open source data mining tool which allows us to use open source machine learning techniques. We will be using this on a kali based HP laptop that has 64 gigabytes of ram and an updated intel i7 processor. We will not be using a graphics card to accelerate the process. This will allow many more researchers that do not have access to upgraded equipment to repeat my results. After all, the goal is to allow researchers to use this as a stepping stone to create and innovate even further than the last. Before we begin, the dataset is raw and in a csv format. we must convert this CSV format into a .arff format due to the fact that weka deals with .arff formatted filed and not .csv formatted files. After we have our data in the correct format, we must truncate the data. This is over 1 million instances which weka cannot handle without dedicated equipment. For this research we will cut the data in half and look at about 600000 instances to simplify the stress on our hardware and the software itself. Once we have that, the data has different data types. We must make this uniform because the classifiers will not be useful if there are nominal, string, and other types of data. For this we will use the built in filters that weka has to offer. First we will convert all data to the string type data. Then we must get rid of values that are labeled as a question mark. This type of data is a blocker in weka. If an attribute has a question mark or no value will we dispose of the data to streamline the process, after all data with no value is useless to us. Now that we have a properly formatted dataset, we can move on to classifying our data.We will use four different classifiers in

our work. We will run the classifiers without feature selection to get a baseline result. After we achieve baseline results we will run two different feature selection techniques. This will allow us to see if accuracy improves and time taken to run the classifiers decreases. The goal is to be able to do this as time efficient as possible while not sacrificing accuracy in the long run. The four different classifiers we will use are known as Adaboost, Logit Boost, J48, and BayesNet. A classifier is a way to implement a technique in machine learning. In a nutshell, a classifier is a model that allows for learning off of a specific training set when compared to itself or a dedicated test set. In our case we have a training set and a test set which allows for improved accuracy in our results. ADAboost is short for adaptive boosting. The reason it is called this is because it takes a subset of many weak attributes and groups them into one strong attribute. This works really well in our case because we have 155 attributes and as I mentioned before, some of these attributes are not great for classification, this method will allow us to group these so called weak attributes into strong attributes so we do make waste of any of these attributes. ADAboost also puts a stronger weight on test cases that are harder to identify. This is an efficient method for our use case. Next, we have Logit Boost. Logit Boost follows the same process as ADAboost. It groups weak classifiers into what's called a decision stump. A decision stump is just a group of weak classifiers grouped into one strong classifier. J48 is a decision tree technique. This is also a great technique for our use case because it groups into categories. In our case we are attempting to group instances into one of 3 categories. Those categories are normal traffic, flooding traffic, and injection traffic. This method used nodes and internodes to classify instances into categories. Finally we have BayesNet. BayesNet is short for Bayesian network or belief network. This network is great for our use case because it looks at events and then tries to correlate it to a specific cause. We have 155 attributes so the network will see that an injection instance is happening then relate it to the attribute or attributes that are responsible for this instance. In a nutshell, it will look at the dataset and see that a certain number of attributes is responsible for a specific event. For example, attribute 1, 5, and 10 are correlated to an injection event. For the feature selection we will use two different combinations of attribute evaluators and search methods. An attribute evaluator measures the worth of an attribute against the desired outcome or class. The search method is the aspect of combining a number of different attributes to select

the features that will be more useful in our classification. Our two different combinations will be cfssubseteval as our attribute evaluator and bestfirst as our search method. The second combination will be classifierattributeval as our attribute evaluator and ranker as our search method. According to Weka, "Cfssubseteval evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them." Weka also adds "Subsets of features that are highly correlated with the class while having low intercorrelation are preferred." ClassiferAttributeEval, "evaluates the worth of an attribute by using a user-specified classifier.", according to Weka. These two attribute evaluators both use different methods of grouping attributes in a manner that will allow the search method to create features that will be best used when classifying. BestFirst, looks at the attributes in two different ways. It can either start empty, then add attributes for feature selection, or start a full amount of attributes, then take attributes away as it sees fit. Ranker is dependent on the type of attribute evaluator you decide to use. Ranker chooses attributes based on a one by one inspection of the evaluation delivered from the attribute evaluator. In the following section we will review the results of the classifiers on their own merit, then look at the results of the classifier used alongside the given feature selection technique. In our case, we will use feature selection and classifying in the same step. You can also use feature selection to find the attributes you want to use then create a separate dataset and run different classifiers against that dataset. This will show drastic time differences if done properly. In previous works, the researchers have done feature selection manually, but that may not be feasible for future work.

Chapter 4: Experiment Analysis and Evaluation

Now we can finally move into the results phase. Here we will go over the results of running the classifiers with and without feature selection. Before we do that, we will explain what some of the results mean and whether high or low values are desired in our use case. Accoridng to Anna KASPERCZUK, "TP Rate is the rate of true positives (instances correctly classified as a given class). FP Rate is the rate of false positives (instances falsely classified as a given class);. Precision is the proportion of instances that are truly of a class divided by the total instances classified as that clases. Recall is the proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate). F-Measure is the general indicator of quality of the model. ROC Curve (ROC Area) is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The accuracy of the test depends on how well the test separates the group being tested into those with and without the disease in question. Accuracy is measured by the area under the ROC curve. Kappa Statistic is a measure of conformity between the proposed allocation instance of the class and the actual, which is about the overall accuracy of the model." For TP rate, we want a high value, for FP rate, we want a low rate. After all we want instances classified correctly which is the meaning of TP rate. For F Measure, Recall and precision, we are looking for values closest to 1 or 100%. For the ROC area, we are also looking for values closest to 1 or 100%. Now that we know what the values mean, and what values we are looking for, we can start to evaluate the results we achieved. First we look it Logit boost, Logit boost achieved its best results when combined with the classifierattrib eval and ranker feature selection. We achieved high results in all categories. We achieved 1.000 recall which is a perfect score for that category. We got .982 in a true positive rate and .42 in a false positive rate which are close to optimal results. Using feature selection, we achieved better results that normal, logit boost achieved good results without feature selection however. The time taken to achieve these results were also optimal. For instance we can look at logit boost ran with no feature selection vs logit boost ran with both feature selection techniques. Without feature selection logit boost takes 5865 seconds to classify. Logit boost ran with cfssubset and best first only takes 4088 seconds. Logit

boost ran with classifierattribeval and ranker only takes 3726 seconds. We can see that running the classifier accompined with feature selection we gain significant time improvements. We alsio ran the classifier and feature selection at the same time so this has an affect on the time results. The time results would have been lower but we did feature selection combined with classification so we can subtract about 800 seconds from that which is the time it took to perform the feature selection. The tables below will illustrate that each classifier gains significant time improvement when ran with feature selection. Table 2, 4, and 6 will be the tables you will want to examine to see the exact time differences when comparing classifiers with and without feature selection. For our results, we show based on all of the figures below, that logit boost is the ideal classifier to use when looking at 5G intrusion detection along with wireless intrusion detection. Next we look at J48, J48 performed well in every category. J48 performed extremely well when used in combination with CFSsubset and BestFirst feature selection. The TP rate was as high as .986. It also came up with .286 for a FP rate which is incredible. Its recall, precision, and f-measure were not as great as logit boost however. Adaboost and bayesnet performed well in certain categories but they did not perform well enough to compete with logit boost and J48. We show that logit boosts strong results in all categories will be the optimal classifier to use when looking at intrusion detection in 5G networks and 802.11 networks. We see that bestfirst combined with CFSsubset and classifierattribeval combined with ranker give us promising results to the fact that these feature selection techniques work best when there are many attributes to search through. The reason for this is that these techniques attempt to group attributes instead of looking at them one by one. This allows for weak attributes to be grouped and either used as one strong attribute or discarded depending on the feature selection technique we are using.

```
              [list of attributes omitted]
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: no attributes
        Search direction: forward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 1277
        Merit of best subset found:    0.284

Attribute Subset Evaluator (supervised, Class (nominal): 146 class):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 46,66,68,77,145 : 5
                     radiotap.datarate
                     wlan.fc.subtype
                     wlan.fc.frag
                     wlan.ta
                     data.len
```

Figure 1: Feature selection using best first and cfs subset evaluator

```
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 146 class):
        Classifier feature evaluator

        Using   Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.rules.ZeroR
        Scheme options:
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5


Ranked attributes:
  0   145 data.len
  0    49 radiotap.channel.type.cck
  0    47 radiotap.channel.freq
  0    48 radiotap.channel.type.turbo
  0    50 radiotap.channel.type.ofdm
  0    72 wlan.fc.protected
  0    51 radiotap.channel.type.2ghz
  0    52 radiotap.channel.type.5ghz
  0    46 radiotap.datarate
  0    45 radiotap.flags.shortgi
  0    44 radiotap.flags.badfcs
  0    43 radiotap.flags.datapad
  0    38 radiotap.flags.cfp
  0    39 radiotap.flags.preamble
```

Figure 2: Feature selection using ranker and ClassifierAttributeEval

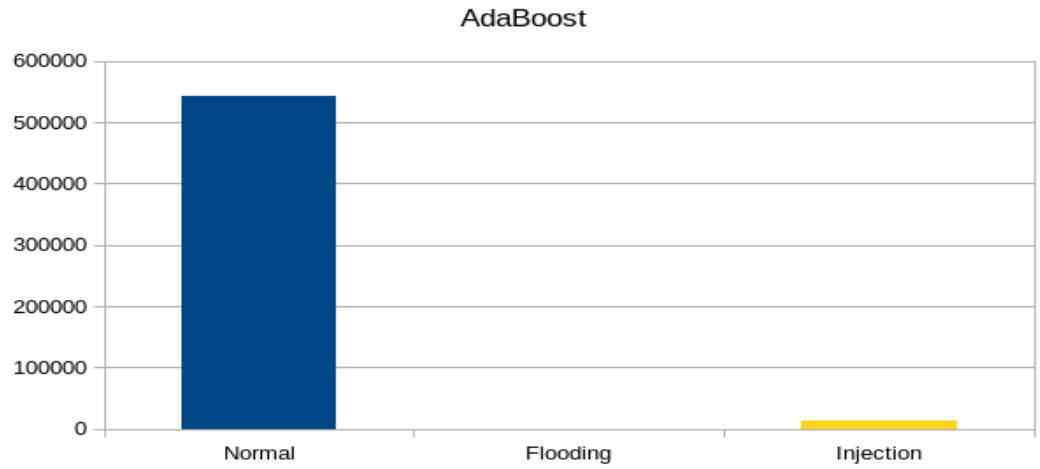

Figure 3: Adaboost Classification

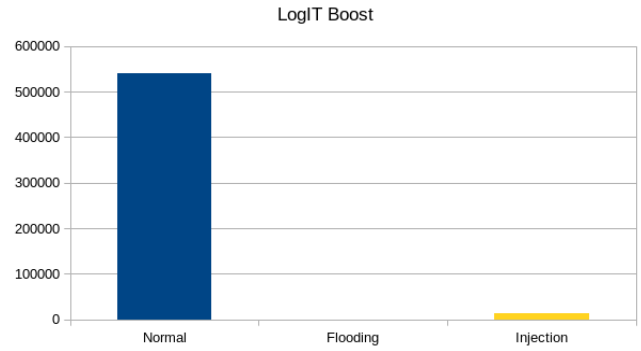

FIGURE 4: LOGIT BOOST CLASSIFICATION

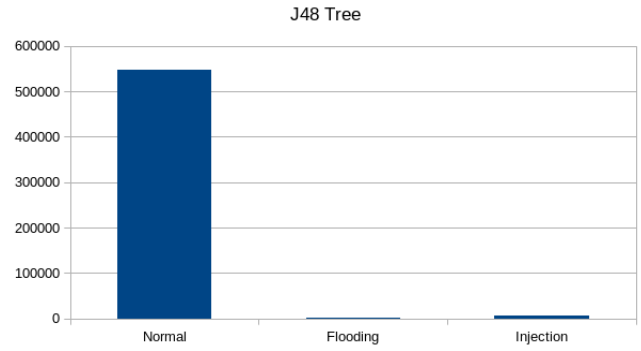

Figure 5: J48 Tree Classification

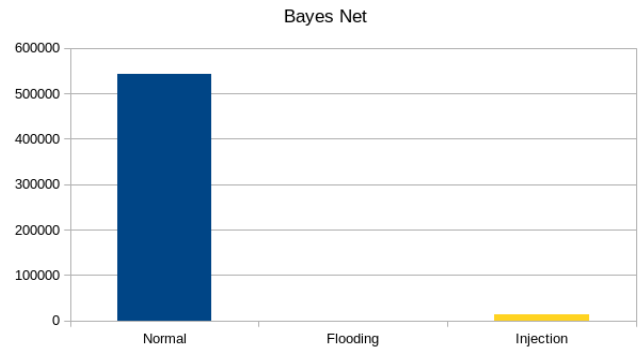

Figure 6: Bayes Net Classification

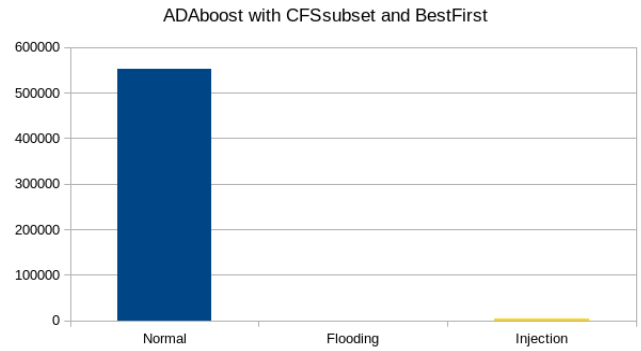

Figure 7: ADAboost with CFSsubset and Best First feature selection

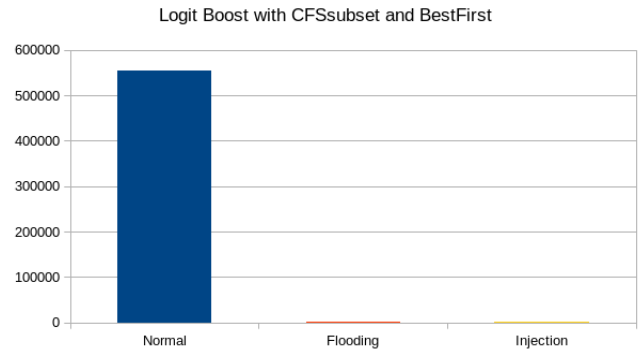

Figure 8: Logit Boost with CFSsubset and BestFirst feature selection

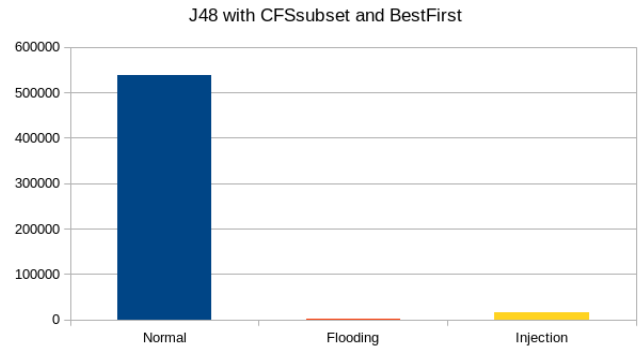

Figure 9: J48 with CFSsubset and BestFirst feature selection

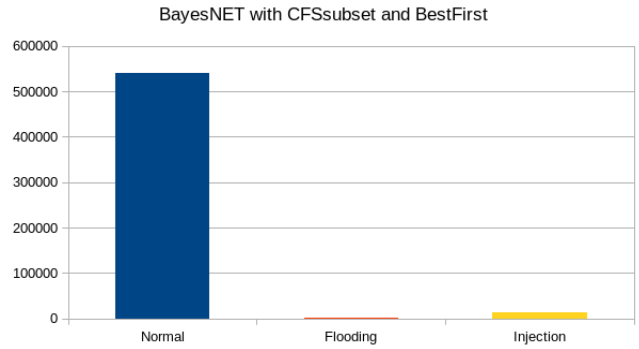

Figure 10: BayesNet with CFSsubset and BestFirst feature selection

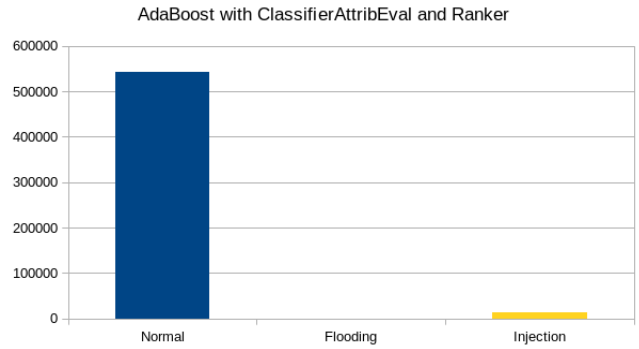

Figure 11: AdaBoost with ClassifierAttribEval and Ranker feature selection

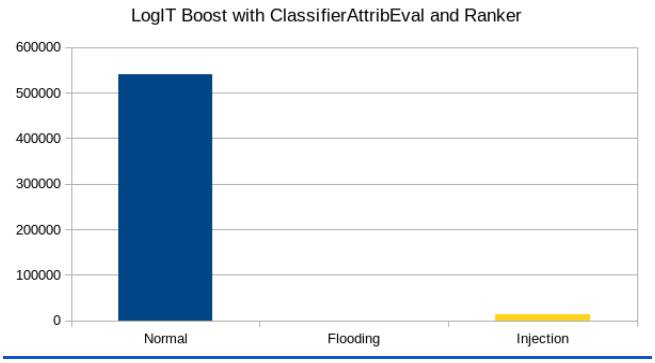

Figure 12: LogIT Boost with ClassifierAttribEval and Ranker feature selection

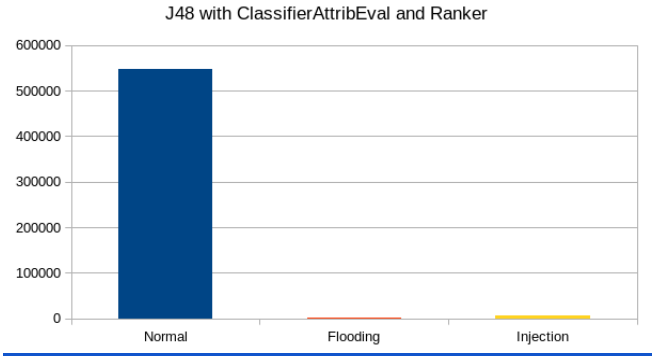

Figure 13: J48 with ClassifierAttribEval and Ranker feature selection

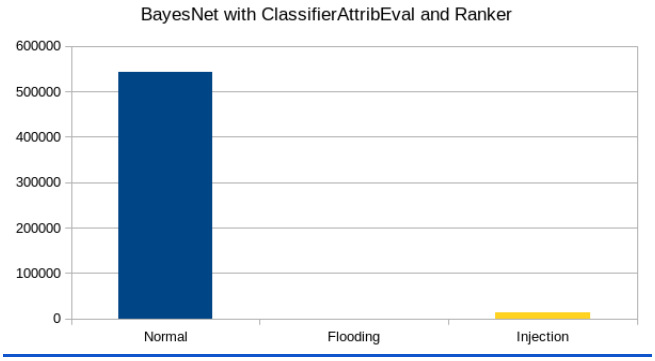

Figure 14: BayesNet with ClassifierAttribEval and Ranker feature selection

Table 1: Results from Classifier without feature selection

	TP Rate	FP Rate	Precision	F-Measure	Roc Area	PRC Area	Recall
Logit	0.981	0.403	0.981	0.974	0.979	0.983	0.981
J48	0.945	0.909	0.924	0.932	0.373	0.863	0.945
ADAboost	0.980	0.417	0.998	0.910	0.937	0.874	0.980
BayesNET	0.981	0.414	0.981	0.974	0.839	0.965	0.981

Table 2: Results from Classifier without feature selection

	Time (sec)	Correct Classification	Incorrect classification	Root MEan squared error	Relative Absolute Error	Root Relative squared error	Kappa Statisitic
Logit	5865.71	545089	10475	.1121	32.87%	66.13%	.7249
J48	3165.07	525161	30403	.1901	66.48%	112.19%	.0503
ADABoost	3960.48	544705	10859	.1257	38.58%	74.17%	.7121
BayesNEt	3319.93	544833	10731	.1134	22.87%	66.93%	.716

Table 3: Results from Classifier with CFSsubset and BestFirst feature selection

	TP Rate	FP Rate	PRecision	F-Measure	ROC Area	PRC Area	Recall
Logit	.956	.932	.937	.936	.922	.957	.956
J48	.986	.294	.986	.980	.514	.902	.986
ADAboost	.952	.925	.957	.976	.819	.819	.952
Baynet	.980	.405	.980	.974	.972	.985	.980

Table 4: Results from Classifier with CFSsubset and BestFirst feature selection

	Time (sec)	Correct Classification	Incorrect Classification	Root MEan Sqaured Error	Relative Absolute Error	Root Relative Sqaured Error	Kappa Statistic
LOGit	4088.62	531051	24513	.162	67.87%	96.085%	.0455
J48	4107.94	547845	7719	.096	26.92%	56.66%	.81
ADAboost	3729.24	528721	26843	.174	66.35%	103.06%	.0209
BAyesNET	4268.8	544631	10933	.110	27.27%	64.24%	.7154

Table 5: Results from Classifier with ClassifierAttribEval and Ranker Feature Selection

	TP RATE	FP RATE	PRESCISION	F-MEASURE	ROC AREA	PRC AREA	RECALL
LOGIT	.982	.422	.981	.990	.980	.998	1.000
J48	.945	.909	.924	.932	.373	.863	.945
ADABOOST	.980	.417	.984	.950	.937	.974	.980
BAYESNET	.981	.414	.981	.974	.839	.965	.981

Table 6: Results from Classifier with ClassifierAttribEval and Ranker Feature Selection

	TIME (SEC)	Correct Classification	Incorrect Classification	Root Mean Squared Error	Relative Absolute Error	Root Relative Squared Error	Kappa Statisitc
LOGIT	3726.09	545089	10475	.112	32.87%	66.13%	.7249
J48	3313.14	525161	30403	.1901	66.482%	112.197%	.0563
ADABOOST	3440.57	544705	10859	.1257	38.58%	74.16%	.7121
BayesNET	3666.15	544833	10731	.1134	22.87%	66.93%	.716

Chapter 5: Conclusion

In this paper we proved that 5G data systems and 802.11 networks have comparable attributes when sending data. We show that using these comparable attributes we can link 5G data systems and 802.11 networks in a way that allows for us to use previous works and datasets to improve upon techniques already laid out. Using these datasets and previous works, we found optimal techniques for identifying three different classes of traffic which were normal, flooding and injection traffic. Unlike previous works, we bought automated feature selection to the table and showed that we can achieve better results while taking less time. In the end, we found results that will allow future researches to build off this work, and to bring in datasets that come from 5G data transmissions, apply our work, and build better security that will benefit us all in the long run.

References

5G Explained - How 5G Works. (n.d.). Retrieved from http://www.emfexplained.info/?ID=25916

5G Network Architecture. (n.d.). Retrieved May 14, 2020, from https://www.huawei.com/minisite/hwmbbf16/insights/5G-Nework-Architecture-Whitepaper-en.pdf

AWID. (n.d.). Retrieved from http://icsdweb.aegean.gr/awid/

Brownlee, J. (2019, December 12). How to Perform Feature Selection With Machine Learning Data in Weka. Retrieved from https://machinelearningmastery.com/perform-feature-selection-machine-learning-data-weka/

Desarda, A. (2019, January 17). Understanding AdaBoost. Retrieved from https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe

Ferrag, M. A., Maglaras, L., Argyriou, A., Kosmanos, D., & Janicke, H. (2018). Security for 4G and 5G cellular networks: A survey of existing authentication and privacy-preserving schemes. *Journal of Network and Computer Applications*, *101*, 55–82. doi: 10.1016/j.jnca.2017.10.017

Geier, J. (2003, May 1). Denial of Service a Big WLAN Issue. Retrieved from https://www.esecurityplanet.com/trends/article.php/2200071/Denial-of-Service-a-Big-WLAN-Issue.htm

Khachatyran, H. (n.d.). LogitBoost. Retrieved from https://www.rdocumentation.org/packages/caTools/versions/1.17.1/topics/LogitBoost

Kirkby, R., & Frank, E. (n.d.). WEKA Explorer User Guidefor Version 3-4-3. *WEKA Explorer User Guidefor Version 3-4-3*. Retrieved from https://weka.sourceforge.io/manuals/ExplorerGuide.pdf?TB_iframe=true

Kolias, C., Kambourakis, G., Stavrou, A., & Gritzalis, S. (2016). Intrusion Detection in 802.11 Networks: Empirical Evaluation of Threats and a Public Dataset. *IEEE Communications Surveys & Tutorials*, *18*(1), 184–208. doi: 10.1109/comst.2015.2402161

LTE 4G & 5G Radio Access Network (RAN). (2019, December 16). Retrieved from https://www.cablefree.net/wirelesstechnology/4glte/lte-4g-5g-radio-access-network-ran/

McCormick, C. (2013, April 16). SVM Tutorial - Part I. Retrieved from https://mccormickml.com/2013/04/16/trivial-svm-example/

Mohr, R. (2019, November 15). Poking Holes in 5G with 5GReasoner. Retrieved from https://www.mobileiron.com/en/blog/5greasoner-5g-vulnerabilities

PortaAll, L. L., & Porta, L. L. (2020, February 27). What is a Man-in-the-Middle (MitM) attack? Signs and Prevention. Retrieved from https://www.wandera.com/what-is-a-man-in-the-middle-attack/

Williams, M. (2019, February 12). How to choose a TV antenna. Retrieved from https://www.techhive.com/article/3214772/which-tv-antenna-should-i-buy.html

Abubakar, A., & Pranggono, B. (2017). Machine learning based intrusion detection system for software defined networks. *2017 Seventh International Conference on Emerging Security Technologies (EST)*. doi: 10.1109/est.2017.8090413

Kholidy, H.A., Fabrizio Baiardi, Salim Hariri: 'DDSGA: A Data-Driven Semi-Global Alignment Approach for Detecting Masquerade Attacks'. The IEEE Transaction on Dependable and Secure Computing, 10.1109/TDSC.2014.2327966, pp:164–178, June 2015.

Kholidy, H.A., Fabrizio Baiardi, "CIDD: A Cloud Intrusion Detection Dataset For Cloud Computing and Masquerade Attacks ", in the 9th Int. Conf. on Information Technology: New Generations ITNG 2012, April 16-18, Las Vegas, Nevada, USA. http://www.di.unipi.it/~hkholidy/projects/cidd/

Kholidy, H.A., Fabrizio Baiardi, "CIDS: A framework for Intrusion Detection in Cloud Systems", in the 9th Int. Conf. on Information Technology: New Generations ITNG 2012, April 16-18, Las Vegas, Nevada, USA. http://www.di.unipi.it/~hkholidy/projects/cids/

Kholidy, H.A., Baiardi, F., Hariri, S., et al.: 'A hierarchical cloud intrusion detection system: design and evaluation', Int. J. Cloud Comput., Serv. Archit. (IJCCSA), 2012, 2, pp. 1–24.

Kholidy, H.A., "PH.D. Thesis: Cloud Computing Security, An Intrusion Detection System for Cloud Computing Systems". https://pdfs.semanticscholar.org/cf8a/14dc638480dbc5304824dd99a631d917d3fe.pdf

Kholidy, H.A., "Autonomous mitigation of cyber risks in the Cyber–Physical Systems", Future Generation Computer Systems, Volume 115, 2021, Pages 171-187, ISSN 0167-739X, https://doi.org/10.1016/j.future.2020.09.002. http://www.sciencedirect.com/science/article/pii/S0167739X19320680

Kholidy, H.A., Abdelkarim Erradi, Sherif Abdelwahed, Fabrizio Baiardi, "A risk mitigation approach for autonomous cloud intrusion response system", in Journal of Computing, Springer, DOI: 10.1007/s00607-016-0495-8, June 2016.

Kholidy, H.A., Abdelkarim Erradi, "A Cost-Aware Model for Risk Mitigation in Cloud Computing SystemsSuccessful accepted in 12th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA), Marrakech, Morocco, November, 2015.

Kholidy, H.A., Ali Tekeoglu, Stefano Lannucci, Shamik Sengupta, Qian Chen, Sherif Abdelwahed, John Hamilton, "Attacks Detection in SCADA Systems Using an Improved Non-Nested Generalized Exemplars Algorithm", the 12th IEEE International Conference on Computer Engineering and Systems (ICCES 2017), December 19-20, 2017.

Qian Chen, Kholidy, H.A., Sherif Abdelwahed, John Hamilton, "Towards Realizing a Distributed Event and Intrusion Detection System", the International Conference on Future Network Systems and Security (FNSS 2017), Gainesville, Florida, USA, 31 August 2017. Conference publisher: Springer. “Industrial control system (ics) cyberattack datasets, http://www.ece.uah.edu/~thm0009/icsdatasets/ PowerSystem_Dataset_README.pdf

Kholidy, H.A.,, Abdelkarim Erradi, Sherif Abdelwahed, Abdulrahman Azab, “A Finite State Hidden Markov Model for Predicting Multistage Attacks in Cloud Systems", in the 12th IEEE International Conference on Dependable, Autonomic and Secure Computing (DASC), Dalian, China, August 2014.

Kholidy, H.A., “Detecting impersonation attacks in cloud computing environments using a centric user profiling approach”, Future Generation Computer Systems, Volume 115, issue 17, Decmenr 13, 2020, Pages 171-187, ISSN 0167-739X, https://doi.org/10.1016/j.future.2020.12 , https://www.sciencedirect.com/science/article/abs/pii/S0167739X20330715