The Promise of Machine Learning in Cybersecurity

James B. Fraley
College of Engineering and Computing
Nova Southeastern University
jf1280@nova.edu

Dr. James Cannady
College of Engineering and Computing
Nova Southeastern University
cannady@nova.edu

Abstract— **Over the last few years' machine learning has migrated from the laboratory to the forefront of operational systems. Amazon, Google and Facebook use machine learning every day to improve customer experiences, suggested purchases or connect people socially with new applications and facilitate personal connections. Machine learning's powerful capability is also there for cybersecurity. Cybersecurity is positioned to leverage machine learning to improve malware detection, triage events, recognize breaches and alert organizations to security issues. Machine learning can be used to identify advanced targeting and threats such as organization profiling, infrastructure vulnerabilities and potential interdependent vulnerabilities and exploits. Machine learning can significantly change the cybersecurity landscape. Malware by itself can represent as many as 3 million new samples an hour. Traditional malware detection and malware analysis is unable to pace with new attacks and variants. New attacks and sophisticated malware have been able to bypass network and end-point detection to deliver cyber-attacks at alarming rates. New techniques like machine learning must be leveraged to address the growing malware problem. This paper describes how machine learning can be used to detect and highlight advanced malware for cyber defense analysts. The results of our initial research and a discussion of future research to extend machine learning is presented.**

Keywords
Belief Propagation, data mining, dynamic analysis, Locality Sensitive Hashing, malware detection, and static analysis.

I. INTRODUCTION

In 2015, a Global Information Security Workforce Study was conducted to better understand the worldwide need for security personnel [1]. The study conducted surveyed nearly 14,000 information security professionals by Information International Information Systems Security Certification Consortium, (ISC)2. The workforce study found that the gap for security personnel is projected to reach over 1.5 million by 2020 [1]. Other organizations have found the demand for cybersecurity resources outpacing supply for security personnel. Michael Brown, CEO of Symantec, said in a recent interview that he believes a global shortfall of 1.5 million will exist before 2019 [2]. Forbes projected that the cybersecurity market will double in size from $75 billion in 2015 to $170 billion by 2020 [2]. A Raytheon survey conducted in 2016 found that the demand for cybersecurity professionals' is growing at a rate 3.5 times faster than the IT job market [2]. In similar studies, Microsoft found that 35% of organizations surveyed were unable to fill key cybersecurity jobs. In the same survey, a majority of these companies expected a cyber-attack within the next 12 months [2].

Specific attacks like malware and Ransomware continue to pose a major challenge for most commercial, government and academic organizations. In 2015, ransomware paralyzed local governments as well as healthcare organizations by encrypting files until a "ransom" was paid [3]. According to Symantec, ransomware attacks grew by 35 percent in 2015 [3]. These highly profitable attacks allow ransomware developers to quickly reap hundreds of millions of dollars in payments [3]. In 2015, ransomware moved from traditional IT targets to smart phones and smart devices [3]. Symantec is convinced that ransomware will move to IoT, wearable devices (e.g. smart watches) and smart homes (e.g. thermostats, televisions, etc.) in the very near future [3].

The ability to train and provide cybersecurity expertise represents a daunting challenge for the global security community [1]. In many cases, these personnel need classroom, hands-on technology training and operational understanding to become proficient [1]. However, the education/training pipeline does not have the requisite volume to fulfill worldwide demand or come close in the next few years [1]. Organizations are facing a daunting challenge of dealing with the millions of new attacks a day. These attacks come in different sizes and different threat vectors including email phishing attacks, common platform and network hacking, polymorphic/metamorphic malware, distributed denial-of-service attacks and ransomware. These multi-vector attacks can quickly overwhelm many organizations [2]. The challenge for most security service and security management organizations. Organizations are also experiencing issues with security staff attrition, employee morale and psyche for cybersecurity analysts [4]. Organization's and the security community as a whole are looking for ways deal with non-stop barrage of attacks on data, networks, systems and mobile devices [4].

It is widely recognized that there is a global shortage of cybersecurity skills and talent [5]. Understanding the shortage has far greater impacts to business, national security, law enforcement and the intelligence community [6]. A number of experts have pointed to cybersecurity for contending with

978-1-5386-1539-3/17/$31.00 ©2017 IEEE

espionage, terrorism, financial crime, business insider threats, drugs and arms trafficking, as well as organized crime [6]. The outlook for closing the gaps looks bleaker after factoring the anticipating the wave of the Internet of Things (IoT) that is due to hit the security landscape by 2018. [2]. These shortages have significant implications across the security community as a whole [1].

The identified skills gap presents a unique opportunity for IT organizations, high school students, and university students who aspire to work in cybersecurity [6]. Cybercrimes that are committed globally such as financial fraud, online child exploitation, and payment scams happen at such frequency that they require a twenty-four-hour international response and cooperation from multi-national law enforcement agencies [6]. In 2014-2015, cybersecurity personnel were required to respond to a massive number of cyber-crimes including Office of Personnel Management, Blue Cross/Blue Shield Anthem, Target, Home Depot and Ashley Madison [7]. Adversaries exploited or hacked government and private computer systems by taking advantages of lax security, holes in security architecture and/or exploiting the vulnerabilities within the IT infrastructure [7].

This study argues that new techniques such as machine learning offers a unique opportunity to close the cyber skills gap by reducing the number of cybersecurity personnel needed to research, analyze and share malware detection information. Machine learning offers individuals interested in moving into the cybersecurity field a focal area for education and training.

II. Related Work

Machine learning applications provides a unique way to address fundamental scientific and engineering questions using computer software [8]. The field of machine learning has progressed dramatically over the past two decades and provides an easy onramp for those new to the field. Machine learning has emerged from a laboratory "black-box" environment to a practical application used by a growing number of commercial companies [8]. Machine learning has developed software applications computer vision, speech recognition, natural language processing, robot control, and other applications [8]. Major companies such as Amazon, Google and Facebook use machine learning to improve user experience, suggest purchases and promote special offerings. Machine learning developers find it far easier to train a system by developing examples of desired output rather than programming the traditional input-process-output [8]. The effect of machine learning has rippled across a number of industries with data-intensive issues like cybersecurity. Machine learning has offered similar opportunities for advancement other fields ranging from biology to cosmology to social science [8]. Machine leaning can process and analyze volumes of experimental data in novel ways [8]. Conceptually, machine learning algorithms can provide unique insight into "big data" and be optimized to improve associated performance metrics in vertical applications. Machine learning algorithms can vary greatly in terms of unique functions (e.g., decision trees, support vector machines, deep learning, neural networks and advanced clustering). However, machine learning offers ways to analyze volumes of data in unique ways such that evolutionary approach can be generated and successive generations of algorithms can provide greater optimization [8].

Machine learning and cybersecurity are ideally suited for processing large volume of data [8]. Networks and platforms are constantly under attack. These attacks are more effective given the number of tools to scan and evaluate targets. Adversaries are now using machine learning to advance their attacks. Most network protection devices provide logs and other traces of information regarding anomalous activities. However, these logs are low priority for most security operations. A best practice for log management is to consume these logs into a Security Incident and Event Management (SIEM) system to highlight events for cybersecurity analysts. Unfortunately, advanced attacks hide their footprint and action to bypass log management processing. McAfee has reported 500 million new types of malware samples in 2015 [7]. In addition, the sophistication of the malware makes detecting more advanced malware far more challenging for cybersecurity professionals [7]. Challenges with detection and signaling the events to cybersecurity analysts becomes more of a challenge [7]. New alerting methods must utilize technology and not labor to address the growing and high volume problem. It will be technology that supports the labor workforce to understand and react to cyber threats. Machine learning offers some promising results for dealing with security events once they occur.

III. Automated Security Event Response

State-of-the-art security devices typically provide system owners with alerts and events based on signature-based monitoring and abnormal behavior [9]. These events typically identify malicious activities and generate alerts at the network devices such as Intrusion Protection Systems/Intrusion Detection Systems (IPS/IDS), network protection devices (firewalls) or endpoint protection (Anti-Virus, Host Intrusion Protections (HIPS), Host Firewall etc.) Security devices that rely solely on signature detection suffer from to key two key limitations: 1) new or unknown attacks go undetected and 2) detection/protection is limited and typically not shared [9]. In the past, these devices have been able to provide adequate protection for large and small organizations by isolating or providing "defense-in-depth" [9]. Coordinating cyber-defense across the enterprise requires better coordination and correlation of diverse events that are seen at the network and host levels [9]. Most system owners are primarily concerned with defending systems and networks against known attacks [9]. However, advanced and mature organizations are equally concerned with identifying zero-day attacks and advanced persistent threats (APTs) [10]. These types of attacks are more difficult to detect as they are "low and slow" attacks which easily get lost in the "noise" of millions of other events.

In order to address the explosion of security events detected across organizations, security event management strategies must be developed in an adaptive approach in order to respond

to new and unique attacks while continuing to address the know attacks [11]. New event management strategies need to become situationally-aware and learn over time [11]. Today's Security Operation Centers (SOC) and Network Operation Centers (NOC) react to detected events by developing context of attacks. Cyber-defense analysts assemble alerts/incidents with attack forensic event information with additional information from sources such as packet capture, netflow data and device logs. Automating security event management processes demands foundational elements such as contextually and semantically rich detection information that assemble of ontologies and logic that can be shared across a security infrastructure [11]. Unfortunately, NOC/SOC analysts are required to assemble, correlate and explore security events manually across heterogeneous devices that provide diverse event information [12]. The modern enterprise typically employs various security point solutions that perform detection and protection independently. New solutions must bring these systems together and share information to promote adaptive learning for security events [12].

In researching security event management there are a number of approaches that enable adaptive learning. Sneath and Sokal offer an approach to analyze categorical similarity measures based on a numerical taxonomy [13]. Sneath and Sokal [13] provide an approach to solve problems and develop solutions based on categorical similarities and relevance. Their research demonstrated that through numerical and categorical analysis, problems can be grouped by determining similarity of categorical elements, features or attributes [13]. In addition, attributes can be measured to determine feature relevance [13]. However, most of the problems solved with this approach are binary in nature and do not lend themselves to complex problems like security events [13]. Understanding that most security events are bipartite or binary – good or bad – the information for determining these events are quite diverse and heterogenous in nature [14]. Wilson and Martinez [15] offer an approach to analyze heterogeneous events through distance functions that group events based on categorical and continuous attributes [15]. This approach provides a supervised learning approach for classes of information where each class has a set of categorical and continuous attributes that can be evaluated through distance metric learning [15]. This approach is extremely useful for small and limited datasets [15]. However, security event management has volumes of data that need to be analyzed. Other learning algorithms have been developed for diverse and large datasets.

Neural Networks (NN) and Deep Neural Networks (DNN) provide a powerful approach for analyzing diverse and large datasets. In addition, NN and DNN provide distinct advantages for complex and abstract layers for data input [16]. NN an DNN are commonly used effectively for natural language processing, speech recognition and image processing [16]. However, using NN and DNN to solve real-world issues such as security event management can be challenging. NN and DNN provide a general process for solving common problems. The process to train NN and DNN models has many issues and practical concerns. Training NN and DNN required specifying and labeling of parameters covering a large portion of the anticipated dataset. The training process must address the various scenarios anticipated by the model. Covering a majority of the data possibilities anticipated for complex datasets can be extremely challenging and time consuming. The training process must attempt to utilize convex optimization techniques to solve nonconvex data model problems [18]. NN and DNN approach solving nonconvex optimization issues by decomposing the larger problem into smaller pieces. This strategy operates on the premise that smaller pieces lend themselves to local convex optimization [18]. Overall, the assembling of local optimized components improves performance by using convex optimization techniques to solve the nonconvex problems [18]. Many times, optimization cannot be fully achieved and other methods must be used to optimize the DNN model. Ad hoc methods such as gradient clipping or batch normalization are need to address outlier issues. Ad hoc processes are many times needed to achieve a coherent model [19]. However, ad hoc methods must tune the network (NN or DNN) using back-propagation a part of training. Tuning the NN or DNN for complex issues is non-trivial. Training NN or DNN models with insufficient data can cause issues with overfitting [19]. This requires the practitioner to understand and construct the NN and DNN model with the right levels of complexity [19]. Therefore, the practitioner must be a subject-matter expert and have the proper data coverage for the problem being solved.

Our paper details the approach and results experienced with developing DNN for high volume security events.

IV. Study Approach

The goal of this study was to determine how machine learning could be leveraged to solve security event management. Many enterprise customers are experiencing millions of alerts a day [20]. McAfee reports that there are as many as 3 million new malicious files launched hourly [20]. The volume of malware and network detections quickly overwhelms cyber defense analysts and security operations. As discussed earlier, there is a shortage of security professionals so people cannot solve this problem. Technology must be leveraged to automate and orchestrate processing of security events. In particular, machine learning can provide an avenue for addressing both the volume and intrinsic knowledge provided by cyber defense analysts.

Events	Daily Issues	Automated Response	# Needing SME Analyst	Projected Manhours
Malware	1,486	1,219	267	669
Email	23,491	20,639	2,852	7,130
Network	4,239	3,193	1,046	2,614
Internet	13,112	10,651	2,461	6,153
Totals	**42,328**	**35,702**	**6,626**	**16,565**
Days of Effort--->				**2,071**

Fig. 1. Days of Effort

In order to understand the problems confronting today's cyber defense analysts, a quick review of real numbers provides deep insight into issues facing NOC and SOC personnel daily. As depicted in Fig. 1, the numbers presented

are real numbers from a medium to large NOC/SOC. The numbers here are typical for a mid-size organization that have under 10,000 users. The numbers also demonstrate the need for automation as the events could potentially consume over 2,000 man-days of effort to clear or evaluate these security events. For this company, the NOC/SOC has 12 analysts. On any given day, the capacity of this group is 96-144 hours. The shortfall is over 1,900 hours. The need for automation is apparent and working with the customers, we have developed an approach to walk through a process to implement machine learning in an operational environment.

Our approach followed distinct six phases:

1) *Develop Business Understanding*
2) *Analyze Data and Data Dependencies*
3) *Engage Subject Matter Experts*
4) *Prepare Dataset*
5) *Develop Model*
6) *Evaluate Model*

Each of the phases are discussed below:

Develop Business Understanding

The team met with stakeholders to discuss the problems and issues. The meeting resulted in the following observations and problems that must be addressed the machine learning project:

- Security analysts are overwhelmed with the sheer number of daily alerts
- Most alerts are common and have little significance or context
- Leadership understands the nature of the alerts but there must be some way to review of the "unknown" or "unconnected" alerts
- No method of quickly prioritizing and responding to alerts

Additional meetings were needed to discuss the project objectives. The following machine learning project objectives were approved by the analysts and stakeholders:

- Improve the efficiency of human security analysts
- Highlight known and evolving common threat vectors
- Enable human analysts to focus on more interesting and complex cyber issues
- Accelerate and highlight potential serious events (e.g. Zero-day and APT) through machine learning techniques and share "lessons learned"

Analyze Data and Data Dependencies

The team gathered representative alerts and analyzed previous work to understand the data, workflow and outcomes. This was a much bigger job than anticipated. The universe of events composed of almost 12,000 alerts. These alerts were originally broken into almost 20 categories. Also, some event had a low density – such that replicating events that were unique did not make sense. It became very clear that there were too many alerts – over 12,000. Initially, these alerts we put into 20 categories. After analyzing the data, the recommendation was made to consolidate the 20 categories into five (5). The recommended five categories include: 1) Malware, Reconnaissance, Denial of Service (DOS), Policy and Exploit. The team was also able to develop sub-categories for the major categories. These sub-categories are shown in Fig. 2.

Sub-category	Sub-category
arbitrary-cmd-execution	non-standard-port
audit	non-standard malware
backdoor	over-threshold
bot	PDF-Emulation
botnet	phishing
brute-force	port-scan
buffer-overflow	privileged-access
Cloud-Analysis-And-Deconstruction-	probe
code-execution	protocol-violation
covert-channel	pup
custom-fingerprinting	read-exposure
ddos-agent-activity	remote-access
dos	restricted-access
evasion-attempt	restricted-application
File-Mismatch	sensitive-content
fingerprinting	service-sweep
File-Reputation	shellcode-execution
Gateway-Anti-Malware	statistical-deviation
host-sweep	TIE-File-Reputation
Malicious-Flash-Analysis-Engine	trojan
multi-aid	unassigned
multi-aid-knownbot	worm
multi-aid-zeroday	write-exposure

Fig. 2. Alerts Sub-category

Engage Subject Matter Experts

Subject-Matter-Experts (SMEs) were engaged to review and validate categorization and sub-categories. SME also provide insight to the actions taken, independent and dependent alerts and the priorities associated with alert categories and subcategories. SMEs also were able to evaluate the frequency and number of alerts. SMEs were also leveraged to provide priority and weighting for the various alerts. After performing a thorough weighting process for over 8,900 alerts, an analysis of the weights to understand the minimum values, maximum values, average, median and min-max average. The analysis and values are presented in Fig. 3.

SME's also wanted to be very clear regarding the outcomes expected from the project. These outcomes are listed below:

- Utilize advanced machine learning techniques to provide a rapid analysis and recognition of common threat vectors
- Classification of incidents using deep neural networks
- Human analysts to focus on more interesting and complex cyber issues by providing the analyst with one or more recommended responses to common security incidents

- Automate various reports (STIX/TAXI) to share Indicators of Compromise
- Refer high priority incidents directly to SMEs

Categorical Weight	Cat/Sub Cat Weight
Min	**Min**
0.75	1.75
Max	**Max**
6.75	19
Avg	**Avg**
4.0825	12.7765
Median	**Median**
4.5	14
MinMaxAVG	**MinMaxAVG**
3.75	10.375

Fig. 3. Weighting Analysis

Prepare Dataset

The it has been found that data preparation may take 75-85% of the total data engineering effort. It is important to note that the dataset used for machine learning must be reviewed, validated and in some cases transformed in order to produce quality results. The dataset provides the critical input to the various NN and DNN algorithms in order to create the learning model. This process involves analyzing the dataset for missing or incomplete data, understanding the distribution of data and range of values contained within the dataset. In order to generate a valid model - data preparation provides three key aspects: (1) highlights hidden patterns, (2) improves erformance and (3) produces higher quality outputs. The dataset used was labeled and all data was validated prior to introducing the data to the model.

Develop Model

The team developed the model from the prepared dataset. There were multiple runs in order to consume the dataset into a single model. TensorFlow was used to create and experiment with the model. TensorFlow was selected for this research because of the distributed architecture. TensorFlow can efficiently utilize hundreds of servers to quickly train and develop advanced NN and DNN models. TensorFlow has a unique ability to deliver and manage both computational and state management through parallelization. This allowed the team to perform various tests and experiments with a single dataset.

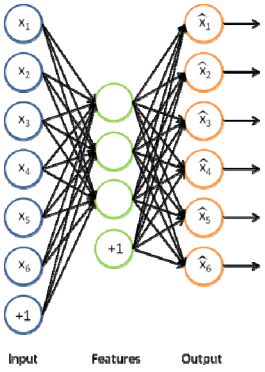


Figure 4. Neural Network Model

Evaluate Model

The evaluation of the model created with the test dataset is measured or evaluated based on the classifier performance. Classifier performance is measured based on common stratified k-fold cross-validation. Stratified cross-validation examines class distribution and how those classes are distributed consistently across each fold. The goal of the cross-validation is to evaluate the classifier's performance for correctly classifying the data instances from the test dataset. Model validation also evaluates instances of the dataset that were correctly labeled.

V. Experimental Environment

Overview

The architecture and dataset are provided for reference.

TensorFlow Architecture

The various TensorFlow components are depicted below.

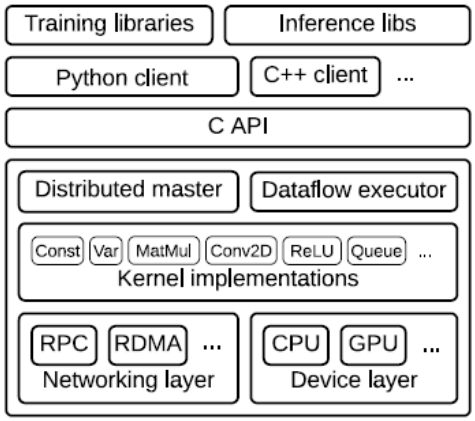
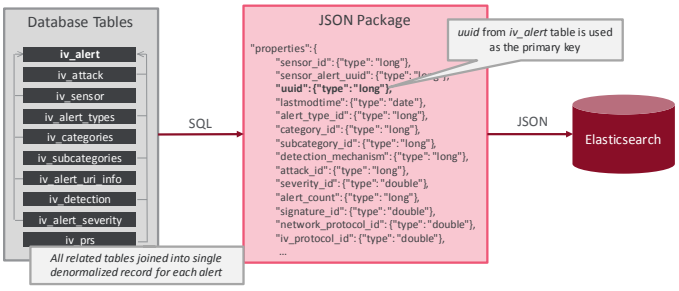


Fig. 5. TensorFlow Architecture

Dataset

The dataset generation is provided below.



VI. Findings and Results

Study Highlights

The machine learning experiment conducted evaluate how machine learning could be leverage for security event management. The experiment focused on evaluating ability of machine learning to accurately predict the actions of a security analyst given the training dataset. Over 9 million alerts were classified using previous analyst decisions for training the neural network. The end result was that the model produced impressive results for classifying alerts: ~99% accurate.

VII. Conclusion

The goal of this research was to better understand how machine learning could be leveraged to classify various security events and alerts. The experiment systematically walked through developing requirements/objectives, assembling data, validating the dataset and generating a neural network model. The end goal for the experiment was to test whether the model would properly react to security events by alerting SMEs, alerting analysts or producing reports depending upon the severity of the security event. The model did perform these functions with very high accuracy (90%). This research used TensorFlow to develop the neural network model. For over 9 million security events the model correctly identified and properly responded to the security events presented in the test dataset. Analyzing the results from the experiment, security analyst time could be reduced by 78% by incorporating machine learning into SOC/NOC operations. The initial problem of over 2,000 hours a day could be dramatically reduced to 455 with machine learning assistance.

Future work should seek to provide additional support for cyber defense analyst to further reduce the time demand for responding to critical security events.

References

[1] ISC2. "The 2015 (ISC) 2 Global Information Security Workforce Study."

[2] S. Morgan, "One Million Cybersecurity Job Openings in 2016. Forbes." (2016).

[3] Enterprise, Symantec. "Internet Security Threat Report 2015." (2016).

[4] L. Fourie, S. Pang, T. Kingston, et al. "The global cyber security workforce: an ongoing human capital crisis." (2014).

[5] K. Evans and F. Reeder. “A Human Capital Crisis in Cybersecurity: Technical Proficiency Matters”. CSIS, 2010.

[6] K. Francis and W. Ginsberg, The Federal Cybersecurity Workforce: Background and Congressional Oversight Issues for the Departments of Defense and Homeland Security.

[7] McAfee Labs Report, March 2016.

[8] M. Jordan and T. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349, no. 6245 (2015): 255-260.

[9] W. Lynn, "Defending a new domain: the Pentagon's cyberstrategy." Foreign Affairs 89, no. 5 (2010): 97-108.

[10] P. Duessel, C. Gehl, U. Flegel, et al. "Detecting zero-day attacks using context-aware anomaly detection at the application-layer." International Journal of Information Security (2016): 1-16.

[11] K. Thakur, S. Kopecky, M. Nuseir, et al. "An analysis of information security event managers." In Cyber Security and Cloud Computing (CSCloud), 2016 IEEE 3rd International Conference on, pp. 210-215. IEEE, 2016.

[12] Bayer, Ulrich, Paolo Milani Comparetti, Clemens Hlauschek, Christopher Kruegel, and Engin Kirda. "Scalable, Behavior-Based Malware Clustering." In *NDSS*, vol. 9, pp. 8-11. 2009.

[11] Bailey, Michael, Jon Oberheide, Jon Andersen, Z. Morley Mao, Farnam Jahanian, and Jose Nazario. "Automated classification and analysis of internet malware." In *Recent advances in intrusion detection*, pp. 178-197. Springer Berlin Heidelberg, 2007.

[12] X. Zhang, M. Clark, K. Rattan, et al. "Controller integrity monitoring in adaptive learning systems towards trusted autonomy." IEEE Transactions on Automation Science and Engineering 13, no. 2 (2016): 491-501.

[13] T. Bhavani, M. Kantarcioglu, K. Hamlen, et al. "A Data Driven Approach for the Science of Cyber Security: Challenges and Directions." In Information Reuse and Integration (IRI), 2016 IEEE 17th International Conference on, pp. 1-10. IEEE, 2016.

[14] G. Granadillo, G. Gonzalez, M. El-Barbori, and H. Debar. "New Types of Alert Correlation for Security Information and Event Management Systems." In New Technologies, Mobility and Security (NTMS), 2016 8th IFIP International Conference on, pp. 1-7. IEEE, 2016.

[15] M. Huang, W. Lin, C. Chen, et al. "Data preprocessing issues for incomplete medical datasets." Expert Systems 33, no. 5 (2016): 432-438.

(Krizhevsky et al., 2012; Sutskever et al., 2014

[16] W. Shang, W. Sohn, D. Almeida, and H. Lee. "Understanding and improving convolutional neural networks via concatenated rectified linear units." In Proceedings of the International Conference on Machine Learning (ICML). 2016.

[17] C. Cortes, X. Gonzalvo, V. Kuznetsov, M. Mohri, and S. Yang. "AdaNet: Adaptive Structural Learning of Artificial Neural Networks." (2016).

[18] A. Johansen, A. Rosenberg, J. Hansen, et al. "Neural Machine Translation with Characters and Hierarchical Encoding." (2016).

[19] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. "Rethinking the inception architecture for computer vision." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826. 2016.