Journal of Industrial Information Technology and Application *ISSN(Online): 2586-0852*
Vol.7 No.4

Deep Learning for Cyber Security Applications: A Comprehensive Survey

Tatiparthy Ramesh reddy[1*], and Dr. D. Usha[2]

[1] CSE DEPT, Dr. M.G.R Educational and Research Institute, Chennai, India.

[2] CSE Dept,Dr.M.G.R. Educational and Research Institute, Chennai, India.

Abstract. Due to its successful use in many traditional artificial intelligence (AI) problems as compared to standard ML algorithms, Deep Learning (DL), a novel form of machine learning, is attracting a lot of research interest (CMLAs). For a variety of applications in the field of cyber security, DL architectures have recently been creatively developed. As researchers explore various cutting-edge DL models and prototypes that may be customised to fit particular cyber security applications, the literature on DL architectures and their modifications is expanding. A thorough review of these research studies is however lacking in the literature. Numerous survey based studies lack a futuristic evaluation and instead concentrate on certain DL designs and specific harmful attack types within a constrained cyber security problem scenario from the past. With regard to next-generation cyber security scenarios involving intelligent automation, the Internet of Things (IoT), Big Data (BD), Blockchain, cloud, and edge technologies, this article intends to provide a comprehensive and well-rounded survey of the past, present, and future DL architectures. This work compares and analyses the contributions and difficulties from many recent research papers to give a tutorial-style thorough analysis of the state-of-the-art DL architectures for various applications in cyber security. First and foremost, the survey is distinctive in that it reports the use of DL architectures for a wide range of cybercrime detection techniques, including intrusion detection, malware and botnet detection, spam and phishing detection, network traffic analysis, binary analysis, insider threat detection, CAPTCHA analysis, and steganography. Second, the survey discusses important DL designs in areas of cyber security such encryption, cloud security, biometric security, IoT, and edge computing. Thirdly, the demand for DL-based research for the next-generation cyber security applications in cyber physical systems (CPS) that leverage BD analytics, natural language processing (NLP), signal and image processing, and blockchain technology for smart cities and Industry 4.0 of the future is discussed. Finally, a critical analysis of current issues and the new DL design that has been presented advances the field of study.

[*] Corresponding author: rameshdba329@gmail.com

Received: Oct 31, 2023; Accepted: Nov 30, 2023; Published: Dec 31, 2023

Keywords. Cyber Security, Machine Learning, Neural Networks, Deep Learning, Communication Networks, Cloud and Edge Computing.

1. Introduction

We are entering Industry 4.0 due to a rapid advancement in cyber physical systems (CPS), which are driven by technologies like cloud computing, mobile computing, edge computing, and the Internet of Things. The internet has become a resource that is necessary for everyone (IoT). Due to the inherent security risks and vulnerabilities that arise as systems become more heterogeneous, sophisticated, and networked, however, the subject of cyber security in CPS is also gaining relevance. 2018 saw a 13% increase in the overall number of vulnerabilities [1]. By 2021, it is anticipated that the number of zero-day exploits observed in the wild would increase from one every week in 2015 to one per day [2]. While the demand for cyber security specialists is rising globally to address this issue, there is a scarcity of qualified researchers and practitioners with a potential shortfall of up to 25% [3]. A survey that acts as a lesson for cyber security experts is required. To assist in solving the significant issue of cyber security for upcoming ICT systems, it is crucial to identify the gaps in the body of literature.

The phrase "cyber security" now refers to a collection of ideas and techniques used to safeguard ICT systems and networks with the aim of maintaining the privacy, accuracy, and accessibility of data in the cyberspace. Computer gear, networks, and software are intentionally assaulted as a result of illegal acts committed in the CPS. In addition, the dangers of unauthorised access, theft, disclosure, and malicious or unintentional harm to data integrity are becoming more and more significant. There have been more criminals and enemies in the cyber security area over time, but the broad categories of threat have remained constant. Security research's major purpose is to stop attackers from achieving their objectives, thus it's crucial to have a thorough understanding of the different kinds of assaults that might be used. Different cyber security strategies, including intrusion detection (ID), social network analysis, malware analysis, advanced persistent threats, online application security, and applied cryptography are being used to combat these dangers. Nevertheless, despite the massive development of CPS towards Industry 4.0, there is still a lack of an adaptive cyber security architecture that can proactively react to changes in systems and physical processes.

The vast amounts of data collected by network sensors, logs, and endpoint agents used by modern cyber security technologies can be analysed effectively utilising data mining (DM) techniques to deliver timely information about harmful activities. Since DM techniques successfully extract the hidden features to distinguish between legitimate and malicious activities, legacy cyber security solutions like host- and network-level firewalls, antivirus software, intrusion detection systems (IDSs), and intrusion protection systems have entered the market (IPSs). However, these tools only work well in identifying established malicious actions and miss new kinds of dangerous activity linked to Industry 4.0 and big data (BD). Due to their lack of domain knowledge and the BD features of massive data in various formats, types, and modalities, such cyber security solutions face numerous difficult problems. BD analytics is capable of gathering, storing, processing, and visualising a huge volume of data. As a result, it is crucial to apply BD analytics to cyber security, which has lately formed a new study direction.

Data created by end-user systems now reveal unidentified patterns due to the development of new CPS technologies, which cannot yet be modelled to determine if they are benign or malicious. Due to their intrinsic processing limits, the resource-constrained IoT devices also experience a number of vulnerabilities and security breaches. In order to incorporate these artificial intelligence (AI) based techniques into systems that make decisions as close to domain experts as possible, cyber security specialists and researchers are exploring cognitive technologies using machine learning (ML) and deep learning (DL) for cyber security [4]. However, it is acknowledged that, as detailed in the literature already in existence [5], ML/DL deployment has the potential to be misleading.

In the fields of computer vision and speech processing, ML and DL have emerged as fundamental tools. In various computer vision and healthcare-related applications, the DL architectures have surpassed domain experts and achieved greater performance than standard ML algorithms (CMLAs) [6]. The fact that CMLAs rely heavily on feature engineering techniques typically prescribed by domain experts is one of their main drawbacks. Security experts are therefore investigating DL techniques to deal with the nefarious behaviours that are always developing. DL scales better than CMLAs for very large amounts of data samples because it can extract the key characteristics from intricate systems like natural language processing (NLP) [5]. The efficiency of DL-based solutions also gets better when the amount of cyber security data keeps expanding daily as a result of technological advancement. As we go into Industry 4.0, a number of DL architectures have recently been offered, thus it's critical to assess their relative adequacy to handle the escalating cyber security challenges.

A. Existing Surveys on ML and DL on Cyber Security

A thorough review of traditional ML frameworks created in the last decade to address cyber security issues can be found in [7], [8], [9], [10], [11], [12]. These, however, do not cover DL techniques. Surveys on DL frameworks have only addressed a small subset of cyber security related applications. Most research studies concentrate on a single cyber security approach, such as malware analysis [15], spam detection [13], anomaly detection [14], intrusion detection [13], and spam detection [14]. Recent research has concentrated on presenting an overview of the work done in order to defend CPS [16], and some studies have covered various ML and DL approaches for protecting IoT technology [17]. Some surveys, such as those that use short tutorials, either have a specific focus on the application of deep reinforcement learning (DRL) to cyber security or use general DL frameworks to detect a variety of attacks, such as malware, spam, insider threats, network intrusions, false data injection, and malicious domain names used by botnets [18] [19], [20]. The scope of earlier surveys was, however, constrained because they only covered a limited number of cyber security scenarios. The use of DL for protecting next

generation communication networks, including blockchain, cloud/edge computing, and autonomous vehicle networks, which are next-generation technologies moving toward 5G network and Industry 4.0, is specifically not included. Despite the fact that there have been numerous surveys on the use of ML and DL in cyber security, to the best of our knowledge, no detailed and thorough survey on the various DL research projects has yet been carried out with the breadth and scope necessary to take into account the complexity of next-generation computing. This essay seeks to close this void in the literature.

B. Research Contribution

Our goal in this study is to give a thorough and in-depth assessment on deep learning (DL) for cyber security, taking into account the shortcomings of the state-of-the-art surveys in literature relevant to DL architectures for next-generation computing. The following is a summary of our survey's main contributions:

- This study presents an overview of the evolution of the DL architectures used for diverse cyber security applications. This survey also provides a thorough grasp of the past, present, and future DL applications in cyber security by summarising, contrasting, and comparing the different DL designs.
- This survey categorises numerous research that have discussed the use of DL for cyber security based on a number of factors, including the type of architecture and its application, the year of the study, text representation, the type of dataset, and performance evaluations against CMLA.
- An overview of the many problems and significant difficulties with off-line and real-time deployment of cyber-security software is given. Also discussed is the value of shared tasks in the sphere of cyber security.

- The most advanced adversarial machine learning (ML) and reinforcement learning (RL) applications in cyber security are looked at.
- Signal and image processing, NLP, DL in BD, and their significance for cyber security are reviewed.
- Smart cities, pervasive computing, biometrics, IoT, fog and cloud computing, and autonomous vehicles are some of the areas where DL architecture-based cyber security is used.

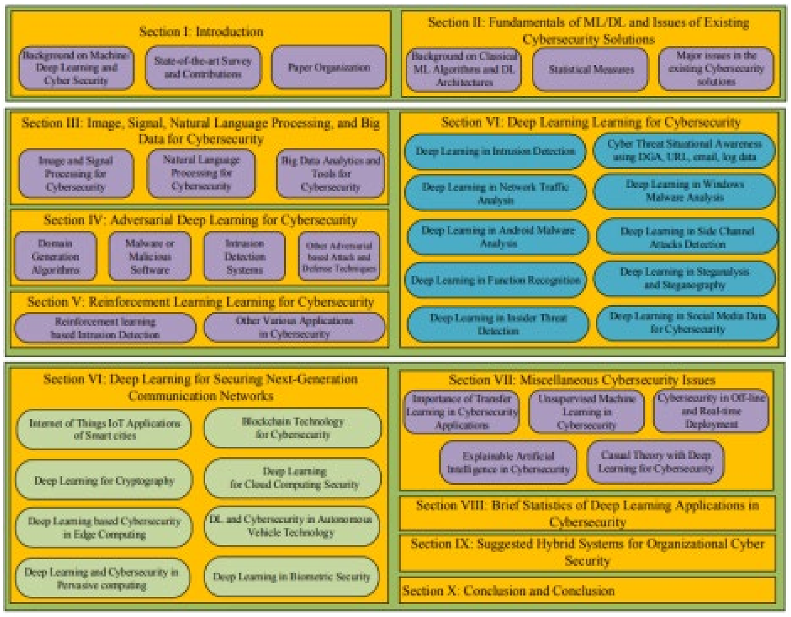
Section I: Introduction
Background on Machine Deep Learning and Cyber Security
State-of-the-art Survey and Contributions
Paper Organization
Section II: Fundamentals of ML/DL and Issues of Existing Cybersecurity Solutions
Background on Classical ML Algorithms and DL Architectures
Statistical Measures
Major issues in the existing Cybersecurity solutions
Section III: Image, Signal, Natural Language Processing, and Big Data for Cybersecurity
Image and Signal Processing for Cybersecurity
Natural Language Processing for Cybersecurity
Big Data Analytics and Tools for Cybersecurity
Section IV: Adversarial Deep Learning for Cybersecurity
Domain Generation Algorithms
Malware or Malicious Software
Intrusion Detection Systems
Other Adversarial based Attack and Defense Techniques
Section V: Reinforcement Learning Learning for Cybersecurity
Reinforcement learning based Intrusion Detection
Other Various Applications in Cybersecurity
Section VI: Deep Learning Learning for Cybersecurity
Deep Learning in Intrusion Detection
Cyber Threat Situational Awareness using DGA, URL, email, log data
Deep Learning in Network Traffic Analysis
Deep Learning in Windows Malware Analysis
Deep Learning in Android Malware Analysis
Deep Learning in Side Channel Attacks Detection
Deep Learning in Function Recognition
Deep Learning in Steganalysis and Steganography
Deep Learning in Insider Threat Detection
Deep Learning in Social Media Data for Cybersecurity
Section VI: Deep Learning for Securing Next-Generation Communication Networks
Internet of Things IoT Applications of Smart cities
Blockchain Technology for Cybersecurity
Deep Learning for Cryptography
Deep Learning for Cloud Computing Security
Deep Learning based Cybersecurity in Edge Computing
DL and Cybersecurity in Autonomous Vehicle Technology
Deep Learning and Cybersecurity in Pervasive computing
Deep Learning in Biometric Security
Section VII: Miscellaneous Cybersecurity Issues
Importance of Transfer Learning in Cybersecurity Applications
Unsupervised Machine Learning in Cybersecurity
Cybersecurity in Off-line and Real-time Deployment
Explainable Artificial Intelligence in Cybersecurity
Casual Theory with Deep Learning for Cybersecurity
Section VIII: Brief Statistics of Deep Learning Applications in Cybersecurity
Section IX: Suggested Hybrid Systems for Organizational Cyber Security
Section X: Conclusion and Conclusion

Figure 1. Diagrammatic view of the organization of this paper.

- 7) Significance of unsupervised learning for cyber security over semi-supervised and supervised learning is explored. Moreover, importance of explainable AI, transfer learning, visualization and hybrid framework in cyber security is summarized.
- 8) Many publicly available datasets used for various cyber security studies are reviewed and suggestions for future research directions are provided.

C. *Paper Organization*

The layout of this survey piece is depicted in Fig. 1. The fundamentals of various CMLAs and DL architectures are presented in Section II, and the main problems with cyber security are also covered. The significance of DL architectures in large data, signal and image processing, and natural language processing methods appropriate for cyber security are examined in Section III. Section V offers a thorough discussion of the uses of RL in cyber security after Section IV presents adversarial DL in that field. Section VI describes the state-of-the-art DL architectures for various cyber security techniques, including intrusion detection, cyber threat situational awareness using domain generation algorithm (DGA), uniform resource allocator (URL), email and security log data analysis, network traffic analysis, Windows/Android malware analysis, side channel attacks detection, insider threat detection, function recognition, steganalysis and steganography, and social media data for cyber security. In Section VII, we provide an overview of the application of DL for various technologies in next-generation communication networks, smart city, blockchain, cryptography, cloud computing, edge computing, autonomous vehicle networks, pervasive computing, and biometric security. In Section VIII, various topics related to the use of DL for cyber security are discussed, such as the value of transfer learning (TL) in applications, the use of unsupervised learning in cyber security, the use of off-line and real-time deployment, the function of explainable AI in cyber security, and the use of causal theory in conjunction with DL for cyber security. Finally, in Section X, we propose a hybrid cyber security framework of submodules of best theories and DL models, and in Section XI, we provide conclusions along with future research directions. Section IX reports detailed statistics of DL applications in cyber security. Section X and Section XI are where we propose the hybrid framework.

2. Fundamentals Of Dl And Issues Of Existing Cyber Security Solutions

A. Basics of Classical ML Algorithms and Deep Learning

Artificial intelligence (AI) was first used by John McCarthy in 1955, and McCarthy described AI as "the science and engineering of constructing intelligent machines." The introduction of ML, a subset of AI, occurred in the same decade as AI, but it gained popularity in the 1990s due to advancements in computing technology and the explosive growth of digital data. In machine learning (ML), mathematical and statistical ideas serve as the fundamental building blocks for a wide range of complicated algorithms that are generally used to find patterns, correlations, and anomalies in data. ML algorithms produce results that are expressed as probabilities and confidence intervals. Due to the limitations in having experts to analyse huge amounts of data, ML algorithms are employed to successfully automate the learning process for AI based automation.

In the last ten years, ML has improved in its use in cyber security [6]. Algorithms used in machine learning (ML) can generally be divided into five categories: supervised, semisupervised, unsupervised, reinforcement learning, and active learning. Algorithms for supervised learning are task-driven and rely on the classification of sample files as malware or not. Preprocessing and feature engineering are necessary for supervised ML algorithms. Naive Bayes (NB), Logistic Regression (LR), Decision Tree (DT), Ada Boost (AB), Random Forest (RF), and Support Vector Machine are examples of CMLAs that are often employed (SVM). Data-driven unsupervised learning involves implicit learning and labelling based on the distribution of the data and simply needs a sample set of data. While the performance of unsupervised models is lower as compared to supervised models, they are preferred in real-time cyber security applications as manual labelling of sample data is a tedious task. Semi-supervised learning combines both the supervised and unsupervised learning to get benefits from both approaches. RL is an environment driven approach which works based on rewards and is improved by a trial and error approach. Most of the DL based real-time systems in current days are based on RL. This is a suitable method for malware and botnet detection in the domain of cyber security. Active learning is a sub method of RL that contacts the user whenever a new data sample is seen.

The three primary components of CMLAs are 1) the collecting of raw data, 2) feature extraction, and 3) classification. In feature engineering, feature extraction is a crucial phase that calls for expertise in the area. The feature extraction is inherently dependent on the classifier's performance. Without human involvement, a neural network (NN) is capable of automatically extracting features and classifying them. The traditional NN performs admirably to a certain extent. However, with advanced NN, sometimes referred to as DL, the feature engineering process can be completely omitted. Due to this, the DL was able to deliver the greatest results in established AI applications across numerous disciplines.

Both security academics and individuals from the security industries have turned to DL as a focal point. DL is presently being used to solve a variety of cyber security issues and outperforms CMLAs in all use cases. According to Fig., there are two types of DL architectures: generative and discriminative. 2. Generative category is composed of deep Boltzmann machine (DBM), deep autoencoder (DAE), deep belief network (DBN), and recurrent structures, whereas discriminative category is composed of recurrent structures and convolutional neural network (CNN). Recurrent structures and CNN most commonly use DL architectures. Both DBN and DBM are based on the restricted Boltzmann machines (RBM). The generative adversarial network (GAN) belongs to both the generative and discriminative DL architecture category. The combination of DL and artificial neural networks (ANN) is called deep neural networks (DNN). While we have provided a pictorial overview of the classification of various DL architectures in Fig. 2, a detailed description and survey reports are available in literature [13], [14], [15].

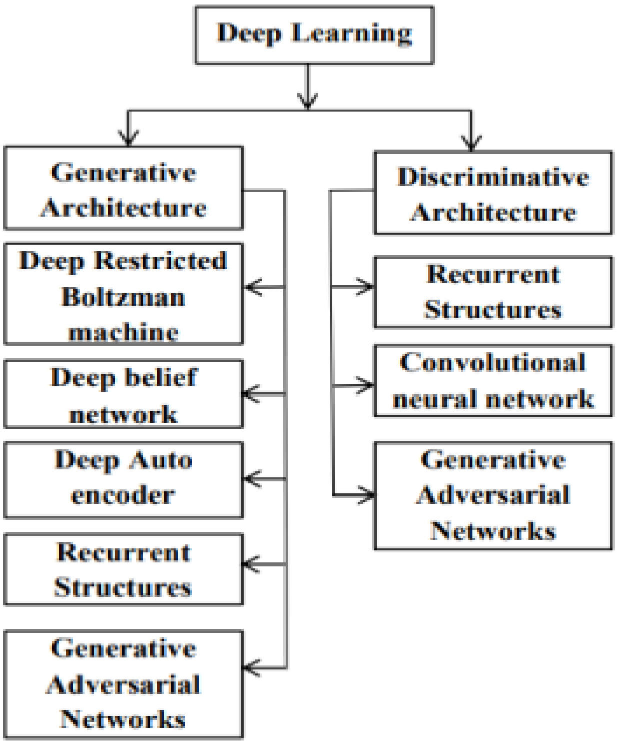
Deep Learning
Generative Architecture
Deep Restricted Boltzman machine
Deep belief network
Deep Auto encoder
Recurrent Structures
Generative Adversarial Networks
Discriminative Architecture
Recurrent Structures
Convolutional neural network
Generative Adversarial Networks

Figure 2. Classification of DL architecture.

Several statistical criteria are employed to assess the performance of the various DL models that are now available. The confusion matrix, which offers the specifics of the classification findings, including the individual differentiated classes reached, is one of the most significant and often used metrics. True Positive (TP) and True Negative (TN) are two metrics used to assess how accurately the DL model categorised the positive and negative data, respectively. False positive (FP) and false negative (FN) measurements also show that the DL model mispredicted the positive and negative data, respectively. Using the confusion matrix, some metrics can be estimated such as accuracy, precision, recall/true positive rate (TPR)/sensitivity, F1-measure/F1- score, false positive rate (FPR), true negative rate (TNR), and false negative rate (FNR) [19]. The values of Accuracy, Precision, Recall, F1- score, FPR, TNR and FNR range from 0 to 1 with larger values representing better performance. Since these measures are correlated, any desire to increase one measure such as TPR may result in an undesired increase of another measure such as FPR. Therefore, during the design phase, an optimal detection accuracy is usually assessed based on a discrimination threshold that reflects the dependency of TPR on FPR, which is represented by the Receiver Operating

Characteristics (ROC) curve. For the purpose of benchmarking, the area under the ROC curve (AUC) is estimated. AUC values typically lie between 0.5 to 1.0, and larger AUCs represent better performance.

B. Key Deep Learning Architectures

Due to their extensive applications in cyber security that have been documented in recent literature, this section summarises the key characteristics of four important DL designs using ANN (DNN).

- Deep Boltzmann machine (DBM) and deep belief network (DBN): DBN is based on generative engineering and is similar to the traditional ANN. It has one input layer, one output layer, and at least one hidden layer. It should be noted that feed forward network and DBN both have one layer (FFN). There should be at least one neuron, which is what scientists refer to as a processing unit, in both the input and hidden layers. Every class that is necessary for the network to classify the inputs has a unit in the output layer. In addition, a network with more than one hidden layer may expend more opportunity for its assembly. For instance, an unsupervised learning component such as the restricted Boltzmann machine (RBM) could take in the minimized element vectors by passing an input vector through at least one of the RBM hidden layers within the preparation stage. DBN's training phase has two steps namely pre-training and reconstruction. Given the training samples without class labels, the pre-training stage propagates the input stochastically across RBM layers. Each layer of RBM learns features which represent the data in the previous layer with associative memory present at the top layer. Conditional distribution is followed by each hidden layer unit to generate binary form feature vectors that are propagated in reverse direction to reconstruct the training samples. This procedure is followed iteratively for all the training samples.

- Autoencoders (AE) are a class of NN architectures with identical input and output layer dimensions that are designed to learn alias representations of input data using linear or nonlinear operations. The reduction of dimensions is the primary goal of AE. Researchers have been found to use many hidden layers in recent literature to uncover representative and discriminative characteristics of raw data. DAE is the name of this network type. These are trained to learn from the input, as opposed to the conventional NN architecture, which is trained to learn predetermined output variables. As a result, the NN develops the ability to recreate the input data on its own. The architecture of an AE is similar to the multi-layer perceptron (MLP), i.e., it has one input layer, one or more hidden layers, and one output layer. If AE has multiple hidden layers, then the features extracted from one layer are further processed to different features that are capable of reconstructing the data. During data reconstruction process, AE aims to minimize the error. Therefore, the outputs of intermediate layers are nothing but an encoded version of their inputs capable of reconstructing the input data under specific conditions.

In typical transformations, a certain set of features is chosen using the data points and then supplied as input to classification algorithms. The unsupervised strategy used by AE, on the other hand, involves extracting various features from various layers and passing them to additional DL layers, including CNN, RNN, and hybrid networks including CNNRNN and CMLAs. There are three well-known AE variants: contractive AE, denoising AE, and sparse AE, which promotes sparsity by having more hidden nodes than inputs and outputs. However, only a portion of the hidden units are activated at a given time. This is accounted for by penalizing the activation of additional nodes. DAE recovers the correct input from a corrupted version to increase the robustness of the model. Contractive AE achieves this by adding an analytic contractive penalty to the reconstruction error function. Overall, the DAE architecture is more robust for noise, while contractive AE can capture the local directions of variation dictated by the data.

- Convolutional neural networks (CNN) are a well-liked, cutting-edge technique that are used in many computer vision applications. Convolutional, pooling, and fully linked layers make up CNN [21]. A convolutional layer extracts the best features, collectively known as feature maps, by moving kernels or filters along the data's various dimensions (1D/2D/3D/4D). The pooling layer is then given access to these feature maps. Because they take into account nearby data, convolutional and pooling layers are both translationally invariant. The feature maps are initially partitioned, and different pooling functions are employed to lower the feature maps' dimensionality, which is nothing more than a non-linear down sampling procedure. The common pooling operations are maximum, minimum, average, stochastic, spatial pyramid, and deformation value from the partition. The stochastic pooling is similar to maximum pooling but it also prevents overfitting by replacing the conventional deterministic pooling operations with a stochastic procedure determined by the activation within each pooling region according to a multinomial distribution. Generally, CNN network can handle only the fixed length input representations. To handle variable length input representations, spatial pyramid pooling can be used as it can handle input images of variable scales, sizes and aspect ratios. A deformation pooling operation can handle deformation in image efficiently when compared to the max and average pooling. The novelty in the DL architecture could be explored by combining the different pooling layers to boost the performance of the CNN architecture. On the ImageNet Large Scale Visual Recognition Challenge, multiple benchmark structures are suggested and evaluated based on CNN (ILSVRC). LeNet, AlexNet, ZFNet, GoogleNet/Inception, VGGNet, SPPNet, ResNet, DenseNet, squeezenet, MobileNet, and NASNet are important CNN-based architectures. All of these architectures have a significant number of parameters and are frequently used with sizable datasets. Data augmentation is used to expand the data samples without adding additional labelling expenses because it is challenging to gather huge datasets for all the classes/tasks in real-time. In newer jobs that need pretraining, it is possible to use CNN architectures rather than random parameter values for parameter initialization. This enhances model generalisation and speeds up the learning process

- Recurrent structures are typically employed in tasks involving the modelling of sequential and temporal data in recurrent neural networks (RNN). A self-recurrent link in the hidden layer of the RNN, an upgraded model of traditional NNs, enables the network to retain the details of the previous phase. When dealing with large time-steps during backpropagation through time, it has disappearing and expanding gradient issues (BPTT). One of the popular solutions to the increasing gradient problems is gradient trimming. To alleviate the vanishing issue, research on RNN progressed on three significant directions: i) Hessian-free optimization for improving the optimization methods, ii) Long Short-term Memory (LSTM) or a variant of LSTM network with reduced parameters set, gated recurrent unit (GRU) for introducing complex components in recurrent hidden layer of network structure, and iii) Identityrecurrent neural network (IRNN) for weight initialization with an identity matrix.

C. Major issues in the existing Cyber security solutions

Due to its BD properties, the availability of multi-core CPUs and GPUs, as well as the development of NNs to train numerous hidden layers, DL serves as the greatest fit for cyber security. However, because to a lack of benchmarking in ML and DL methods as well as datasets, adoption is still in its infancy. The difficulties and problems associated with using ML/DL approaches in cyber security have recently been highlighted [24]. We include the threat detection techniques investigated and the benchmark datasets used in Table I. Finding a satisfactory dataset for cyber security use cases is often troublesome due to four main reasons: 1) the vast majority of publicly available datasets are outdated, 2) they are not genuine agent datasets, 3) most researchers follow different splitting methodologies to divide data into train, validate and test categories, and 4) they are not broadly accessible to the research community due to security and privacy reasons. This leads to experimental results that are not reproducible. Due to these issues, the use cases of cyber security do not have a standard approach and most enterprises avoid using ML/DL solutions for improving their cyber security applications [23].

The most recent method for improving a system's performance is to arrange the shared duties as a component of a conference and workshop. Shared tasks are contests where researchers or groups of researchers submit solutions to certain, predetermined problems. The training dataset is distributed among the participants as the first stage of the collaborative task. Evaluation of trained models is performed utilizing the testing dataset. Finally, the results are made publicly available with an option for publication. Shared tasks are most familiar in the field of NLP, computer vision and speech recognition. Recently, CDMC11 , IWSPA-AP12, DMD 201813, and AICS 201914 are 4 shared tasks in the field of cyber security.

3. Image/Signal Processing, Natural Language Processing, And Big Data For Cyber Security

A. Image and Signal Processing for Cyber security

For the purpose of classifying malware, a number of feature engineering techniques from the signal and image processing area have been effective. As an alternative to malware binaries represented in Hexadecimal or text forms, the malicious features are represented as a signal or grayscale image [115]. Each signal's range is [0, 255]. (0: black, 255: white). When it comes to images, the height can change depending on the file size while the width is fixed. Malware analysis based on signals and images is quick and does not need disassembling, unpacking, or running binary code. Recently, novel feature engineering methods such as spectral flatness, mel frequency cepstrum coefficients (MFCC), chroma features are proposed to accurately extract important features from signals and images. Current methods of signal and image-based malware detection exhibit two major problems: i) characterization of malware using signal and image based features does not give much information about the actual behavior of the malware and ii) since the approach relies on instance-based learning, its main limitation is that it can only detect or classify malware similar to what has already been observed. However, zero-day or new unseen malware attacks cannot be prevented. Hence, feature engineering mechanisms are employed by DL architectures to enhance the performance of malware analysis and detection.

B. Natural Language Processing for Cyber security

The goal of NLP is to simplify human-computer interaction by analysing and extracting information from natural languages. The availability of language in the form of data is essential for NLP success in the field of cyber security. In the field of cyber security, text data comes from a variety of sources, including emails, transaction logs from different systems, and online social networks. Utilizing NLP approaches has a direct impact on giving situational awareness based on user activity and other network event logs. Text can be encoded up to the word or letter level using a variety of methods, including vector space models, distributional representation, and distributed representation. As the first step of word/character level text encoding, preprocessing is followed by tokenization. This involves data cleaning and transformation of unnecessary and unknown words/characters, followed by word/character level tokenization. Non-sequential and sequential inputs are the two main types of text representation. Bag of words (BoW), term document matrices (TDM), and term frequency-inverse document frequency matrices (TFIDF) belong to non-sequential representation. N-gram, Keras embedding, Word2vec, Neural-Bag-of words, and FastText belong to sequential representation which has the capability to extract similarities in word meaning. In cyber security, capturing the sequential information is more important as compared to the similarities in word meaning due to the fact that most data contain time and spatial information. Hence, DL approaches could be adopted for an effective malware detection.

C. Big data Analytics for Cyber security

In order to safeguard data, computer systems, networks, and IoT from harmful behaviour, numerous cyber security applications require real-time analysis of BD created in CPS. To process and handle very large amounts of data effectively, due to the extraordinary rate at which unstructured and noisy data are being generated, breakthrough technologies in GPU and cluster computing frameworks are needed [22]. In order to store, process, and analyse data using ML approaches, such infrastructure is the key element in BD technologies. BD technologies are typically split into two categories: batch processing, like Hadoop, and stream processing, like InfoSphere. The Hadoop framework consists of Hadoop Distributed File System to store large files and MapReduce programming model to work on largescale data processing problems. Hadoop tools that adopt ML frameworks include Hive (an SQL-friendly query language), Pig (a platform and a scripting language for complex queries), Mahout and RHadoop. New frameworks such as Spark 4 are designed to improve the performance of DM and ML algorithms by repeated reuse of the working dataset. Hence, databases specifically designed for efficient storage and query of BD include NoSQL databases such as CouchDB, Cassandra, HBase, Greenplum Database, Vertica, and MongoDB. While batch processing has a dominant mature technology such as Hadoop, stream processing is still in its infancy to embrace ML/DL approaches. Complex Event Processing (CEP) is one of the models for stream processing where highlevel events are produced by aggregating and combining notification of events which are considered from the information flow. Storm, InfoSphere Streams, and Jubatus are few other implementations of stream technologies.

Due to its considerable dimensionality reduction in cyber security, the autoencoder (AE), a generative model, is thought to be a good technique for network traffic analysis. It learns the latent representation of various feature sets in an unsupervised manner. This is helpful with BD because it must quickly process big amounts of data without losing any information. Commonly used traditional methods for dimensionality reduction include singular value decomposition (SVD) and principal component analysis (PCA). The difficulty of managing large datasets and the demanding computational demands included in their processing are the primary obstacles restricting the advancement of AI and DL. Large datasets were traditionally handled by computer clusters equipped with specialised high-end CPUs. With a number of platforms specifically designed for the creation of ML and DL applications, the power of GPUs has recently been harnessed and has demonstrated to be faster and better for large-scale data processing. DL architectures are implemented on well-known platforms like TensorFlow, Theano, Torch, Caffe, and DeepDist. Table II lists the most common DL frameworks in alphabetical order. The majority of well-known systems employ C++ and have Python frontends while supporting high level ML computations in the backend. Additionally, Keras, a high-level API that supports TensorFlow, CNTK, and Theano, has grown in popularity.

References

[1] Farhat, Danyal, and Malik Shahzad Awan. "A brief survey on ransomware with the perspective of internet security threat reports." 2021 9th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2021.

[2] Khiralla, Fatma Abdalla Mabrouk. "Statistics of cybercrime from 2016 to the first half of 2020." Int. J. Comput. Sci. Netw. 9.5 (2020): 252-261.

[3] Axon, Louise, et al. "Data presentation in security operations centres: exploring the potential for sonification to enhance existing practice." Journal of Cybersecurity 6.1 (2020): tyaa004.

[4] Cave, Stephen. "The problem with intelligence: its value-laden history and the future of AI." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

[5] Chen, Jiefeng, et al. "Towards Evaluating the Robustness of Neural Networks Learned by Transduction." arXiv preprint arXiv:2110.14735 (2021).

[6] Du, Yi-Lun, et al. "Identifying the nature of the QCD transition in relativistic collision of heavy nuclei with deep learning." The European Physical Journal C 80.6 (2020): 1-17.

[7] Salloum, Said A., et al. "Machine learning and deep learning techniques for cybersecurity: a review." The International Conference on Artificial Intelligence and Computer Vision. Springer, Cham, 2020.

[8] Shafiq, Muhammad, et al. "Data mining and machine learning methods for sustainable smart cities traffic classification: A survey." Sustainable Cities and Society 60 (2020): 102177.

[9] Serinelli, Benedetto Marco, Anastasija Collen, and Niels Alexander Nijdam. "Training guidance with kdd cup 1999 and nsl-kdd data sets of anidinr: Anomaly-based network intrusion detection system." Procedia Computer Science 175 (2020): 560-565.

[10] Erlacher, Felix, and Falko Dressler. "On high-speed flow-based intrusion detection using snort-compatible signatures." IEEE Transactions on Dependable and Secure Computing (2020).

[11] Shamshirband, Shahab, et al. "Computational intelligence intrusion detection techniques in mobile cloud computing environments: Review, taxonomy, and open research issues." Journal of Information Security and Applications 55 (2020): 102582.

[12] Martínez Torres, Javier, Carla Iglesias Comesaña, and Paulino J. García-Nieto. "Machine learning techniques applied to cybersecurity." International Journal of Machine Learning and Cybernetics 10.10 (2019): 2823-2836.

[13] Salloum, Said A., et al. "Machine learning and deep learning techniques for cybersecurity: a review." The International Conference on Artificial Intelligence and Computer Vision. Springer, Cham, 2020.

[14] Ferrag, Mohamed Amine, et al. "Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study." Journal of Information Security and Applications 50 (2020): 102419.

[15] Apruzzese, Giovanni, et al. "On the effectiveness of machine and deep learning for cyber security." 2018 10th international conference on cyber Conflict (CyCon). IEEE, 2018.