ISSN: 2454-9940

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

E-Mail :
editor.ijasem@gmail.com
editor@ijasem.org

www.ijasem.org

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

DEEP LEARNING-DRIVEN CYBER THREAT DETECTION USING EVENT PROFILING

Dr P Rama Koteswara Rao[1], B.Kiran[2], V.Karthik[3],P.Pavankumar[4],K.Praveen kumar[5]

Professor, Department of Computer Science and Engineering[1]

Student, Department of Computer Science and Engineering[2,3,4,5]

Sree Dattha Institute of Engineering and Science, Sheriguda, Telangana. [1,2,3,4,5]

ABSTRACT

Among the most critical concerns in cyber security is the creation of an effective automated system for identifying cyber threats. Using neural networks, we provide an AI approach to cyber-threat detection in this research. The proposed method employs a detection algorithm based on deep learning to enhance cyber-threat identification by converting a flood of recorded security events into distinct event profiles. We developed an AI-SIEM system that employs event profiling for data preprocessing and several ANN algorithms, including FCNN, CNN, and LSTM, as part of our study. The primary goal of the system is to help security analysts differentiate between legitimate and false positive signals so that they can react more rapidly to cyber attacks. Every experiment is carried out by the authors utilising NSLKDD and CICIDS2017, two benchmark datasets in addition to two real-world datasets. We compared it to five popular machine-learning algorithms: support vector machine (SVM), kernel neural network (k-NN), random forest (RF), natural Bayes (NB), and decision tree (DT). In conclusion, the experimental findings show that our proposed techniques outperform existing machine-learning approaches, even when used in reality, and that they may be employed as learning-based models for network intrusion-detection.

Index Terms: Cyber-threat detection, deep learning, AI-SIEM, event profiling, neural networks, network intrusion detection.

I.INTRODUCTION

Several studies have demonstrated the effectiveness of learning-based techniques in cyber threat identification, bolstered by advancements in AI methodologies. However, with cyberattacks continually evolving, safeguarding IT systems from threats and malicious network behaviors remains a significant challenge. Given the prevalence of network breaches and harmful activities, establishing reliable solutions has underscored the importance of robust defenses and heightened security concerns [1, 2, 3, 4].

Historically, two primary approaches have been used for detecting cyber threats and network intrusions. Corporate Intrusion Prevention Systems (IPS) utilize signature-based methods to monitor network protocols and flows, aiming to prevent intrusions by identifying security events such as alarms triggered by potential intrusions. These systems typically communicate detected events to Security Information and Event Management (SIEM) systems, which focus on aggregating and processing IPS alarms. Security analysts play a critical role in assessing collected security events and logs, actively seeking suspicious signals based on predefined rules and thresholds, and correlating events to identify malicious activities [5].

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 2, 2024

Detecting and identifying breaches against sophisticated network attacks remains challenging due to the sheer volume of security data and the prevalence of false alarms [6, 7]. Consequently, recent research in intrusion detection has increasingly focused on machine learning and artificial intelligence systems. These advanced AI approaches enable security experts to rapidly and automatically detect network intrusions, albeit requiring prior training with historical threat data to identify new cyber threats effectively [8, 9].

Learning-based methods capable of autonomously detecting attacks within data are particularly valuable for swiftly assessing a large number of occurrences. According to [10], information security typically relies either on analyst-driven solutions or machine learning-based technologies that detect anomalous patterns. While analyst-driven approaches set foundational rules, machine learning technologies excel in identifying emerging cyber risks by detecting unexpected patterns [10].

Despite the effectiveness of learning-based approaches in detecting cyber attacks, several challenges persist. Firstly, acquiring labeled data essential for training and evaluating models in supervised learning systems is challenging, as many commercial SIEM systems do not store labeled data [10]. Secondly, the theoretical learning features often do not seamlessly integrate into real-world network security systems, complicating practical implementation [3]. To address these limitations, researchers have explored automating intrusion detection using deep learning techniques with benchmark datasets like NSLKDD [11], CICIDS2017 [12], and Kyoto-Honeypot [13], yet these datasets may not fully capture real-world complexities.

Thirdly, anomaly-based intrusion detection technologies, while effective in uncovering novel cyber threats, risk generating false alarms that demand extensive investigation efforts [6]. Moreover, hackers frequently adjust their tactics to evade detection, further challenging the efficacy of detection models [10], [14].

In response to these challenges, our initiative focuses on developing an AI-SIEM system leveraging deep learning techniques to differentiate between genuine threats and false alarms. This proposed strategy aims to enable security analysts to swiftly respond to distributed cyber threats across multiple security events by extracting patterns from data, particularly by linking clusters of events sharing contemporaneous attributes. By establishing event profiles, our system enhances the capability of deep neural networks to operate effectively with minimal data. Additionally, analysts can optimize their time by comparing current data against historical event records, thereby enhancing cyber threat detection and response capabilities over extended periods.

II.LITERATURE SURVEY

- Sheraz Naseer, Dr. Yasir Saleem, and Mr. Khawar proposed that due to the monumental growth of Internet applications in the last decade, the need for information network security has increased manifold. As a primary defense of network infrastructure, an intrusion detection system is expected to adapt to a dynamically changing threat landscape. Many supervised and unsupervised techniques have been devised by researchers in the fields of machine learning and data mining to achieve reliable detection of anomalies. Deep learning, an

area of machine learning that applies neuron-like structures for learning tasks, has profoundly changed the way we approach various tasks, delivering significant progress in disciplines like speech processing, computer vision, and natural language processing. It is only relevant that this new technology be investigated for information security applications. The aim of this paper is to investigate the suitability of deep learning approaches for anomaly-based intrusion detection systems. For this research, we developed anomaly detection models based on different deep neural network structures, including convolutional neural networks, autoencoders, and recurrent neural networks. These deep models were trained on the NSLKDD training dataset and evaluated on both test datasets provided by NSLKDD, namely NSLKDDTest+ and NSLKDDTest21. All experiments in this paper are performed by the authors on a GPU-based test bed. Conventional machine learning-based intrusion detection models were implemented using well-known classification techniques, including extreme learning machine, nearest neighbor, decision tree, random forest, support vector machine, naive bayes, and quadratic discriminant analysis. Both deep and conventional machine learning models were evaluated using well-known classification metrics, including receiver operating characteristics, area under the curve, precision-recall curve, mean average precision, and accuracy of classification. Experimental results of deep IDS models showed promising results for real-world applications in anomaly detection systems.

- Bang-Cheng Zhang, Guan-Yu Hu, and Zhi-Jie Zhou proposed that intrusion detection is very important for network situation awareness. While a few methods have been proposed to detect network intrusions, they cannot directly and effectively utilize semi-quantitative information consisting of expert knowledge and quantitative data. Hence, this paper proposes a new detection model based on a directed acyclic graph (DAG) and a belief rule base (BRB). In the proposed model, called DAG-BRB, the DAG is employed to construct a multi-layered BRB model that can avoid the explosion of combinations of rules due to a large number of intrusion types. To obtain the optimal parameters of the DAG-BRB model, an improved constraint covariance matrix adaptation evolution strategy (CMA-ES) is developed that can effectively solve the constraint problem in the BRB. A case study was used to test the efficiency of the proposed DAG-BRB. The results showed that, compared with other detection models, the DAG-BRB model has a higher detection rate and can be used in real networks.

- Ming Zhu, Yongzhong Huang, and Xiaozhou Ye proposed that the development of an anomaly-based intrusion detection system (IDS) is a primary research direction in the field of intrusion detection. An IDS learns normal and anomalous behavior by analyzing network traffic and can detect unknown and new attacks. However, the performance of an IDS is highly dependent on feature design, and designing a feature set that can accurately characterize network traffic is still an ongoing research issue. Anomaly-based IDSs also have the problem of a high false alarm rate (FAR), which seriously restricts their practical applications. In this paper, we propose a novel IDS called the hierarchical spatial-temporal features-based intrusion detection system (HAST-IDS), which first learns the low-level spatial features of network traffic using deep convolutional neural networks (CNNs) and then learns high-level temporal features using long short-term memory

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 2, 2024

networks. The entire process of feature learning is completed by the deep neural networks automatically; no feature engineering techniques are required. The automatically learned traffic features effectively reduce the FAR. The standard DARPA1998 and ISCX2012 datasets are used to evaluate the performance of the proposed system. The experimental results show that the HAST-IDS outperforms other published approaches in terms of accuracy, detection rate, and FAR, which successfully demonstrates its effectiveness in both feature learning and FAR reduction.

- Mohammed Khudhur Hussein, Nasharuddin Bin Zainal, and Aws Naser Jaber proposed that distributed computing has become an effective approach to enhance the capabilities of an institution or organization and minimize the need for additional resources. In this regard, distributed computing helps in broadening an institution's IT capabilities and is now an integral part of most expanding IT business sectors. It is considered a novel and efficient means for expanding business. As more organizations and individuals start to use the cloud to store their data and applications, significant concerns have developed to protect sensitive data from external and internal attacks over the Internet. Due to security concerns, many clients hesitate to relocate their sensitive data to the cloud despite significant interest in cloud-based computing. Security is a significant issue since much of an organization's data provides a tempting target for hackers, and these concerns will continue to hinder the development of distributed computing if not addressed. Therefore, this study presents a new test and insight into a honeypot, a device that can be classified into two types: handling and research honeypots. Handling honeypots are used to mitigate real-life dangers, while research honeypots are utilized as exploration tools to study and identify online threats. The primary aim of this research project is to conduct an intensive network security analysis through a virtualized honeypot for cloud servers to attract attackers and provide a new means of monitoring their behavior.

III. PREVIOUS WORK:

Cybersecurity has exploded in prominence as a contemporary security issue due to the proliferation of linked devices, the growth of computer networks, and the plethora of practical applications. Because of this, a solid intrusion detection system that can spot many forms of cybercrime and network anomalies is becoming more important. Although accurate, a lot of earlier studies used benchmark datasets that don't have enough data to be useful in the actual world. To get over these limitations, deployed learning models must be evaluated using real-world datasets. Thirdly, there is a risk of a high volume of false alarms when using anomaly-based intrusion detection technologies, even if they may help find cyber threats that were previously unknown.

IV. PROPOSED MODEL:

In this article, we provide a neural network–based artificial intelligence approach to cyber risk detection. The proposed method employs a detection algorithm based on deep learning to enhance cyber-threat identification by converting a flood of recorded security events into distinct event profiles. We developed an AI-SIEM system that employs event profiling for data preprocessing and several ANN algorithms, including FCNN, CNN, and LSTM, as part of our study. By separating genuine from false positive signals, the technology mainly aims to help security experts react swiftly to cyber attacks. we conducted experiments

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

using the five commonly used ML methods: decision tree, random forest, k-nearest neighbour, and support vector machine. Results from the study's experiments show that the approaches we proposed function as learning-based models for IDSs for networks.

V.SYSTEM ARCHITECTURE:

Information about the AI-powered SIEM system's architecture and workflow that has been meticulously planned. Data preprocessing, an AI-based learning engine, and real-time threat detection are the three main components of an AI-SIEM system. Gathering raw data into concise inputs for several deep neural networks is the main objective of the system's initial preprocessing stage, event profiling. In order, the AI-SIEM system performs the following operations: data preparation, parsing-based data aggregation, TF-IDF data normalisation, and event profiling. Next, event data sets, event vectors, and event profiles are all generated from the output, as shown in Figure. In a real-time intrusion detection system, this step comes before learning the data and before the raw security events are transformed into input data for the deep-learning engine.

In the second AI-driven learning engine, three ANNs are used for modelling. Using the preprocessed data, three ANNs are trained in the data learning stage to choose the best accurate model. Finally, each ANN model uses the trained models to automatically classify security raw events for real-time threat detection. Security analysts may decrease the number of false alerts by using the dashboard to present only the acknowledged legitimate ones.

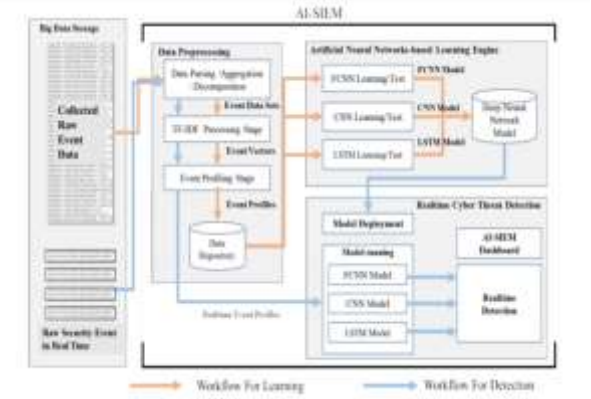


Figure.1 System Architecture Diagram

VI.MODULES DESCRIPTION:

Propose algorithms consists of following module

1) Data Parsing: This module take input dataset and parse that dataset to create a raw data event model
2) TF-IDF: using this module we will convert raw data into event vector which will contains normal and attack signatures
3) Event Profiling Stage: Processed data will be splitted into train and test model based on profiling events.
4) Deep Learning Neural Network Model: This module runs CNN and LSTM algorithms on train and test data and then generate a training model. Generated trained model will be applied on test data to calculate prediction score, Recall, Precision and FMeasure. Algorithm will learn perfectly will yield better accuracy result and that model will be selected to deploy on real system for attack detection.

Datasets which we are using for testing are of huge size and while building model it's going to out of memory error but kdd_train.csv dataset working perfectly but to run all algorithms it will take 5 to 10 minutes. You can test remaining datasets also by reducing its size or running it on

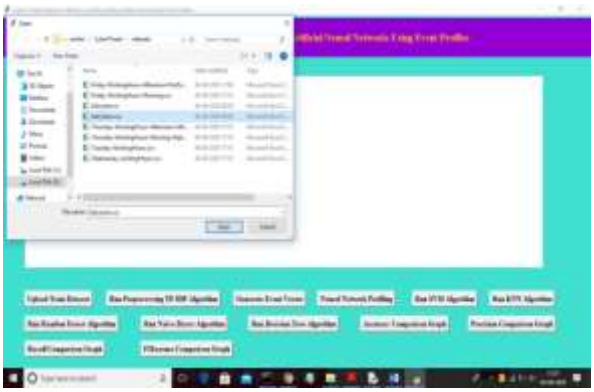
ISSN2454-9940
www.ijasem.org
Vol 18, Issue 2, 2024

high configuration system.

VIII.RESULTS

To run project double click on 'run.bat' file to get below screen

In above screen click on 'Upload Train Dataset' button and upload dataset

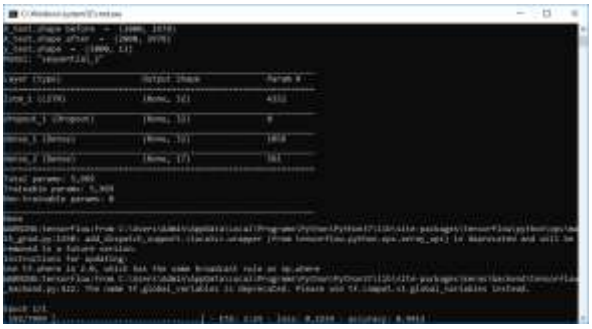


In above screen uploading 'kdd_train.csv' dataset and after upload will get below screen

In above screen we can see dataset contains 9999 records and now click on 'Run Preprocessing TF-IDF Algorithm' button to convert raw dataset into TF-IDF values

In above screen TF-IDF processing completed and now click on 'Generate Event Vector' button to create vector from TF-IDF with different events

In above screen we can see total different unique events names and in below we can see dataset total size and application using 80% dataset (7999 records) for training and using 20% dataset (2000 records) for testing. Now dataset train and test events model ready and now click on 'Neural Network Profiling' button to create LSTM and CNN model



In above screen LSTM model is generated and its epoch running also started and its starting accuracy is 0.94. Running for entire dataset may take time so wait till LSTM and CNN training process completed. Here

INTERNATIONAL JOURNAL OF APPLIED SCIENCE ENGINEERING AND MANAGEMENT

ISSN2454-9940

www.ijasem.org

Vol 18, Issue 2, 2024

dataset contains 7999 records and LSTM will iterate all records to filter and build model.

In above selected text we can see LSTM complete all iterations and in below lines we can see CNN model also starts execution

In above screen CNN also starts first iteration with accuracy as 0.72 and after completing all iterations 10 we got filtered improved accuracy as 0.99 and multiply by 100 will give us 99% accuracy. So CNN is giving better accuracy compare to LSTM and now see below GUI screen with all details

In above screen we can see both algorithms accuracy, precision, recall and FMeasure values. Now click on 'Run SVM Algorithm' button to run existing SVM algorithm

In above screen we can see SVM algorithm output values and now click on 'Run KNNAlgorithm' to run KNN algorithm

In above screen we can see KNN algorithm output values and now click on 'Run Random Forest Algorithm' to run Random Forest algorithm

In above screen we can see Random Forest algorithm output values and now click on 'Run Naïve Bayes Algorithm' to run Naïve

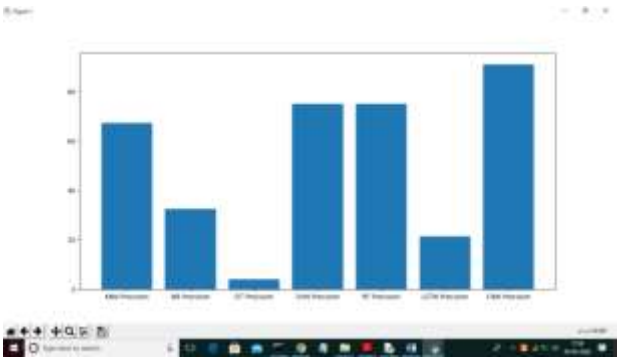
ISSN2454-9940
www.ijasem.org
Vol 18, Issue 2, 2024

Bayes algorithm

In above screen we can see Naïve Bayes algorithm output values and now click on 'Run Decision Tree Algorithm' to run Decision Tree Algorithm

Now click on 'Accuracy Comparison Graph' button to get accuracy of all algorithms

In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and from above graph we can conclude that LSTM and CNN perform well. Now click on Precision Comparison Graph' to get below graph



In above graph CNN is performing well and now click on 'Recall Comparison Graph'

In above graph LSTM is performing well and now click on FMeasure Comparison Graph button to get below graph

From all comparison graph we can see LSTM and CNN performing well with accuracy, recall and precision.

ISSN2454-9940
www.ijasem.org
Vol 18, Issue 2, 2024

IX.CONCLUSION

In this study, we introduced the AI-SIEM system, which makes use of event profiles and artificial neural networks. To enhance cyber-threat identification, our novel strategy compresses large datasets into event profiles and employs detection methods grounded in deep learning. The AI-SIEM technology enables analysts to efficiently manage critical security alerts by comparing past security data. In order to respond faster to cyber threats that are spread out across several security events, it is necessary to limit the number of false positive alerts. In order to evaluate performance, we evaluated results on two real-world datasets and two benchmark datasets (NSLKDD, CICIDS2017). In the first place, we compared our techniques to others using popular benchmark datasets, and the results showed that our learning-based model could successfully identify intrusions in networks. The second piece of good news is that we used two real datasets to show that our system outperformed conventional machine learning methods in terms of accurate classifications.

X.FUTURE ENHANCEMENT

In order to address the growing problem of cyber attacks, our future focus will be on improving early threat estimates by using several deep learning techniques to discover long-term patterns in historical data. Furthermore, in an effort to construct first-rate learning datasets and improve the supervised-learning dataset's accuracy, several SOC analysts will endeavour to record the labels of raw security events sequentially throughout the period of several months.

XI.REFERENCES

[1] S. Naseer, Y. Saleem, S. Khalid, M. K. Bashir, J. Han, M. M. Iqbal,K. Han, "Enhanced Network Anomaly Detection Based on DeepNeural Networks,"*IEEE Access*, vol. 6, pp. 48231-48246, 2018.

[2] B. Zhang, G. Hu, Z. Zhou, Y. Zhang, P. Qiao, L. Chang, "NetworkIntrusion Detection Based on Directed Acyclic Graph and Belief RuleBase", *ETRI Journal*, vol. 39, no. 4, pp. 592-604, Aug. 2017

[3] W. Wang, Y. Sheng and J. Wang, "HAST-IDS: Learning hierarchicalspatial-temporal features using deep neural networks to improveintrusion detection,"*IEEE Access*, vol. 6, no. 99, pp. 1792-1806,2018.

[4] M. K. Hussein, N. Bin Zainal and A. N. Jaber, "Data security analysisfor DDoS defense of cloud based networks,"*2015 IEEE StudentConference on Research and Development (SCOReD)*, KualaLumpur, 2015, pp. 305-310.

[5] S. Sandeep Sekharan, K. Kandasamy, "Profiling SIEM tools andcorrelation engines for security analytics,"*In Proc. Int. Conf.Wireless Com., Signal Proce. and Net.(WiSPNET)*, 2017, pp. 717-721.

[6] N.Hubballiand V.Suryanarayanan,‘‘False alarm minimizationtechniques in signature-based intrusion detection systems: Asurvey,’’ *Comput. Commun.*, vol. 49, pp. 1-17, Aug. 2014.

[7] A. Naser, M. A. Majid, M. F. Zolkipli and S. Anwar, "Trusting cloudcomputing for personal files,"*2014 International Conference onInformation and Communication Technology Convergence (ICTC)*,Busan, 2014, pp. 488-489.

[8] Y. Shen, E. Mariconti, P. Vervier, and Gianluca Stringhini, "Tiresias:Predicting Security Events Through Deep Learning,"*In Proc. ACMCCS 18*, Toronto, Canada, 2018, pp. 592-605.

[9] Kyle Soska and Nicolas Christin, "Automatically detectingvulnerable websites before they turn malicious,", *In Proc.*

USENIXSecurity Symposium., San Diego, CA, USA, 2014, pp.625-640.

[10] K. Veeramachaneni, I. Arnaldo, V. Korrapati, C. Bassias, K. Li,"AI2: training a big data machine to defend,"*In Proc. IEEEBigDataSecurity HPSC IDS*, New York, NY, USA, 2016, pp. 49-54.

[11] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu and Ali A. Ghorbani,"A detailed analysis of the kdd cup 99 data set,"*In Proc. of theSecond IEEE Int. Conf. Comp. Int. for Sec. and Def. App.*, pp. 53-58,2009.

[12] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, "Toward generatinga new intrusion detection dataset and intrusion trafficcharacterization", *Proc. Int. Conf. Inf. Syst. Secur. Privacy*, pp. 108-116, 2018.

[13][online]Available: http://www.takakura.com/Kyoto_data/

[14] N. Shone, T. N. Ngoc, V. D. Phai and Q. Shi, "A deep learningapproach to network intrusion detection,"*IEEE Trans. Emerg. TopicsComput. Intell.*, vol. 2, pp. 41-50, Feb. 2018

[15] R. Vinayakumar, Mamoun Alazab, K. P. Soman, P. Poornachandran,Ameer Al-Nemrat and Sitalakshmi Venkatraman, "Deep LearningApproach for Intelligent Intrusion Detection System,"*IEEE Access*,vol. 7, pp. 41525-41550, Apr. 2019.

[16] W. Hu, W. Hu, S. Maybank, "Adaboost-based algorithm for networkintrusion detection,"*IEEE Trans. Syst. Man B Cybern.*, vol. 38, no.2, pp. 577-583, Feb. 2008.

[17] T.-F. Yen et al., "Beehive: Large-scale log analysis for detectingsuspicious activity in enterprise networks", *Proc. 29th Annu.Comput. Security Appl. Conf.*, New York, NY, USA, 2013, pp. 199-208.

[18] K.-O. Detken, T Rix, C Kleiner, B Hellmann, L. Renners, "Siemapproach for a higher level of it security in enterprise networks", *InProc. IDAACS*, Warsaw, Poland, 2015, pp. 322-327.

[19] en.wikipedia.org,"Security information and event management,"2016 [Online] Available:https://en.wikipedia.org/wiki/Security-_information_and_event_management.

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-basedlearning applied to document recognition,"*Proc. IEEE*, vol. 86, no.11, pp. 2278-2324, Nov. 1998.