Cyber Security Intrusion Detection Using Deep Learning Approaches, Datasets, Bot-IOT Dataset

Iram Manan
Department of Information Technology
Khawaja Fareed University of Engineering and Information Technology
Rahim Yar Khan, Pakistan
iramrana019@gmail.com

Faisal Rehman
Department of Computer Science & Information Technology
Lahore Leads University
Lahore, Pakistan
faisalrehman0003@gmail.com

Hanan Sharif
Department of Computer Science & Information Technology
Lahore Leads University
Lahore, Pakistan
hanankhan386@gmail.com

Chaudhry Nouman Ali
Department of Computer Science & Information Technology
Lahore Leads University
Lahore, Pakistan
noumanali0322@gmail.com

Rana Rashid Ali
Department of Computer Science & Information Technology
Lahore Leads University
Lahore, Pakistan
rashidrana092@gmail.com

Amjad Liaqat
Department of Computer Science & Information Technology
Lahore Leads University
Lahore, Pakistan
4130417@gmail.com

2023 4th International Conference on Advancements in Computational Sciences (ICACS) | 978-1-6654-6377-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACS55311.2023.10089688

***Abstract*—Cyber Security is a crucial point of the current world; it is used to analyze, defend, and detect network intrusion systems. An intrusion detection system has been designed using Deep learning techniques, which helps the network user to detect malicious intentions. The dataset plays a crucial part in intrusion detection. As a result, we describe various well-known cyber datasets. Mainly we have analysed the IoT traffic-based dataset with some other datasets. We have also analyzed deep learning DL models, including Feed forward deep neural network (FDNN), deep auto-encoder, De-noising auto-encoder, deep migration, stacked de-noising auto-encoders, Replicator Neural networks, and Self-Taught Learning. We observe the effectiveness of models individually in two different types (multiclass and binary) through real-world traffic datasets, such as Bot-IoT dataset. Moreover, we evaluate the effectiveness of various methods based on the most critical key performance indicators, namely correctness, rate of false alarms and detection rate.**

Keywords—Cyber Security, Intrusion Detection, Deep Learning, Intrusion Detection, Bot-IoT dataset

I. INTRODUCTION

Neural network-based algorithms have proven helpful in detecting different inconsistencies, including such network intrusions, with the rapid development of computing systems. Intrusion detection and security operation are utilized to detect network inconsistency and attacks. Several detection systems have been developed that use machine learning (ML) [1] techniques to detect malicious intentions of users on the network. Intrusion datasets are critical for recognizing, validating, and testing appropriate measure detection methods. The efficiency of techniques for detecting anomalies is affected by the quality of data collected. Numerous datasets used mainly for intrusion detection seem to be freely available to the public.

As part of the system's secondary line of defence, an intrusion detection system (IDS) [2] is typically deployed. Security measures such as intrusion detection systems (IDS), network services, authentication procedures, and encryption methods are also necessary to properly defend systems from cyberattacks. An IDS may be able to switch between good and evil behaviour by analysing patterns of benign traffic, routine activities, or rules that characterise a particular attack [3]. In light of this, data mining [4], which is used to characterise recognition of understanding, may aid in the creation and deployment of IDS with increased correctness and robust actions as conventional IDS, which shall not be as successful as modern and complex cyber threats [5].

Furthermore, several academics have been challenged to locate extensive & accurate datasets to assess & analyze their suggested methodologies, which is a substantial difficulty in and of itself. To validate the effectiveness of such systems, reliable datasets that contain both benign and numerous attacks, satisfy real-world circumstances, and are accessible to the public are necessary [6] [7]. This paper describes our research as follow.

- Examine systems for detecting intrusions that employ deep learning techniques.
- Examine deep learning algorithms using 2 models: deep selective and un-supervised frameworks.
- Investigate the presence of a DL approach spending a novel practical traffic dataset, the Bot-IoT.
- They evaluate effectiveness of DL methods on the performance of four ML approaches [8]: "Naive Bayes (NV), SVM-Support Vector Machines, Artificial neural networks (RNN), and Random forests (RF)" [7].

978-1-6654-6377-5/23/$31.00 ©2023 IEEE

Other structured paper sections are given below: Section 2 summaries the literature review. Deep learning approach based on IDS is discussed in Section 3. Section 4 gives Open datasets with the Bot-IoT dataset, which has been used in DL approach articles for ID. Section 5 experiments, and Section 6 examines the performance matrix's effectiveness.

II. Literature Review

There are several relevant research in the literature that work with ML [9] algorithms for systems that detect intrusions. We classified the research based on the given criteria in Table 1:

- DL techniques: indicates whether or not the study focuses on Deep Learning methods for systems that detect intrusions.
- ML techniques: shows whether the study evaluated ML methods for intrusion detection.
- DL approach assessment: specifies whether or not the study investigates deep-learning methods for systems that detect intrusions.
- Machine learning approach evaluation: specifies whether or not the study assesses machine learning techniques for detecting intrusions.
- IDS datasets used: describe on condition that research engrossed on intrusion detection system datasets.

Characteristics of each can vary depending on the dataset used. For IoT traffic-based datasets, "these properties are categorised as Basic Knowledge, Assessment, Recording Surroundings, Volume Of data, and Data Structure" [10]. Intrusion detection systems make use of ML algorithms [11]. The datasets used in this investigation were classified as either "package," "NetFlow," or "open." Futhermore, the study provided an estimated computational time for every ML technique used in an ID system. Assess the relative merits of various IoT identification techniques [12]. Research categorises it according to its concentration on detection methods, the importance of IDS to its effectiveness, and the potential threat to security. Standard assessment parameters, such as workloads, metrics, and approach, were analysed by Milenkoski et al. [13] to reveal routines for cybersecurity intrusion detection. In our studies and four published publications [14], we focus on applying deep learning techniques to the problem of cyber security ID. However, several articles fail to offer a comparative study of DL algorithms applied to data sets. In the first study of its kind, ours is the first to examine deep learning (DL) methods, datasets, and a comparison study for systems that IDS for ID establish on DL approaches.

A. Feed Forward DNN

FFDNN [17] is used to filters feature selection strategy to design appropriate subgroups of characteristics with little repetition for wireless networks. The suggested IDS divided the primary training dataset into two parts. Then, a characteristic conversion and bidirectional normalising operation are executed.

III. Deep Learning Approaches-Based IDS

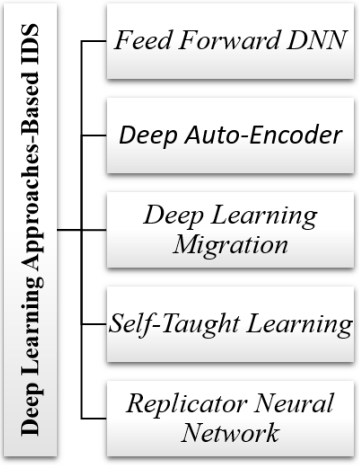

Fig. 1. Deep Learning Approache-Based ID

In conclusion, the FFDNN is used for both the testing and training of models in the proposed approach. Tools such as KDDTest+ and KDDTrain+, along with the NSL-KDD dataset, were used in the said study. Based on the evaluation of system performance, it was advised that 30 neurons be distributed across three hidden layers, and a learning rate of 0.05 would produce accurate results.

A. Deep Auto-Encoder

Identifying hackers using a deep autoencoder [18]. Using an auto-encoder with asymmetrical hidden layers, they surpass deep belief networks in classification. The study uses the NSL-KDD and KDD Cup '99 datasets together with five performance measures (correctness, exactness, memory, false alarm, and F-score). With an accuracy of 97.85 percent, the suggested model exceeds earlier attempts on the KDD Cup '99 dataset. With an average accuracy of 85.42 percent, the proposed model is 5% more accurate than the deep belief network model, according to the assessment of the NSL-KDD dataset.

The TSDL IDS is based on a two-phase DL model [20]. Input, hidden, and output are the three main layers of the TSDL model, which employs a stacked auto-encoder and a soft-max classifier. These three layers utilise a feedforward neural network, which is functionally equal to a multilayer perceptron, to accomplish this. The research employs the publicly available KDD99 and UNSW-NB15 datasets. Recognition rates of 99,996 percent accuracy have been reached for the KDD99 dataset. High identification rates of up to 89.134% are likewise reached in the UNSW-NB15 dataset's findings.

A self-learning and independent abuse IDS [21] is developed using autoencoder algorithms. Four steps (monitor, analysis, planning, performance, and expertise) make up the proposed system. Any event in which an intrusion detection system needs to adapt is detected in the monitoring phase. At this stage, we employ network audit technologies to interpret the data and network flows was converted from raw network traffic. The planning stage uses a sparse autoencoder because

that is what is needed for the input data. The data storage function is handled by the execute phase. Two datasets, NSL-KDD and KDDCup'99, are used in the paper for the evaluation. The results show that the fixed model is only 59.71% accurate, whereas adaptive model has 77.99% precision.

DNN and an upgraded conditional vibrational autoencoder to detect cyber security intrusions [22]. The suggested study is divided into three stages: training, attack generation, and attack detection. The training step contains optimizing the encoder and decoder losses. The distribution used in generating new attacks is a multivariate Gaussian distribution. DNN is used in the detection phase to detect attacks. The UNSW-NB15 and NSL-KDD datasets are employed to validate the suggested model, and the Adam optimist's default learning rate is 0.001. On the UNSW-NB15 dataset, the findings reveal the most fantastic accuracy of 89.08% and detection accuracy of 95.68%.

A noise removal autoencoder is used as a fundamental unit for cyber security ID to develop a deep neural network [23]. The de-noising auto encoder is utilized to recover partial feedback from intrusion detection systems. The performance evaluation uses the 1999 KDD Cup sets of data, and findings demonstrate which suggested model may reach sensing correctness of upwards over 95%. It employs stacked de-noising auto-encoders to detect fraudulent JavaScript code [24].

B. Deep Learning Migration

DL migration is used to detect cyber security intrusions. The research explicitly blends a DL approach with a system for intrusion detection. This research categorizes deep migration learning into four types: parameter migrating methodology, sample immigration methodology, associated knowledge immigration methodology, and feature extraction migration approach. The work employs the 1999 KDD Cup sets of data as experimental data added, with 10% of the learning dataset serving as test information. Ten thousand randomly chosen datasets are used as training sets in the experiment, and 10,000 randomly selected datasets are used as actual experimental sets. The results indicate a 91.05% detection accuracy and a 0.56% number of false alarms.

C. Self-Taught Learning

Its function is to identify threats to computer security. The proposed classification method includes steps like learning attributes presentation and using the previously learnt model for the categorization task. The NSL-KDD dataset is used for the evaluation of performance in this study. There are now three different classifications that make use of the current scheme: (1) two classes (normal and unusual), (2) five classes (normal plus four assault kinds), and (3) twenty-three classes. The study used 10-fold cross-validation on the data sets itself to assess the classification effectiveness of self-taught learning for these three categorization types. In the end, the f-measure value comes out to be 75.76 percent.

D. Replicator Neural Network

RNN models are used to identify cyber security intrusions. For abnormalities, detection investigation used the dropout technique. The entropy extraction consists of 3 phases: packet aggregation, dividing the flows into the time frame and Picking out interesting nodes along the flow. Incorporating the new simulated attacks into the performance analysis using the MAWI data sets is a great way to see how well the system is doing overall.

TABLE I. DL APPROACHES-BASED ON IDS

DL Methods	Datasets	Performance metrics
Feedforward DNN	NSL-KDD dataset	Correctness, Exactness, Memory
Deep auto-encoder		Correctness, Exactness, Memory, False Alarm, F-score
		Correctness, F-Measure, Precision, Recall
Deep auto-encoder	UNSW-NB15 dataset	Correctness, Exactness, Memory, F-measure, FAR
auto-encoder that denoises	1999 KDD Cup sets of data	Correctness, Check categorization fault
Deep learning migration		Detection rate, FAR, Exactness, Missing rate
Self-Taught Education		Correctness, Exactness, Memory, F-measure
Stacked de-noising auto-encoders	Heritrix dataset	Correctness, Categorization error, Exactness, Memory, F-measure
Replicator Neural Network	MAWI dataset	Anomalies Detection, Injected attacks Detection

IV. OPEN DATASETS

The table lists the fair representation of deep learning approaches articles that were reviewed for intrusion detection, such as the many times they were referenced and used dataset. Therefore, researchers use the data from the 1999 KDD Cup, the data from the UNSW-NB15 conference, and the data from the NSL-KDD conference. However, cyber security ID can also be used with other datasets.

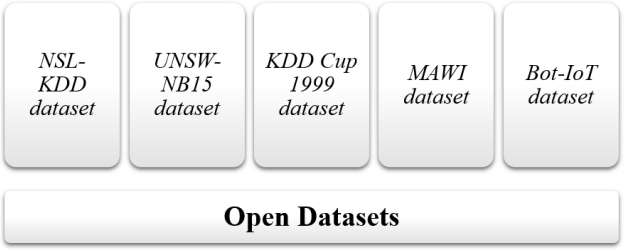

Fig. 2. Open Sets of Data

A. NSL-KDD dataset

This dataset was used to resolve the problems link in the 1999 KDD Cup sets of data. The NSLKDD sets of data improves on the original KDD sets of data in several ways. For example, no duplicated records, no duplication, and the amount of selected record keeping are organized as a percentage of the total value of records and appropriate amount of data. It is worth noting that several articles on ID use all sets of data to analyze performance.

B. NB15 dataset for UNSW

These data were collected using four different programmes: IXIA PerfectStorm, Tcpdump, Argus, and Bro-IDS. Assaults such as denial-of-services, exploits, generic attacks, reconnaissance, shellcode, and worms can be developed using these tools. The UNSW-NB15 dataset contains about 2 million vectors and 540,044 characteristics. This dataset also contains the test dataset (82,332 vectors) and the test sets (175,341 vectors).

C. 1999 KDD Cup Dataset

This dataset, which was derived from the DARPA'98 IDS assessment programme, has over 4,900,000 vectors culled from seven weeks' worth of network traffic. Remote-to-local (R2L), User-to-root (U2R), probing, and denial-of-service are the four categories of simulated assaults (DoS). The 41 characteristics present in the KDD Cup 1999 dataset may be categorised as follows. The first set of features pertains to traffic management, the second set to the essentials, and the third set to the content. A TCP/IP connection provides access to the system's fundamental features. Same-host or same-service criteria are used to classify the traffic. Some of the content's traits may suggest suspicious behaviour in the supporting data. For IDS systems, there is no better benchmark than this dataset.

D. MAWI dataset

These datasets document, in packet form, the daily traffic patterns on a connection between the United States and Japan across the Pacific. The MAWI dataset is useful for researching Internet traffic, anomaly detectors, and traffic classification techniques.

E. Bot-IoT dataset

There are around 72,000 records in this collection, and they include a wide variety of attacks like DDoS, DoS, OS and service scanning, keylogging, and data extraction. When compared to other datasets, the bot-IoT sets are completely new for the IoT infrastructure. For this purpose, the team used Node-red, a tool for modelling network activity in the Internet of Things. Datasets use the MQTT, a straightforward protocol for establishing M2M connections.

TABLE II. TYPES OF ATTACKS IN BOT-IOT DATASET

Type	Attacks	Training	Slide count	Analysis
Gathering	OS Fingerprinting	28,662	358275	7166
DDoS Attack	DDoS TCP	1,563,808	19,547,603	390,952
	DDoS UDP	1,517,208	18,965,106	379,302
	DDoS HTTP	1582	19771	395
BENIGN	BENIGN	7634	9543	1909
Information	Keylogging	1175	1469	294
DoS Attack	DoS TCP	985,280	12,315,997	246,320
	DoS UDP	1,652,759	20,659,491	413,190
	DoS HTTP	2376	29706	594
Theft	Data theft	94	118	24
Information	Service scanning	117,069	1,463,364	29,267

TABLE III. TRAINING TIME (TT) OF DDM WITH LEARNING IN BOT-IOT DATASET AND DIFFERENT HIDDEN NODES AND

Criterion		TT	DNN	RNN	CNN
LR	0.01	Time	56.5	70.7	65.3
LR	0.1	Time	66.6	92.6	91.3
LR	0.5	Time	88.1	102.5	101.1
LR	0.01	Time	88.1	102.5	101.1
LR	0.1	Time	102.2	150.4	144.2
LR	0.5	Time	170.3	222.1	221.7
LR	0.01	Time	250.8	331.2	339.6
LR	0.1	Time	302.9	377.1	366.2
LR	0.5	Time	391.1	451.2	412.2
LR	0.01	Time	600.2	801.5	812.2
LR	0.1	Time	711.9	1001.8	1022.1

TABLE IV. ACCURACY OF DEEP DISCRIMINATIVE MODELS WITH VARIOUS HIDDEN NODES AND LEARNING IN THE BOT-IOT DATASET

Criterion	Corrections		DNN	CNN	RNN
HN	ACC	15	96.446%	96.900%	96.765%
HN	ACC	15	96.651%	96.912%	96.882%
HN	ACC	15	96.651%	96.910%	96.884%
HN	ACC	30	96.611%	96.919%	96.877%
HN	ACC	30	96.655%	96.921%	96.882%
HN	ACC	30	96.661%	97.101%	96.898%
HN	ACC	60	96.766%	97.102%	96.955%
HN	ACC	60	96.922%	97.212%	96.974%
HN	ACC	60	97.102%	97.881%	97.291%
HN	ACC	100	97.221%	97.991%	97.618%
HN	ACC	100	97.501%	98.121%	97.991%
HN	ACC	100	98.221%	98.371%	98.311%

V. EXAMINATION

When conducting experiments, we make use of recently developed real-time traffic databases, more notably Bot-IoT data sets, which contain attack statistics gathered during the dataset's Training and Test phases. The study is conducted using Google's Colab platform. First, use a GPU and TensorFlow in Python. It shows how the IDS was used in the experiment in detail. Datasets, pre-processing, training, and testing are the many components of the procedure.

A. Data-set pre-processing

More than 72.0 million records may be found in the 74 papers that make up the Bot-IoT dataset, with 46 attributes per row. We use 5% of the full dataset for both our training and testing versions. In order to build a sample of testing and training files, we use Py-Mongo 3.7.2 to combine the documents into a single JSON file.

VI. PERFORMANCE METRICS

Confidential nodes are also displayed in Table II, along with the accuracy and learning rates for unsupervised model learning on the Bot-IoT dataset. Using 100 hidden nodes and a learning rate of 0.5, a convolutional neural network (CNN)

may attain an accuracy of 98.371%. The time of training for deep neural networks is also significantly less than that of similar methods in every case. An improved accuracy of 98.394% is attained by the deep auto encoders. In deep discriminative models, deep auto encoders perform better than three other methods.

CONCLUSION AND FUTURE DIRECTIONS

The article contrasts various DL approaches to intrusion detection, including deep racist and discriminatory models and unsupervised models. For this study, we focused on various forms of deep learning, including deep neural networks, deep auto-encoders, de-noising auto-encoders, deep migration, stacked de-noising auto-encoders, neural networks with replicators, and self-taught learning. The Bot-IoT dataset includes three critical performance indicators: false alarm rate, accuracy, and detection rate, and it is used to compare ML techniques to new datasets. To maximise accuracy and detection rates in future studies, we want to employ the most recent live dataset of IoT traffic in conjunction with a variety of DL models.

REFERENCES

[1] N. Riaz, S. I. A. Shah, F. Rehman, and M. J. Khan, "An intelligent hybrid scheme for identification of faults in industrial ball screw linear motion systems," IEEE Access, vol. 9, pp. 35136–35150, 2021.

[2] A. Ahmim, M. Derdour, and M. A. Ferrag, "An intrusion detection system based on combining probability predictions of a tree of classifiers: An intrusion detection system based on combining probability predictions of a tree of classifiers," Int. J. Commun. Syst., vol. 31, no. 9, p. e3547, 2018.

[3] A. Ahmim, L. Maglaras, M. A. Ferrag, M. Derdour, and H. Janicke, "A novel hierarchical intrusion detection system based on decision tree and rules-based models," in 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS), 2019.

[4] Z. Dewa and L. A., "Data mining and intrusion detection systems," Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 1, 2016.

[5] B. Stewart et al., "A novel intrusion detection mechanism for SCADA systems which automatically adapts to network topology changes," EAI endorsed trans. ind. netw. intell. syst., vol. 4, no. 10, p. 152155, 2017.

[6] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proceedings of the 4th International Conference on Information Systems Security and Privacy, 2018.

[7] M. A. Ferrag, L. A. Maglaras, H. Janicke, and R. Smith, "Deep learning techniques for cyber security intrusion detection : A detailed analysis," 2019.

[8] A. Ashfaq, M. Kamran, F. Rehman, N. Sarfaraz, H. U. Ilyas, and H. H. Riaz, "Role of artificial intelligence in renewable energy and its scope in future," in 2022 5th International Conference on Energy Conservation and Efficiency (ICECE), 2022.

[9] N. Riaz, S. I. A. Shah, F. Rehman, S. O. Gilani, and E. Udin, "A novel 2-D current signal-based residual learning with optimized softmax to identify faults in ball screw actuators," IEEE Access, vol. 8, pp. 115299–115313, 2020.

[10] M. Ring, S. Wunderlich, D. Scheuring, D. Landes, and A. Hotho, "A survey of network-based intrusion detection data sets," Comput. Secur., vol. 86, pp. 147–167, 2019.

[11] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Commun. Surv. Tutor., vol. 18, no. 2, pp. 1153–1176, 2016.

[12] B. B. Zarpelão, R. S. Miani, C. T. Kawakani, and S. C. de Alvarenga, "A survey of intrusion detection in Internet of Things," J. Netw. Comput. Appl., vol. 84, pp. 25–37, 2017.

[13] A. Milenkoski, M. Vieira, S. Kounev, A. Avritzer, and B. D. Payne, "Evaluating computer intrusion detection systems: A survey of common practices," ACM Comput. Surv., vol. 48, no. 1, pp. 1–41, 2015.

[14] D. S. Berman, A. L. Buczak, J. S. Chavis, C. L. Corbett, S. Mahdavifar, and A. A. Ghorbani, "Application of deep learning to cybersecurity: asurvey," information, vol. 10, no. 4, 2019.

[15] S. M. Kasongo and Y. Sun, "A deep learning method with filter based feature engineering for wireless intrusion detection system," IEEE Access, vol. 7, pp. 38597–38607, 2019.

[16] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi, "A deep learning approach to network intrusion detection," IEEE trans. emerg. top. comput. intell., vol. 2, no. 1, pp. 41–50, 2018.

[17] K. Alrawashdeh and C. Purdy, "Toward an online anomaly intrusion detection system based on deep learning," in 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), 2016.

[18] F. A. Khan, A. Gumaei, A. Derhab, and A. Hussain, Tsdl: a twostage deep learning model for efficient network intrusion detection. .

[19] D. Papamartzivanos, F. Gomez Marmol, and G. Kambourakis, "Introducing deep learning self-adaptive misuse network intrusion detection systems," IEEE Access, vol. 7, pp. 13546–13560, 2019.

[20] H. Sharif, F. Rehman, and A. Rida, "Deep Learning: Convolutional Neural Networks for Medical Image Analysis - A Quick Review," in 2022 2nd International Conference on Digital Futures and Transformative Technologies (ICoDT2), May 2022, pp. 1–4. doi: 10.1109/ICoDT255437.2022.9787469.

[21] A. Abusitta, M. Bellaiche, M. Dagenais, and T. Halabi, "A deep learning approach for proactive multi-cloud cooperative intrusion detection system," Future Gener. Comput. Syst., vol. 98, pp. 308–318, 2019.

[22] Y. Wang, W.-D. Cai, and P.-C. Wei, "A deep learning approach for detecting malicious JavaScript code: Using a deep learning approach to detect JavaScript-based attacks," Secur. Commun. Netw., vol. 9, no. 11, pp. 1520–1534, 2016.

[23] D. Li, L. Deng, M. Lee, and H. Wang, "IoT data feature extraction and intrusion detection system for smart cities based on deep migration learning," Int. J. Inf. Manage., vol. 49, pp. 533–545, 2019.

[24] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), 2016.