CDMC'19—the 10th International Cybersecurity Data Mining Competition

Shaoning Pang[1], Tao Ban[2], Youki Kadobayashi[3]
Jungsuk Song[4], Kaizhu Huang[5], Geongsen Poh[6], Iqbal Gondal[1]
Kitsuchart Pasupa[7] and Fadi Aloul[8]

[1]School of Science Engineering and Information Technology
Federation University Australia (Email: p.pang@federation.edu.au)
[2]The National Institute of Information and Communications Technology, Japan
[3]Nara Institue of Science and Technology, Japan
[4]Korea Institute of Science and Technology Information, Republic of Korea
[5]Xi'an Jiaotong-Liverpool University, China
[6]Malaysian Institute of Microelectronic Systems, Malaysia
[7]King Mongkut's Institute of Technology Ladkrabang, Thailand
[8]American University of Sharjah, United Arab Emirates

Abstract CDMC-International Cybersecurity Data Mining Competition [1] is a world unique data-analytic competition sitting in the transdisciplinary area of artificial intelligence and cybersecurity. In this paper, we summarize CDMC'19 — the 10th cybersecurity data mining competition, which was held in Sydney Australia — together with a coupled workshop event, the Artificial Intelligence and Cyber Security (AICS) workshop 2019. We introduce the scope and background of the CDMC competition, the competition organizer, International Cyber Security Data-mining Society (ICSDS), and the rules that we followed to manage the competition. We reveal details of CDMC'19 regarding the competition tasks, participating teams, and the results the participants have achieved. Moreover, we publish the collection of CDMC's 10-year competition datasets as the CDMC Cybersecurity Dataset Repository via `http://archive.csmining.org`. Finally, we conclude the paper with an outlook on the future activities of CDMC.

1 Introduction

Cybersecurity has become more and more a data-driven industry — data becomes both the goals and means of the all the activities in the cyber space. On the one hand, data in the form of digital assets and intellectual properties have grown into the most valuable resources for business and life and thus became the major targets of the cyber attacks. This not only gives rise to the ever-evolving attacks and record-breaking number of attack campaigns toward these data, but also brings chance to cyber security innovation for the digital economy, e.g., Data Loss Prevention (DLP) and Cloud Security. On the other hand, with the

[1] `http://www.csmining.org`

advance of digitization technology, e.g., Industry4.0, AI, 5G, software-defined networking, we are more than ever capable of recording the footprint of any cyber-attacks to enable corresponding defence operations. Hence, the solution to cybersecurity highly relies on systematic collection, management, analysis, interpretation and application of data.

Up to now, Artificial intelligence (AI) has been widely used in analyzing massive cyberspace data, creating so called cyber-threat intelligence to help security operations analysts to identify potential threats and thus stay a step ahead of big incidents. AI also helps to dramatically reduce the incident response time at security operation centers, by instantly extracting insights from the noise of thousands of daily threat alerts [1]. In this intelligence production process, various forms of advanced computational algorithms including statistical analyses, machine-learning algorithms, and deep-learning networks are engaged.

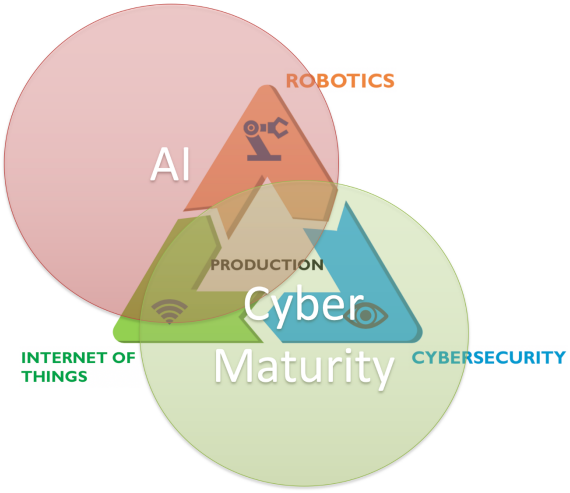

Figure 1: An illustration of balanced AI and cybersecurity development based on [6].

AI has been attracting investors, inventors, as well as academic researchers worldwide. In the past decade, the publication of AI-focused academic papers, has outpaced the amount of published researches on computer science; and the number of AI-focused startups backed by venture capital was more than doubled, outpacing the increase of the overall pool of startups. Yet as we employ more and more AI and automation technologies in our life and business, e.g. Internet of Thing (IoT) and robotics, they may introduce a potential new opening for electronic intruders. Cybersecurity must be carefully refactored before further application to more major but critical services. Towards a sustainable world through a smart digital transformation, we foster a balanced development scheme between AI and Cybersecurity, as shown in Figure 1, or namely, cyber maturity

of AI. The interdisciplinary field, "AI × Cyber Security" focus on researches to develop AI-enabled defense against increasingly sophisticated cyber attacks.

2 The Activities of ICSDS

In 2008, a remarkable collaboration between National Institute of Information and Communications Technology (NICT) Japan and Auckland University of Technology (AUT) New Zealand led to the formation of a dedicated international academic society in the trans-disciplinary area of computational intelligence and information security, i.e., the International Cyber Security Data-mining Society (ICSDS). ICSDS was founded in 2008, after ICONIP'08 in Auckland, under the leadership of Professor Paul S. Pang of AUT. Tao Ban of NICT, and Prof Youki Kadobayashi of Nara Institute of Science and Technology, Japan. So far, it has grown into a full-fledged research association with 15 governing board members in 10 regions: New Zealand, Japan, Korea, China, Malaysia, Thailand, Australia, United Arab Emirates, Singapore, and Canada. ICSDS seeks more active participation from researchers and professionals specially in the Asia Pacific region.

Aiming at promoting more active interactions of researchers, scientists, and industry professionals, ICSDS engaged in a variety of international research activities soon after it has started.

Since 2008, ICSDS has been hosting the International Workshop on Data Mining and Cybersecurity, which is reforged as the International Workshop on AI and Cybersecurity to incorporate most recent progresses from AI field after 2017. The purpose of AICS is to raise the awareness of cybersecurity, promote the potential of industrial applications, and give young researchers exposure to the key issues related to the topic and to ongoing works in this area. AICS provides a forum for researchers, security experts, engineers, and students to present latest research, share ideas, and discuss future directions in the fields of data mining, artificial intelligence, and cybersecurity. During the past years, we had AICS2010 in Sydney Australia, AICS2011 in Hangzhou China, AICS2012 in Doha Qatar, AICS2013 in Daegu South Korea, AICS2014 in Kuala Lumpur Malaysia, AICS2015 in Istanbul Turkey, AICS2016 in Kyoto Japan, AICS2017 in Guangzhou China, AICS2018 in Siem Reap, Cambodia, and AICS2019 in Sydney Australia respectively. This year's AICS2020 will be held on November 18-22 in Bangkok Thailand.

ICSDS started the CDMC competition since 2010, which turns out to be a popular competition on cybersecurity which is attractive for young researchers. Refer to more detail information, e.g., tasks, evaluation, and yearly statistics, in the next section.

In 2017, ICSDS newly launched the first AI x Cyber Security Summit (ACSS) as an ICSDS-leading international high-tech forum, which is featured as an engagement of academia, industry and venture capital. ACSS'17 was hosted by Xi'an Jiaotong-Liverpool University, in Suzhou, China, and ACSS'18 was hosted by Chongqing University of Science and Technology, in Chongqing, China.

ICSDS collaborates with other organizations and conferences to promote academic and technical activities within its scope of interests. These collaborators includes New Zealand Embassy Beijing, New Zealand Consulate ChengDu, ICONIP conferences, Asia Pacific Neural Network Society (APNNA), IEEE New Zealand Section, Europe Neural Network Society (ENNS), International Neural Network Society (INNS), etc. Over the years, the events hosted by ICSDS have gathered hundreds of researchers, scientists, and professionals who are working in the field of artificial intelligence and/or cybersecurity from more than 68 countries and regions.

This initiative provided funding and infrastructure to foster coordination of a society of international experts, and launch the first ever international Cybersecurity Data Mining Competition (CDMC) in 2010, which not only gathered hundreds of young researchers, but also brought together experts from around the world to facilitate collaboration and accelerate research progress. Since then, great progress has been made. In the following we review the 10-year activities of ICSDS society.

3 CDMC Annual Competition

The CDMC is a challenging, research and practice competition, focusing on application of knowledge discovery and computational intelligence techniques to address cyber security challenges in real world applications. The competition is open to worldwide research teams or individuals, particularly welcomes university students, undergraduate or postgraduate, in the field of data science, network engineering, cyber security, and artificial intelligence.

3.1 Past Statistics

Across the history of this competition series, a wide range of cyber security problems covering 10 different categories of challenges has been studied. In addition to that, a couple of pattern recognition tasks have also been included in the competition series. In total, over 30 original datasets acquired from industry or research experiments were used in the CDMC history. Over 1276 teams/individuals from 68 different countries have registered and/or participated the CDMC over the last 10 years. See Figure 2 for a record of yearly participation for CDMC.

Figure 3 illustrates the distribution of historical participants by country. As seen, more than half of participants came from the Asia Pacific countries. USA and Canada contributed to 9% of the participants. Europe also contributed to 10%, from which UK is the most active participation country. The remaining 28% participants scattered sparsely in the other part of the world. The participants included commercial ICT companies, universities, and research institutes. It's worth noting that some participants have participated more than one time. Table 1 gives a list of the institution of the winning teams for the annual competition.

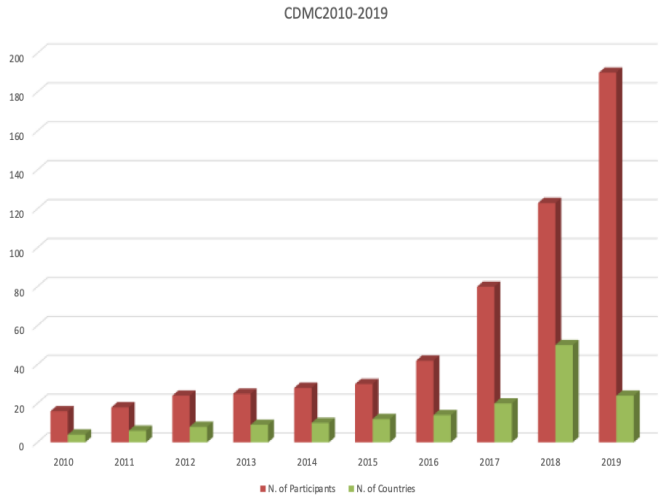

Figure 2: A summary of 10-year CDMC participation.

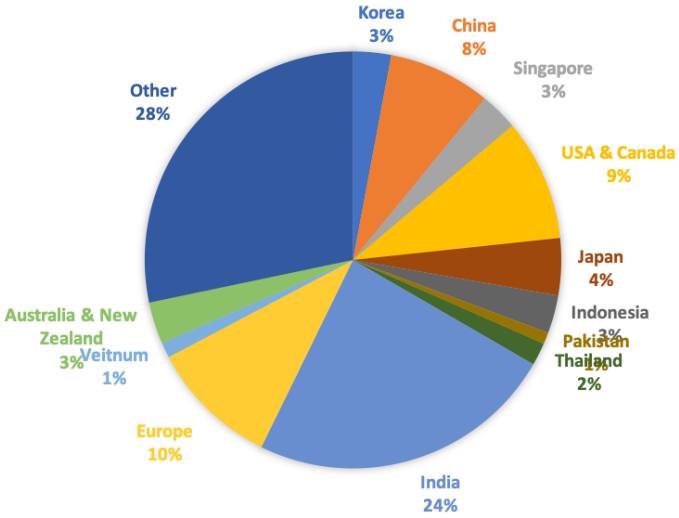

Figure 3: Distribution of participants by country.

Table 1: A list of institution of the winning teams for the annual CDMC competition.

Institution	Country	Year	Rank
Shandong Uni,	China	2010	1
Tongji Uni.	China	2011	1
Fujitsu R&D Center Co., Ltd.	Japan	2012	1
Uni. of Ottawa	Canada	2013	1
Inst. for Infocomm Research	Singapore	2014	1
Kyoto Women's University	Japan	2015	1
Austral University	Argentina	2016	1
Uni. of Edinburgh	UK	2016	2
Kyoto Women's University	Japan	2017	1
Uni. of Queensland	Australia	2017	2
National Uni. of Defense Tech.	China	2018	1
Washington Uni.	USA	2018	2
Nara Inst. of Sci. & Tech.	Japan	2019	1
Kle Technological University	India	2019	2

3.2 CDMC Cybersecurity Repository

CDMC has been addressing a number of specific cyber security challenges in the area of network security, IoT security, mobile security, social engineering, and hardware security, as well as a list of pattern recognition tasks towards leveraging AI techniques to meet industry needs. The acquired data range from numerical data, text, image, video/image series, structured & un-structured data, binary & multiple class data, as well as single & multi-task labelled data.

As the 10-years anniversary of CDMC, we publish in this paper the whole collection of CDMC datasets as the CDMC Cybersecurity Repository at `http://archive.csmining.org`. Table 2 gives the list of datasets in the repository. Note that a few tasks which are subject to specific restriction of publication after the competition are excluded from the repository.

3.3 Citation

If you publish your work based on datasets of this repository, you have to acknowledge the contributors of this repository. This will encourage other researchers to conduct a comparison study of different approaches on the same datasets and thus benefit your research as well. We suggest the following pseudo-APA reference format to cite this repository:

Pang, S. Ban, T. Kadobayashi, J. Song, Y. Huang, K. Gondal A. Poh, G. Pasupa, K. and Fadi, A. (2020). CDMC Cybersecurity Dataset Repository `http://archive.csmining.org`, International Cybersecurity Data-mining Society (ICSDS), hosted by the School of Engineering Information Technology and Physical Sciences, Federation University, Australia.

Note, a few datasets have additional citation requirements which can be found at the bottom of the dataset's web page.

4 CDMC 2019

Taking CDMC 2019 as an example, this section provides more detail information about the events hosted by ICSDS.

4.1 Competition Process

The steps of CDMC process are summarized as:

(a) **Obtaining the Tasks:** The competition tasks will be made available at the CDMC website `wwww.csmining.org` on the starting date of the competition. To enter the competition, all participants must register and download the assigned tasks at the official website.
(b) **Result Submission:** Submission of the results can be done using the submission form at the competition website. A valid submission should include the predicted results for the testing samples in plain .txt files, where the predictions need to be in the same definitions as the training datasets. In addition to result submission, participants are encouraged to submit a 2-8 page short paper to the AICS workshop.
(c) **Evaluation and Ranking:** The performance evaluation criteria include *p*recision, *r*ecall, F-measure, and *A*ccuracy, as defined in section 4.3. Note that while multiple submissions are allowed for one participant/team, only the last valid entry of result submissions will be used for performance evaluation and ranking. Here, a valid entry must include the results for all competition tasks.
(d) **Method Verification:** To prevent cheating, the top ranking teams will be required to fill out a fact sheet to describe their methods used for the competition. The ICSDS governing board will review the method and confirm the ranking. Note that participants might be required to provide their compiled software and/or source code for validation purpose.
(e) **Awarding and Prize Giving:** CDMC will announce 1st-place winner at the AICS workshop, which is collocated with the International Conference on Neural Information Processing (ICONIP). The awarding of cash prize and winner certificate will be at the banquet of the ICONIP conference. The CDMC cash prize is normally set as $3000NZD, the amount of which may be subject to the yearly sponsorship grant received by CDMC.

In tradition, CDMC awards only the 1st-place winner of the competition. The ICSDS governing board reserves the right to also present winner certificates to a short list of top ranking teams.

4.2 Competition Tasks

CDMC'19 came with three competition tasks which includes Task 1&2: SADAVS-sensor array data for autonomous vehicle safety, and Task 3: IoT malware classification [4].

1. SADAVS-Sensor Array Data for Autonomous Vehicle Safety [5]
   Vehicle-based accident detection systems monitor a network of sensors to determine if an accident has occurred. Instances of high acceleration/deceleration are due to a large change in velocity over a very short period of time. In the context of autonomous vehicles, the speeds are hard to attain since a vehicle is not controlled by a human driver. The presented data captured originally in New Zealand gives a collection of a sensor array (160x144) values in monitoring the status of moving vehicle. The objectives of these competition tasks are for early detection of any potential road accidents in two different scenarios.
2. IoT Malware Classification
   Based on the sequence of system calls as discriminant features and the malware families of the programs as training labels, the participants are required to perform a classification task to predict the malware families of the test samples. The dataset consists of 8442 samples generated following the procedure below: First, a collection of potentially malicious Linux programs in CEF format are collected from various sources. Then, each of these programs is executed in a sandbox environment hosted by an emulator that provides the required runtime environment for it. During the runtime, the *strace* command is used to monitor and record the interactions between the processes initialized by the program and the Linux kernel. This process yields a log file that contains lines of system calls. On each line, *strace* records the time stamp, the invoked system call, as long as parameters and results of the calls.

4.3 Performance Evaluation

The results of classification are first represented in a confusion matrix composed of *true positives*, *false positives*, *true negatives*, and *false negatives*. And then, to embrace competition tasks which have imbalanced class distribution, we perform performance evaluation using multiple criteria including *precision*, *recall*, *F-measure*, and *accuracy*.

Precision is a metric that calculates the accuracy for the minority class [3]. For an imbalanced binary classification problem, precision is calculated as the number of true positives divided by the total number of true positives and false positives:

$$Precision = \frac{TP}{(P + FP)}. \tag{1}$$

In an imbalanced classification problem with multiple classes, precision is calculated as the sum of true positives across all classes divided by the sum of true

Table 2: The list of datasets in CDMC Cybersecurity Repository

Category	Dataset Name/CDMC Year
Social Attacks	SNSSR/CDMC'19 e-News2016/CDMC'16 PSI/CDMC'14 e-News2013/CDMC'13 LingSparm /CDMC'10 e-News2015/CDMC'15
Sentiment Analysis	Trademe Sentiment/CDMC'15
DDoS Attacks	DDoS-ADENS/CDMC'18
IoT Malware	IoT-Malware/CDMC'19
Network Security	Packet Identification/CDMC'12
Intrusion Detection	IDS-Korea2014/CDMC'14 IDS-Korea2013/CDMC'13
Mobile Security	Android-API/CDMC'17 Android-Malware/CDMC'16
Cloud Security	UniteCloud-UTM/CDMC'17 UniteCloud-Log/CDMC'16
Physical System Security	SADAVS/CDMC'19
Financial Fraud	FDFT/CDMC'17
Pattern Recognition	AAP/CDMC'18 ESMC/CDMC'14 DMLI-MTPR/CDMC'13 5-Disease Diagnosis/CDMC'15

positives and false positives across all classes:

$$Precision = \sum_{c \in C} \frac{TP_c}{\sum_{c \in C}(TP_c + FP_c)}. \tag{2}$$

Recall is calculated for binary classification as the number of true positives divided by the total number of true positives and false negatives:

$$Precision = \frac{TP}{TP + FN}. \tag{3}$$

For multiple classes problem, recall is calculated as the sum of true positives across all classes divided by the sum of true positives and false negatives across all classes:

$$Precision = \sum_{c \in C} \frac{TP_c}{\sum_{c \in C}(TP_c + FN_c)}. \tag{4}$$

F-Measure combines precision and recall into a single measure that captures both properties, i.e.,

$$F - Measure = \frac{2Precision \times Recall}{Precision + Recall}. \tag{5}$$

Accuracy is calculated as the number of correctly classified instances divided by the total number of instances:

$$Accuracy = \frac{1}{\|C\|} \sum_{c \in C} \frac{TP_c + TN_c}{TP_c + TN_c + FP_c + FN_c}, \tag{6}$$

where $\|C\|$ is the total number of classes, TP_c+TN_c and $TP_c+TN_c+FP_c+FN_c$ are the correctly classified instances and the total number of instances of the ith class, respectively.

4.4 Results

CDMC'19 had total 190 teams/participants from 24 different countries. Compared to CDMC'18, the number of participants increased by 67, while the number of countries dropped from 50 to 24. Table 3 gives the top 10 teams and the results they had achieved.

Table 3: Top 10 teams of CDMC'19 and their results

Rank	Name	Institution	Country	Accuracy
1	Masataka Kawai	Nara Institute of Scienc and Technology	Japan	75.22%
2	Aditya Pandey	Kle Technological University	India	73.42%
3	Shivam Ralli	Kle Technological University	India	73.33%
4	Inzamam Sayyed	Nil	India	72.12%
5	Teoh John	University of Glasgow	United Kingdom	70.63%
6	Qianguang Lin	Hainan University	China	70.41%
7	Syukron Abu Ishaq Alfarozi	King Mongkut's Institute of Technology Ladkrabang	Thailand	69.52%
8	Vadim Borisov	University Tuebingen	Germany	68.93%
9	Binh Nguyen	University of Science, Ho Chi Minh City	Vietnam	66.94%
10	Yoshino Ozawa	Kyoto Women's University	Japan	64.07%

5 Conclusion

Data is driving force for both AI and cybersecurity. It offers opportunities for researchers to discover rules from the practices, for cyber security analysts to find ways to adopt new policies to fortify cyber resilience, and for AI practitioners to explore solutions to transform learned models to businesses. CDMC takes

advantage of such interdisciplinary insights, and has contributed to the society in providing a premium forum for discussion and exchange of these experiences.

In 2020, despite of the COVID-19 situation, the 11th CDMC, and the 13th AICS workshop have been set up. You are cordially invited to submit papers to the 13th International Workshop on Artificial Intelligence and Cybersecurity (AICS2020) and participate in the 11th International Cybersecurity Data Mining Competition (CDMC 2020). The two events are associated with the 27th International Conference on Neural Information Processing (ICONIP 2020) as a special session. ICONIP will be organized in Bangkok, Thailand, November 18-22, 2020. For more information about the CDMC2020, please refer to the competition website at `http://wwww.csmining.org`. We look forward to meeting you in Bangkok, Thailand.

Acoknowledgement

The authors would like to acknowledge all the participants who had ever take part in the competitions over the last 10 years. We would like to express our great appreciation to Auckland University of Technology, New Zealand, Unitec Institute of Science and Technology, New Zealand, and National Institute of Information and Communications Technology, Japan for their financial sponsorship to CDMC in the past 10 years, and to the Asia Pacific Neural Network Society (APNNS) for 10 years partnership in making CDMC a world known competition in the area of AI $\times$ Cybersecurity.

References

1. Aktayeva A., Niyazova R., Muradilova G., Makatov Y., Kusainova U.: Cognitive Computing Cybersecurity: Social Network Analysis. In: Sukhomlin V., Zubareva E. (eds) Convergent Cognitive Information Technologies. Convergent 2018. Communications in Computer and Information Science, vol. 1140, pp. 28–43. Springer, Cham (2020)
2. Sokolova M. and Lapalme G.: A systematic analysis of performance measures for classification tasks. Information Processing & Management, vol. 45, no. 4, pp. 427–437, (2009)
3. Fernandez A. Garcia S. Galar M. Prati R. C. Krawczyk B. and Herrera F.: Learning from Imbalanced Data Sets. Springer, (2018)
4. Pang, S. Ban, T. Kadobayashi, J. Song, Y. Huang, K. Gondal A. Poh, G. Pasupa, K. and Fadi, A. (2020). CDMC Cybersecurity Dataset Repository [http://archive.csmining.org/], International Cyber Security Data-mining Society (ICSDS), hosted by the Federation University Australia, School of Engineering Information Technology and Physical Sciences.
5. Pang S. and Huang Y.: Sensor Array Data for Autonomous Vehicle Incident Detection, the 10th International Cyber Security Data Mining Competition (CDMC2019), Unitec Institute of Technology, New Zealand, (2019)
6. `https://projects.tuni.fi/uploads/2019/03/8506038d-erf2019_trinity_cybersecurity_robotics_workshop_slides.pdf`