Proceedings of the 5th Joint Symposium on Computational Intelligence

Editors

Phayung Meesad

Kitsuchart Pasupa

September 7th, 2018

Faculty of Information Technology

King Mongkut's University of Technology North Bangkok

ISBN: 978-616-368-065-5

PREACE

The Joint Symposium Computational Intelligence (JSCI) is a biannual event, which was first organized in 2016. IEEE Computational Intelligence Society Thailand Chapter (IEEE-CIS Thailand) aims to support research students and young researchers to create a place enabling participants to share and discuss on their research prior to publish their works initiated the event. This event is open to all researchers who want to broaden their knowledge in the field of computational intelligence. It provides a very good opportunity to share and discuss your work with other researchers in the field of computational intelligence.

The 5th Joint Symposium on Computer Intelligence (JSCI5) was held on September 7th, 2018 at the Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand. JSCI5 main contributions are in the field of machine learning, natural language processing, time series analysis, signal processing, and image processing. Theses topic are among the trends of research and applications of computational intelligence. At the symposium, we had three invited professor talks. For regular papers, at least two members of the Technical Program Committee (TPC) reviewed all papers, in which a positive vote for presentation at the symposium and included in the proceedings published in electronic format.

We would like to thank all the authors and committee members for their excellent support and scientific collaboration. Finally, the authors, contributors and participants all made the JSCI again successful. We hope they find the common and fruitful debate of competitive thinking as well as new ideas for further development in the fields of computational intelligence and applications.

September 6th, 2018 *Phayung Meesad*
Bangkok *Kitsuchart Pasupa*

ORGANIZING COMMITTEE

Advisory

Chanboon Sathitwiriyawong, King Mongkut's Institute of Technology Ladkrabang

Chair

Phayung Meesad, King Mongkut's University of Technology North Bangkok

Technical Program Committee

Jonathan H. Chan, King Mongkut's University of Technology Thonburi

Kitsuchart Pasupa, King Mongkut's Institute of Technology Ladkrabang

Kiyota Hashimoto, Prince of Songkla University

Kuntpong Woraratpanya, King Mongkut's Institute of Technology Ladkrabang

Vithida Chongsuphajaisiddhi, King Mongkut's University of Technology Thonburi

Friday September 7th, 2018
King Mongkut's University of Technology North Bangkok
Faculty of Information Technology, 4th Floor, Room 4A07

Time	Activities	Topics/Speakers/Affiliations
08:30-08:50	Registration	
08:50-09:00	*Opening Speech* *Assoc.Prof.Dr. Phayung Meesad, Dean of Faculty of Information Technology, KMUTNB*	
09:00-09:10	Group Photograph	
09:10-10:10	Professor Keynote Speech 1	**Optimal Document Clustering with Feature Selection and Centroid Allocation** *Prof. Dr. Lance Chun Che Fung* *Murdoch University, Western Australia*
10:10-10:30	Student Speaker 1	**Development of an Anomaly Prediction System for Multivariate Time-Series from Sensor Data** *Thasorn Chalongvorachai and Kuntpong Woraratpanya* *King Mongkut's Institute of Technology Ladkrabang*
10:30-10:45	Coffee Break	
10:45-11:45	Professor Keynote Speech 2	**Communication in Decentralised Systems** *Prof.Dr.-Ing.habil. Herwig Unger* *FernUniversität in Hagen, Germany*
11:45-12:05	Student Speaker 2	**Facial Recognition Attendance Checker** *Thanet Prompinit, Pornchai Mongkolnam, Salisa Cheawcharnthong and Jonathan H. Chan* *King Mongkut's University of Technology Thonburi*
12:05-13:10	Lunch	
13:10-14:10	Professor Keynote Speech 3	**Advances in Biomedical Text Mining** *Assoc.Prof.Dr. Jonathan H. Chan* *King Mongkut's University of Technology Thonburi*
14:10-14:30	Student Speaker 3	**A Framework for Stock Selection Using Association Rules on Combined Cash Flow and Accrual Financial Indicators** *Amontep Wijitcharoen, Praisan Padungweang and Bunthit Watanapa* *King Mongkut's University of Technology Thonburi*
14:30-14:45	Coffee Break	
14:45-15:05	Student Speaker 4	**Essential Processes for Electroencephalography** *Srisakul Chanyalikhit and Kuntpong Woraratpanya* *King Mongkut's Institute of Technology Ladkrabang*
15:05-15:25	Student Speaker 5	**Incremental Learning with Deep GMDH Neural Network for Data Stream Mining** *Panida Lorwongtrakool and Phayung Meesad* *King Mongkut's University of Technology North Bangkok*
15:25-15:45	Student Speaker 6	**Database Creation by Natural Language Processing** *Chalermpol Tapsai, Phayung Meesad and Choochart Haruechaiyasak* *King Mongkut's University of Technology North Bangkok*
15:45-16:15	Student Speaker 7	**The Fatigue Monitoring System using the EEG Signals** *Worawut Yimyam and Mahasak Ketcham* *King Mongkut's University of Technology North Bangkok*
16:15-17:30	Closing/Committee Meeting	

TABLE OF CONTENTS

Keynote Speakers

Student Papers

Keynote Speaker

Optimal Document Clustering with Feature Selection and Centroid Allocation

Lance Chun Che FUNG
School of Engineering and Information Technology
Murdoch University
Murdoch, Western Australia
L.Fung@murdoch.edu.au

***Abstract*— Effective Document clustering system aims to improve the tasks of documents analysis, grouping, and retrieval. Its performance depends on documents preparation and allocation of centroids in the clusters. Optimal document clustering is a combinatorial NP-hard optimization problem and it becomes necessary to utilize non-traditional methods to look for optimal or near optimal solutions. This research investigated supervised and unsupervised feature selection methods as well as two centroid allocation methods to improve the document clustering process.**

BIOGRAPHY

Professor Emeritus Lance C.C. Fung received his technical training as a Marine Radio and Electronic Officer from the Hong Kong Polytechnic and Brunel Technical College, Bristol UK in 1972-74 and 1975-76 respectively. After serving on the high seas, he completed a Bachelor of Science Degree in Maritime Technology with First Class Honors in 1981, and a Master of Engineering in System Test Technology in the 1982 from the University of Wales, Institute of Science and Technology (UWIST). Subsequently, he received his PhD degree from the University of Western Australia (UWA) in 1994 with a thesis on the Application of Artificial Intelligence to problems in Electrical Power System Engineering. Lance has lectured at the Singapore Polytechnic (1982-1988), Curtin University (1989-2003) and Murdoch University (2003-2015). He received his Emeritus Professor appointment which enables him to continue his academic engagement since retirement in 2015. In September, 2017, he received an Honorary PhD Degree in Information Technology from Walailak University, Thailand, in recognition of his contributions towards the development and advancement of the University's research and postgraduate programs. Lance has been an active IEEE volunteer for over two decades, served various positions in many committees, boards and conferences. He is the current Chair of the IEEE Conference Quality Committee (CQC) and the IEEE Technical Program Integrity Committee (TPIC); Chair-Elect of the IEEE New Initiatives Committee (NIC) and Chair of the IEEE SMC Society Chapter Coordinator Committee. Lance has published over 325 articles in academic journals, conference proceedings and book chapters. He has also supervised to completion over 30 doctoral and higher degree candidates. His research interest is in the development and applications of innovative intelligent technologies and advanced techniques to solve practical problems. His passion is to nurture postgraduate research students and he continues dedicating his time to supervise postgraduate students locally and abroad.

978-616-368-065-5

Keynote Speaker

Communication in Decentralised Systems

Herwig Unger and Mario Kubek
Chair of Communication Networks
FernUniversität in Hagen, Universitätsstr
Hagen, Germany
{herwig.unger, mario.kubek}@fernuni-hagen.de

***Abstract*— Modern applications as they are used for instance within Industry 4.0 systems rather build on peer-to-peer than client-server-based system architectures and functional principles. The highly dynamic character of those systems requires many new configuration and maintenance activities. New communication principles based on random walkers and new kinds of group communication are the probate tools to satisfy the respective requirements. After a discussion of some innovative communication principles in the beginning of the talk, it is intended to show on two examples, how self-organising, adaptive and flexible structures can be emerge and how they can be used for a variety of tasks.**

BIOGRAPHY

Prof. Dr.-Ing. habil. Herwig Unger received his PhD with a work on Petri Net transformation in 1994 from the Technical University of Ilmenau and his doctorate (habilitation) with a work on large distributed systems from the University of Rostock in 2000. Since 2006, he is a full professor at the FernUniversität in Hagen and the head of the Chair of Communication Networks. His research interests are in decentralised systems and self-organization, natural language processing, as well as large scale simulations. He has published more than 140 publications in refereed journals and conferences, published or edited more than 25 books and gave over 35 invited talks and lectures in 12 countries.

Dr.-Ing. Mario Kubek is a researcher at the Chair of Communication Networks of the FernUniversität in Hagen. He received his PhD in 2012 with a thesis on locally working agents to improve the search for web documents. His research focus is on natural language processing, text mining and semantic information retrieval in large distributed systems. His further research interests include topic and trend detection in diachronic text corpora and contextual information processing in mobile computing environments.

978-616-368-065-5

Keynote Speaker

Advances in Biomedical Text Mining

Jonathan H. Chan
Data Science and Engineering Laboratory (D-Lab)
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok, Thailand
jonathan@sit.kmutt.ac.th

***Abstract*— Biomedical text mining, generally referred to as BioNLP, is the application of data mining and machine learning techniques to extract useful information and knowledge in the biomedical as well as molecular biology literature. As the number of biological and biomedical publications are increasing exponential in PubMed and other indices/repositories, it is paramount to develop better natural language processing (NLP) tools and techniques, especially in conjunction with biological resources. This talk will provide an overview on the development of BioNLP and some recent advances and applications in this field with an emphasis on data mining and machine learning.**

BIOGRAPHY

Dr. Jonathan H. Chan is an Associate Professor at the School of Information Technology, King Mongkut's University of Technology Thonburi (KMUTT), Thailand. Jonathan holds a B.A.Sc., M.A.Sc., and Ph.D. degree from the University of Toronto and was a visiting professor there back in 2007, 2009 and 2016; he was also a visiting scientist at The Centre for Applied Genomics at Sick Kids Hospital in Toronto in several occasions. Dr. Chan is a member of the editorial board of Neural Networks (Elsevier), Heliyon (Elsevier), International Journal of Machine Intelligence and Sensory Signal Processing (Inderscience), International Journal of Swarm Intelligence (Inderscience), and Proceedings in Adaptation, Learning and Optimization (Springer). Also, he is a reviewer for a number of refereed international journals including Information Sciences, Applied Soft Computing, Neural Networks, BMC Bioinformatics, and Memetic Computing. He has also served on the program, technical and/or advisory committees for numerous major international conferences. Dr. Chan has organized/co-organized many international conferences, and he is the Past-President of the former Asia Pacific Neural Network Assembly (APNNA) and a Governing Board member of the current Asia Pacific Neural Network Society (APNNS). He is a senior member of IEEE and INNS, and a member of ACM and the Professional Engineers of Ontario (PEO). His current research interests are in the interdisciplinary field of data science, including but not limited to intelligent systems, biomedical informatics, and systems biology.

978-616-368-065-5 ©2018 JSCI

Development of an Anomaly Prediction System for Multivariate Time-Series from Sensor Data

Thasorn Chalongvorachai
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
58070047@kmitl.ac.th

Kuntpong Woraratpanya
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand
kuntpong@it.kmitl.ac.th

***Abstract* – A long short-term memory recurrent neural network (LSTM-RNN) is applied for an anomaly detection in multivariate time-series from sensor data. This paper extends its ability to an anomaly prediction by means of a fuzzy logic technique for data modification. The experimental results give us satisfaction in a low root-mean-squared error (RMSE) and the anomaly events are predicted.**

Index Terms – LSTM-RNN, Anomaly prediction, Time-series data, Sensor data, Multivariate time-series data

I. Introduction

In the past, industrial sensors were installed in machinery to detect anomaly events or malfunctions and then alarm to engineers, technicians, or workers who were responsible to those problems. The problem-solving was based on human intelligence. Nowadays, artificial intelligence (AI) becomes a new wave technology of predictive diagnostics in current industries. It can detect anomaly incidents before the actual event happens in the future. One of the existing works on anomaly detection in time-series data is the use of a long short-term memory recurrent neural network (LSTM-RNN) [1]. This paper applied the recurrent neural network (RNN) and LSTM-RNN to various time-series datasets for an anomaly detection performance comparison. Nevertheless, the solution of this paper is in the form of detection, not a prediction. In other words, this approach can only detect abnormal incidents but cannot forecast anomaly events which may happen in the future. Detection is insufficient for anomaly prevention monitoring in industries that require online learning for real-time prediction and updating.

As mentioned above, it becomes a challenge to extend the ability of LSTM-RNN for multivariate time-series sensor data-sets. Responding to this challenge, we apply the fuzzy logic technique to the dataset by providing a suitable membership function for event labelling. This makes the labeled event possible to forecast and notified anomaly event in advance.

II. Methodology

This section describes the proposed method as schematically shown in Fig. 1. In the scheme, a real sensor multivariate time-series dataset is modified with a fuzzy logic method to make the model more suitable with the problem-solving. Then the LSTM-RNN is used to train the modified dataset. An overall framework can be visualized as shown in Fig. 1.

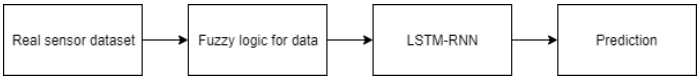

Fig. 1 An overall framework of an anomaly prediction system for multivariate time-series from sensor data.

A. Modification of a real sensor dataset with fuzzy logic

A real dataset is collected from the sensors installed in machinery, especially the motor used for a cooling system in the power plants. This dataset is a multivariate time-series platform that contains 15 features from 15 sensors classified into 9 temperature sensors, 2 vibration sensors, 2 electricity power sensors, 1 electricity current sensor, and 1 humidity sensor. This dataset gives us detected values from each sensor and time ranges of anomaly events when the cooling system starts working abnormally. It contains one anomaly event occurred from a complete breakdown of machinery and two abnormal events occurred from system shutdown by engineers before the actual abnormal event happens.

According to the previous approach [1], the algorithm can only detect anomaly events, but cannot forecast abnormal incidents in advance. In order to make the algorithm predictive, the fuzzy logic technique is applied to label the dataset. With the fuzzy logic labelling, data can indicate the increase of anomaly incidents sign before actual event occurs in regression form, so prediction and notification of the abnormal event in advance are easily detected. The event label is defined by membership function in (1).

$$f(x) = \begin{cases} 0 & ,x < a \quad or \quad x > c \\ \frac{x-a}{b-a} & ,a \le x \le b \\ 1 & ,b \le x \le c \end{cases} \tag{1}$$

where x is the current time in time-series data, a is the beginning of anomaly event, b is the starting point of the absolute abnormal event, and c is the ending point of the anomalous incident as shown in Fig. 2.

As a result, the machinery status is labeled with 0 if the machine works normally. On the other hand, the machinery status is labeled with 1 if the machine is breakdown.

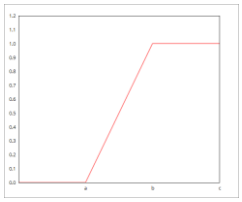

Fig. 2 A membership function for event labelling.

In order to validate our approach, we modified data when the sign of breakdown appeared with an interval between 0 and 1 by giving additional constant values per minutes.

B. *Long Short-Term Memory – Recurrent Neural Network*

The concept of a feed-forward neural network is the state-of-art solution for artificial intelligence prediction in various works. However, the feed-forward neural network has limitation. It can work only with non-sequential data, but cannot work well with sequential data such as time-series. In 1990, Elman proposed a recurrent neural network (RNN) solution [2]. To summarize the concept of this solution, he used the output as input again in the hidden state. This would make a neural network able to learn with sequential data. However, this approach still has a vanishing gradient problem. This means that the longer data will not be updated by the model. In 1997, Hochreiter proposed long-short term memory recurrent neural network [3]. They claimed that his model is more immune to the vanishing gradient problem. The concept behind architecture is using a cell state and four gates for calculation, including input gate, forget gate, update gate, and output gate. These gates are for updating the next hidden state to eliminate the vanishing gradient and return effective output.

III. Experiment

In the experiment, inputs for training and testing are variables from 15 real sensor data with event labeling. The data are divided into 62.5% for training and 37.5% for testing. Our models are developed with Python language and Keras library and are tested on Notebook with processor Intel(R) Core(TM) i7-4720HQ CPU 2.60 GHz, 2601 MHz, 4 Core(s), 8 Logical Processor(s). After training and testing with different units, batch size and layers; the results of root mean squared error are compared to each model as reported in Table 1.

TABLE I: Performance comparisons of test models.

Model	LSTM-RNN Units*	Epoch	Batch size	RMSE	Computational Time (Second)
a	(100)	100	100	0.02758	716
b	(100-100)	100	100	0.00886	1571
c	(100)	100	50	0.06096	1560
d	(100-100)	100	50	0.00536	2722
e	(100-100)	200	50	0.01849	3112
f	(200)	100	50	0.03055	3451
g	(200-100)	100	50	0.01520	5150
h	(200-200)	100	50	0.01636	5886

*(200-100) can be interpreted as 200 units in the first hidden layer and 100 units in the second hidden layer.

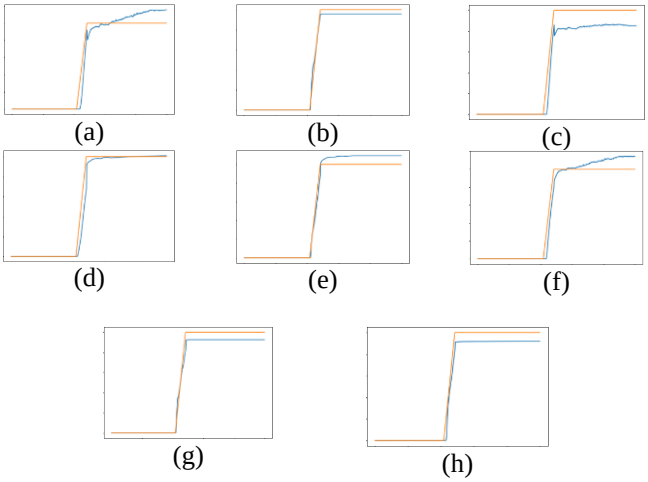

Fig. 3 Comparisons of actual abnormal data (orange) and predictions results from models (blue).

In prediction task, we show the visualization of a normal event compared with an abnormal event as shown in Fig. 3.

According to models (b) and (d), with a low RMSE and similar result of a prediction line compared with a test line, we could say that this model returns satisfaction results.

IV. Conclusion

In this study, we introduce an anomaly prediction method for multivariate time-series sensor data. A LSTM-RNN architecture in conjunction with a fuzzy-logic technique is implemented for anomaly event prediction. The results are satisfactory. However, this work still requires a diversity of abnormal event datasets for training and testing in order to improve the model. In future work, we are going to improve our model by modifying another architecture neural network such as GRU-RNN. We also eager to develop the online learning method by LSTM-RNN for prediction and updating in real-time.

References

[1] P. Malhotra, L. Vig, G. Shroff, P. , Long Short Term Memory Networks for Anomaly Detection in Time Series, Agarwal, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN, 2015, pp. 89-94

[2] J. L. Elman, Finding Structure in Time, Cognitive Science, University of California, 1990, pp.179-211.

[3] S. Hochreiter and J. Schmidhuber, Long Short-Term Memory, Neural Computation, 1997, pp. 1735-1780.

Facial Recognition Attendance Checker

Thanet Prompinit, Salisa Cheawcharnthong, Pornchai Mongkolnam and Jonathan H. Chan
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok 10140, Thailand
thanet.prom@mail.kmutt.ac.th, salisa_cct@hotmail.com, pornchai@sit.kmutt.ac.th, jonathan@sit.kmutt.ac.th

***Abstract*— This work presents a novel facial recognition framework for attendance checking, by using the student's smartphone and Bluetooth-low-energy (BLE) beacons placed in the classroom. Facial recognition is used for authentication based on the Active Appearance Model (68-point facial landmarks) that is converted to a 128-dimensional vector space. One or more beacons are used to track the sitting position of each student. Then image processing is used to analyze the current facial sentiment of the student. Facial recognition models based on Naïve Bayes, Support Vector Machine, and Random Forest were compared. The best model was found to be Random Forest with an accuracy of about 98% on the test data. A web-based application has been deployed and tested on Android smartphones to connect with the API for the needed services.**

I. Introduction

According to a recent (2016) statistical survey, only about 3 percent of the Thai population, or just fewer than 2 million students, pursued higher education [1,2]. However, the global approaches to higher education learning have been evolving rapidly with advances in information technology, leveraging the use of videos, graphics and animation, text messages, as well as social networking applications. For example, the traditional e-learning or distance-based education has evolved in form of a more comprehensive Massive Open Online Course (MOOC); such offering is available for anyone with Internet access for free, with commercial entities offering certification for a small fee [3]. Even with the rising popularity of MOOCs, online learning in general lacks the physical interactive element found in a traditional classroom. Currently, in most Thai universities, attendance is still mandatory. Nonetheless, one can argue that it is a burden for the instructors to manually keep students' attendance, even with the use of attendance checking devices such as fingerprint reader or radio frequency identification (RFID) card reader. Also, students still do not have timely access to their attendance record. In order to address the above issues, an application was developed as a proof of concept based on Android smartphone, beacon(s) and a rudimentary facial recognition procedure for efficient classroom attendance checking.

The aim of this work is to assist both students and instructors to seamlessly keep attendance and participation records to improve and facilitate the classroom learning environment, with the added bonus of reducing the instructors' burden in keeping the attendance. One or more beacons with Bluetooth Low Energy (BLE) are used to triangulate each student's position during the class. The instructor can also initiate an authentication protocol at any given time, if he/she wishes, to make sure that a student is present in class. In addition, more subtle analyses such as facial sentiments, individual and group sitting positions, can be used to help educators understand their students' behaviors in classrooms better.

II. Methodology

Our proposed SMATCH: Smart Classroom Attendance Checker system comprises a web-based API facial recognition server, a Firebase server, and an Android smartphone with Internet access and connected to a beacon [4]. One or more beacons are placed inside the classroom for the purpose of attendance checking. At least one beacon is needed but the uncertainty can be up to one meter. Thus we have tested up to three beacons in order to improve precision of student position triangulation. For the purpose of analytics though, an approximate sitting position that depicts the front of the classroom, middle area, or back of the classroom may be adequate. Each student is required to register his or her phone and number with our system. This process could be initiated before a semester begins. When the student is coming to class and within the beacon signal coverage area, the attendance check-in and sitting position data are recorded to the online Firebase server. An instructor can perform a facial check-in at a specific time or at a random interval. For example, whenever there is a quiz or in-class exercise, students need to check their attendance with facial images taken by the smartphones. When initiated, the instructor's smartphone sends a push notification to the students' smartphones, which asks for a facial check-in in order to authenticate the genuinely enrolled students, not someone else. In facial recognition, our method is based on the Active Appearance Model (AAM) and some well-known data mining techniques for classifying students' faces. The system's facial recognition framework is shown in Fig. 1.

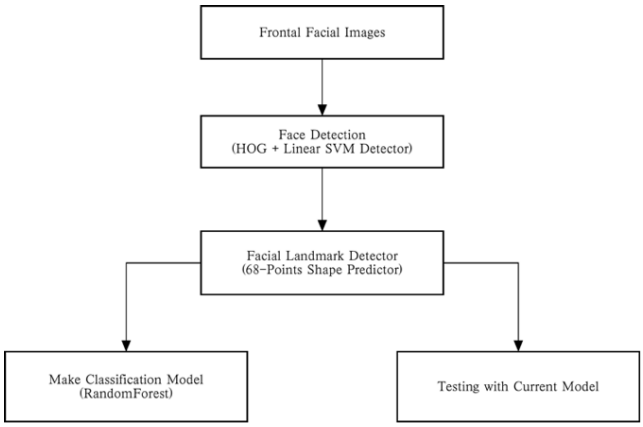

Fig. 1. Framework of basic facial recognition system

A more detailed description about the framework of the facial recognition system is provided as follows.

A. Frontal Facial Images

When students use the system for the first time, they must take 5 self-facial images for making a recognition model at the positions of center, right, left, up, and down. The next time the students would only need to use a single self-facial

image for comparison with the existing pre-trained model created earlier.

B. *Face Detection*

When the system gets an image from students. The system makes detection and crops just a facial area from whole image by using HOG + Linear SVM Detector techniques from the DLib Library.

C. *Facial Landmark Detector*

In this step, the system detects 68-point facial landmarks on the face image and then extracts these data for classification in the next step.

D. *Classification Model*

When the system receives facial landmark data for the first time, data mining techniques are used for making a model of classification of each student for authentication purpose. For the next time of usage, we would have unseen data for authentication purpose.

For the facial recognition process, Python was used to develop a web-based API server, with sending and receiving messages using the HTTP protocol, and the data were stored in the Java Script Object Notation (JSON) format. This process may be termed a RESTful web service [5]. The Python web framework known as Flask [6] was used. For the feature extraction task in the facial recognition framework, we used DLib [7] for both face detection and face landmark, and then we used the NumPy module [8] to convert the 68-point face landmarks to 128-dimensional data. The popular WEKA data mining software [9] was used for classification purpose. A connector between Python and WEKA was developed for the facial recognition process. Fig. 2 shows the schematic of the facial recognition system.

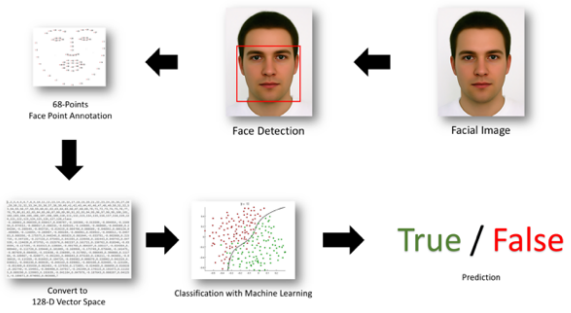

Fig. 2. Schematic of facial recognition system.

III. Results

In training a facial dataset, we applied the AAM technique used for feature extractions in order to obtain the 68-point face landmarks and converting them to 128 dimensions for subsequent classification [10,11]. For the dataset, we used 5 correct face samples and 5 incorrect face samples, for each of the 10 test subjects, and that resulted in a total of 100 sample faces. The samples were obtained from the publicly available extended Cohn-Kanade face database [12]. The dataset was trained with three common classifiers: Naïve Bayes, Support Vector Machine, and Random Forest. Table 1 shows comparisons of the three classifiers in terms of the accuracy of prediction. As can be seen from the results, Random Forest performed better than the other classifiers, for both training and testing data (shown in bold). The testing dataset was randomly generated from the same face database (not part of those 100 sample faces), which included 5 correct facial images and 5 incorrect facial images for each person, resulting in another total of 100 faces used for model testing.

Classifier	Training	Testing
Naïve Bayes	99%	68%
Support Vector Machine	76%	96%
Random Forest	**100%**	**98%**

Table 1. A comparison of the accuracy of different classifiers

References

[1] Office of The Higher Education Commission, Thailand. (2016). *Total Student Information.* Retrieved April 6, 2017, from Office of The Higher Education Commission, Thailand: http://www.info.mua.go.th/information/download.php?file_id=201706301624.xlsx&stat_id=344&id_member=.

[2] The Bereau of Registration Administration (BORA), Department of Provincial Administration, Thailand. (2016). *Number of Population in The Kingdom, Following by The Evidence of Registration*. Retrieved July 20, 2017, from The Bereau of Registration Administration (BORA), Department of Provincial Administration, Thailand: http://stat.bora.dopa.go.th/stat/y_stat59.htm.

[3] Glušac, D., Karuović, D., Milanov, D., Gluac, D., Karuovi, D., & Milanov, D. (2015). Massive open online courses - Pedagogical overview. *Carpathian Control Conference (ICCC), 2015 16th International*, (vi), 142–146. https://doi.org/10.1109/CarpathianCC.2015.7145063.

[4] Prompinit, T., Mongkolnam, P., Cheawcharnthong, S., & Chan, J. H. (2018). SMATCH: Smart Classroom Attendance Checker Using Beacon, Smartphone and Data Mining Techniques. *International Conference on Learning Innovation in Science and Technology (ICLIST 2018)*, March 21-24, 2018, Hua Hin, Thailand, pp. 1-8.

[5] Richardson, L., & Ruby, S. (2008). *RESTful Web Services*. https://doi.org/0596554605.

[6] Lewandowski, C. M. (2015). *Flask Web Development. The effects of brief mindfulness intervention on acute pain experience: An examination of individual difference* (Vol. 1). https://doi.org/10.1017/CBO9781107415324.004.

[7] King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research*, *10*, 1755–1758. https://doi.org/10.1145/1577069.1755843.

[8] Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*(2), 22–30. https://doi.org/10.1109/MCSE.2011.37.

[9] Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I. H., & Trigg, L. (2005). WEKA: A Machine Learning Workbench for Data Mining. *Springer US. In: Data Mining and Knowledge Discovery Handbook*, 1305–14. https://doi.org/10.1007/0-387-25465-X_62.

[10] Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 681–685. https://doi.org/10.1007/BFb0054760.

[11] Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Review of FaceNet : A Unified Embedding for Face Recognition and Clustering. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–2. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Schroff_FaceNet_A_Unified_2015_CVPR_paper.pdf (Hess, 2017).

[12] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., & Matthews, I. (2010). The extended Cohn-Kanade dataset (CK+): A complete facial expression dataset for action unit and emotion specified expression. *Cvprw*, (July), 94–101. https://doi.org/10.1109/ISIEA.2010.567950

A Framework for Stock Selection Using Association Rules on Combined Cash Flow and Accrual Financial Indicators

Amontep Wijitcharoen*, Bunthit Watanapa†, and Praisan Padungweang‡
School of Information Technology
King Mongkut's University of Technology Thonburi
Bangkok 10140 Thailand
Email: *amontep.wj@mail.kmutt.ac.th, †bunthit@sit.kmutt.ac.th, ‡praisan.pad@sit.kmutt.ac.th,

***Abstract*—Traditionally, decision makers use accrual-based financial indicators, which are dependent on historical information, to evaluate stocks. This passive information can provide only partial intelligence to the decision makers. Additional cash-based insight would be needed for more complete information to identify any stock that could beat the stock market trend. In this study, we propose a systematic framework for financial information mining that actively considers the interaction between traditional financial ratios and cash flow signaling in identifying potential outperforming stocks in the coming period. Association detection method can help to ensure the mined rules be reliable and practical. To demonstrate the performance of the mining rules, we plan to collect the financial statement of selected stocks in the Service Industry from the Stock Exchange of Thailand during 2014 to 2016 for inducing a set of rules. The accuracy test of the model may then be performed based the latest actual outcome in 2017.**

Keywords— Association rule, stock selection, cash flow, financial analysis

I. INTRODUCTION

Suitable stock selection plays an important role in financial portfolio construction. Selecting the right mixture of stocks at the right time could generate a satisfactory return to the investor. Over the recent decades, financial analysis is becoming a popular tool for analyzing stocks using financial indicators such as Return on Equity (ROE), Return on Assets (ROA), Price-to-Earnings ratio (P/E), Earnings per Share (EPS), and so on [1]. The drawback of using such accrual financial indicators only is the inability to detect cash-based signaling which complements to provide a fuller picture of a company's operation performance [2]. For example, when a company sold a product, in accrual accounting, it could mean the company already take a profit even though the customer still does not pay the money. As a complementary technique, cash flow (CF) has been used to give insight into the liquidity of the company in a given period of time [3,4]. In this study, we propose a framework to combine the traditional financial ratios and the cash flow measures to identify effective rules that can effectively suggest stocks in a selected industry that could generate a satisfactory return rate.

II. METHOD

The proposed framework is shown in Fig.1. It breaks down the relevant tasks into four sequential steps. First, a decision maker (DM) has to extract and transform all the financial indicators and cash flow into usable forms of codes or measures for the rules mining purpose. For example, the price of stock would be transformed as % change in price or the return of stock. The available traditional financial indicators used in this study are as listed in Table I. They can be categorized into five performance measures of profitability, liquidity, leveraging, asset utilization and growth. Cash flow manipulated in this study are operating cash flow, investing cash flow and free cash flow. We shall exclude those stocks with incomplete data for ensuring the validity of the mining process.

Second, both financial ratios and cash flow are absolute numerical value for an individual firm. Thus, prior to the mining process, these measures must be encoded and normalized for compatibility according to the defined schemes shown in Table II. The encoding is required for meaningful interpretation of data, e.g. for the dependent data, the outperforming stock is coded as '1' if its return is larger than the % change of the market index, otherwise '0'. The normalization is to standardize the cash flow data so that the difference in size and scale of the companies is mitigated, e.g. the operating CF is considered in terms of % of revenue.

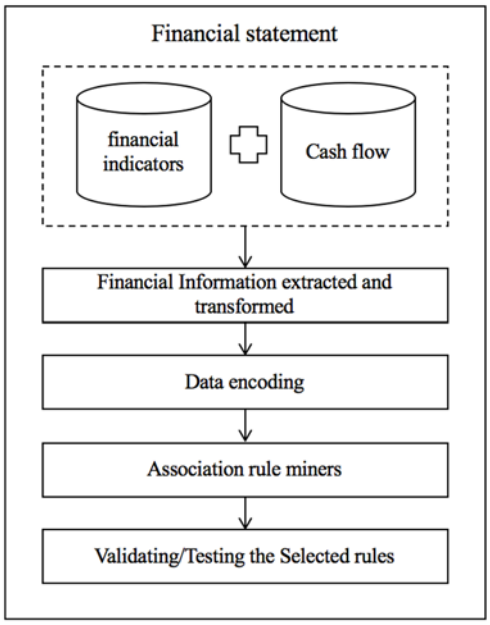

Fig.1 The proposed framework for stock portfolio selection.

Third, we use the coded values obtained from the previous step to identify the combinations of financial ratios and cash flows which are associated with the outperformance of stocks. The desired level of reliability of the discovered rules should be heuristically identified by investors via the tuned values of Support, Confidence, and Lift parameters. [5].

Last, the validation of the rule should be performed by testing the usability of the obtained rule(s). A metric of accuracy when applying rules to the real and unseen set of the financial information should be reported.

TABLE I. LIST OF TRADITIONAL FINANCIAL RATIOS USED

Performance measures	Financial Ratios
Asset utilization	Receivable turnover
	Asset turnover
Liquidity	Current Ratio
	Quick Ratio
	Debt to equity ratio
Leverage	Leverage ratio
	Solvency Ratio-I
Profitability	Return on equity
	Return on assets
	Net profit margin
	Earnings per Share
Growth	Revenue growth rate

TABLE II. BINARY CODING INSTRUCTION FOR ALL VARIABLES: 1 INDICATES THE PRESENCE OF A CHARACTERISTIC AND 0 INDICATES THE ABSENCE.

	Variables	Coding	
	Dependent variable:		
1	Outperforming stock	If the stock's return at the end of the year greater than the percent change of Market index: coded as 1	If not: 0
	Independent variables:		
	Financial indicators		
1	Receivable turnover (RT)	If the RT greater than the average of the sector: coded as 1	If not: 0
2	Asset turnover (AT)	If the AT greater than the average of the sector: coded as 1	If not: 0
3	Current Ratio (CR)	If the CR greater than the average of the sector: coded as 1	If not: 0
4	Quick Ratio (QR)	If the QR greater than the average of the sector: coded as 1	If not: 0
5	Debt to equity ratio (DE)	If the DE is within the acceptable level (< +1 S.D.): coded as 1	If not: 0
6	Leverage ratio (LR)	If the LR greater than the average of the sector: coded as 1	If not: 0
7	Return on equity (ROE)	If the ROE greater than the average of the sector: coded as 1	If not: 0
8	Return on assets (ROA)	If the ROA greater than the average of the sector: coded as 1	If not: 0
9	Net profit margin (NP)	If the NP greater than the average of the sector: coded as 1	If not: 0
10	Price Earning (P/E)	If the P/E greater than the average of the sector: coded as 1	If not: 0
	Cash flows		
1	*Operating cash flow/Revenue (OCF/R)*	If the OCF/R greater than the average of the sector: coded as 1	If not: 0
2	*Investing cash flow/revenue (IFC/R)*	If the ICF/R lower than the average of the sector: coded as 1	If not: 0
3	*Free cash flow/Revenue (FCF/R)*	If the FCF/R greater than the average of the sector: coded as 1	If not: 0

III. AN INSIGHT INTO EXPERIMENTAL DESIGN

For validating the model in this framework, the data set could be collected from the public data source such as Yahoo Finance, Morningstar, or The Stock Exchange of Thailand (SET). We plan to use the historical financial statements of the Service industry of SET during 2014 to 2016 for mining rules and will validate the obtained set of rules by selecting stocks in 2017.

The Service industry is one of the big three industry groups in the Thai stock market, consisting of five sectors namely Commerce, Health, Media and Publish, Tourism and Transportation. The comparative accuracy of the proposed system could be calculated from the number of the selected stocks that have an annual return greater than that of the SET.

IV. CONCLUSION AND FUTURE WORK

The integration of cash flow and the financial indicators can hypothetically fill the gap of using only accrual-based analytic information to select stocks, in order to outperform the average return of the stock market. A novel framework is proposed to mine these aforementioned information using Association rule method, so that investors may gain insight into the key features of a company's performance in a particular industry and generate effective rules for selecting stocks.

For future work, we plan to set up an experiment for discovering rules using Apriori algorithm using real data set acquired from the Service industry in SET. The concept of Apriori that prunes infrequent itemsets regarding the Support enables efficient mining and is extensible for dealing with Confidence and Lift [6]. The validation of the model in terms of rules accuracy will then be performed and the results will be disseminated accordingly.

REFERENCES

[1] Beaver, W. H. Financial ratios as predictors of failure. Journal of Accounting Research, 71-111, 1966.

[2] Barua, Suborna, and Anup Kumar Saha. "Traditional Ratios vs. Cash Flow based Ratios: Which One is Better Performance Indicator?." *Advances in Economics and Business* 3.6 (2015): 232-251.

[3] Sloan, R., 1996. Do stock prices fully reflect information in accruals and cash flows about future earnings? The Accounting Review 71, 289–315.

[4] Lee, T. 'Laker Airways - The Cash Flow Truth', Accountancy, 115-116, 1982.

[5] Agrawal, A., Mannila, H., Srikant, R., Toivonen, H. and Verkamo, A. Fast Discovery of Association Rules. In: Fayyad et al; 1996. p. 307–328.

[6] Mandave, P., Mane, M. and Patil, S., 2013. Data mining using Association rule based on APRIORI algorithm and improved approach with illustration. *International Journal of Latest Trends in Engineering and Technology (IJLTET), ISSN.*

Essential Processes for Electroencephalography

Srisakul Chanyalikhit
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand, 10520
srisakul.tat@gmail.com

Kuntpong Woraratpanya
Faculty of Information Technology
King Mongkut's Institute of Technology Ladkrabang
Bangkok, Thailand, 10520
kuntpong@it.kmitl.ac.th

***Abstract*—Electroencephalography (EEG) is important for studying brain activities and detecting brain disorders. This paper presents an overview of essential processes for EEG signal processing, including data acquisition, preprocessing, feature extraction, and classification.**

Keywords—Electroencephalography, Wavelet transform

I. INTRODUCTION

Human brain is the most important organ which consists of billion neurons. These neurons create small electrical signals to communicate between brain and body. The electroencephalography (EEG) signal [1] is non-invasive electrical signals from the scalp of the brain. EEG signals can be used to detect brain activity disorders such as Alzheimer's disease, Epilepsy, sleep disorder, etc. The EEG signals are collected from multiple channels, which record all brain activity signals. To detect a disorder or an activity of the brain, the EEG signals are eliminated any interference and then are transformed to the appropriate frequency range by means of the following essential processes: data acquisition, preprocessing, feature extraction, and classification as shown in Fig. 1.

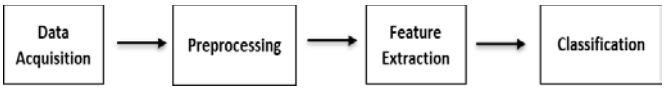

Fig. 1. EEG signal processing for classification

II. DATA ACQUISITION

To collect brain activity signals, it requires the specific device and the understanding of the proper use of it.

A. EEG Device

Nowadays, EEG signals can be collected by a small headset device which is portable and easy to use. The EEG device has 4 main components – electrodes, conductive gels, amplifiers, and analog to digital converter. There are many neuroheadset devices from many companies, such as Mind-Wave from NeuroSky, Muse from InteraXon, and EPOC from Emotiv.

B. 10-20 System

Each part of the brain controls different functions of the body. For example, the left frontal lobe of the brain is res-ponsible for speech and language [1]. Therefore, the loca-tions of the electrodes are important for collecting the acti-vity of the brain.

The electrodes are used to conduct the electrical signals from the scalp of the brain. The 10-20 System is a method to describe the position of the electrodes. The '10' and '20' refer to the distance between adjacent electrodes that are either 10% or 20% from front to back or right to left of the skull [2]. Each electrode placement is represented by the letters (F-Frontal, T-Temporal, C-Center, P-Parietal, O-Occipital, z-midline region) and the digits (odd number-left brain, even number-right brain) which indicate the position of the brain lobe as shown in Fig. 2.

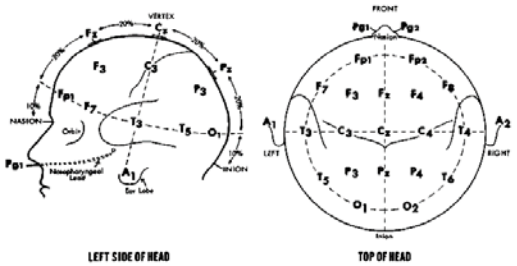

Fig. 2. 10-20 System for locating the electrodes

III. PREPROCESSING

Preprocessing is a procedure for eliminating the unnecessary information and noise.

According to [1], most of EEG signals are in range 0.1–100 Hz. EEG data filtering depends on frequency band of each brain activity. For example, the muscles movement and eyes movement require 2.5 ~ 25 Hz bandpass filter.

IV. FEATURE EXTRACTION

Feature extraction is one of the most important parts of all processes. The better feature extraction gives the better result of classification. There are many options for feature extraction such as Fourier transform and wavelet transform. Fourier transform such as Short Time Fourier transform cannot work for the multi-resolution signals, while wavelet transform can.

Wavelet transform is one of the most popular methods for EEG signal feature extraction. There are two forms of wavelet transform for feature extraction: continuous wavelet transform (CWT) and discrete wavelet transform (DWT) [3].

A. Continous Wavelet Transform

Continuous wavelet transform of an analog signal f is expressed as

$$C_{b,a} = \left(W_{\psi} f\right)(b,a) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} f(t)\, \overline{\psi\left(\frac{t-b}{a}\right)}\, dt \qquad (1)$$

where $C_{b,a}$ is the wavelet coefficient, $\psi(t)$ is the basic wavelet or also called mother wavelet, b is the translation of $\psi(t)$, and a is the dilation or scale factor of $\psi(t)$.

CWT uses Scalogram plots to represent the continuous variation of translation and dilation factors of EEG signals by using its colors. Fig. 3 is an example of Scalogram.

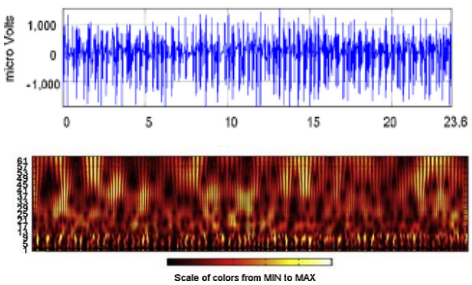

Fig. 3. Example of CWT Scalogram of Epileptic.

Discrete wavelet transform decomposes a raw EEG signal into detail coefficients and approximation coefficient are expressed as

$$C_{b,a} = \left(W_{\psi} f\right)\left(\frac{k}{2^j}, \frac{1}{2^j}\right) \quad (2)$$

In order to calculate $W_{\psi} f(b,a)$, dyadic value $b = k/2^j$ and value $a = 2^{-j}$. This method can also call "dyadic wavelet transform".

If wavelet decomposition goes through the 5th level, the original signal (S) is decomposed into five detail coefficients (D1-D5) and one approximation coefficient (A5) as shown in Fig. 4.

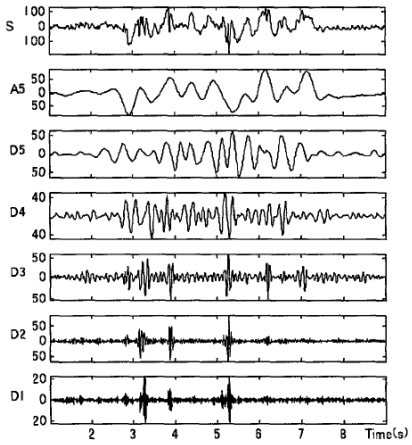

Fig. 4. 5th level of wavelet decomposition

With a comparison between CWT and DWT in [4], wavelet transform is often used more than other methods. For wavelet transform, DWT is often used more than CWT for researches in Epileptic seizure detection. From 81 researches, 40 papers selected DWT, 21 papers selected CWT for their researches, and remainders selected others methods.

V. CLASSIFICATION

Classification is essential process for detecting brain activities or brain disorders. Each classifier typically has different behaviors and provides different results. There is a study on a comparison of several classifiers in [5] such as linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), kernel fisher discriminant (KFD), support vector machine (SVM), multilayer perceptron (MLP), learning vector quantization (LVQ), neural network, k-nearest-neighbor (k-NN), and decision tree (DT) with dataset III and IV in BCI competition 2003. Dataset III and IV are for motor imagery signals and finger movement signals, respectively. Almost of results from the same dataset and feature extraction have nearly accuracy with the fixed time window (fixed starting point and length of time) and each classifier has the best result with different temporal filtering. The result shows that the optimal starting point of time window is between 550–560 for dataset III and 43–46 for dataset IV. The authors suggest that linear classifiers are the first choice because of its simplicity but they seem to be more sensitive with different temporal filter.

The performance of classifiers [6] can be measured by accuracy, sensitivity, and specificity. In this review, *TP*, *TN*, *FP*, and *FN* refer to true positive, true negative, false positive, and false negative, respectively.

Accuracy is the rate of positive and negative outcomes are correctly calculated.

$$Accuracy\ (\%) = \frac{TP + TN}{TN + TP + FN + FP} * 100 \quad (3)$$

Sensitivity is the rate of positive outcomes correctly calculated.

$$Sensitivity\ (\%) = \frac{TP}{TP + FN} * 100 \quad (4)$$

Specificity is the rate of negative outcomes correctly calculated.

$$Specificity\ (\%) = \frac{TN}{TN + FP} * 100 \quad (5)$$

VI. CONCLUSION

This paper reviews about the essential processes for EEG signals. Data acquisition is the process of collecting raw EEG data. Preprocessing is the procedure to eliminate noise and filtering the unwanted data. Feature extraction is the process of transforming signals to the appropriate form for classification. In this review, we pay attention to wavelet transform because of its popularity and providing the multi-resolution signals. Classification is the process to classify the brain activities or disorders.

REFERENCES

[1] J. Satheesh Kumar, and P. Bhuvaneswari, "Änalysis of Electroencephalography (EEG) Signals and Its Categorization – A Study", International Conference on Modeling, Optimizing and Computing (ICMOC 2012)

[2] Trans Cranial Technologies ldt., "10/20 System Positioning Manual"

[3] C. Yamaguchi, "Fourier and Wavelet Analyses of Normal and Epileptic Electroencephalogram (EEG)", 1st International IEEE EMBS Conference on Neural Engineering, Capri Island, Italy, March 20-22, 2003.

[4] O. Faust, U. Rajendra Acharya, H. Adeli, and A. Adeli, "Wavelet-based EEG processing for computer-aided seizure and epilepsy diagnosis", 2015 British Epilepsy Association.

[5] B. Wang, C. Man Wong, F. Wan, P. Un Mak, P. In Mak, and M. I Vai, "Comparison of Different Classification Methods for EEG-Based Brain Computer Interfaces: A Case Study", 2009 IEEE International Conference on Information and Automation, June 22-25, Zhu hai/Macau, China.

[6] V. Sutha Jebakhumari, D. Shanthi, D. Devaraj, "Development of Neural Network Classifier For Classification of epileptic seizures in EEG signals", the 2nd International Conference on Communication and Electronics Systems (ICCES 2017).

Incremental Learning with Deep GMDH Neural Network for Data Stream Mining

Panida Lorwongtrakool
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
panidajlo@gmail.com

Phayung Meesad
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
pym@kmutnb.ac.th

***Abstract*— This research aims to create an incremental learning with Deep Group Method of Data Handling (GMDH) Neural Network for data stream mining. The signal derived from the e-nose and e-tongue consists of different sensors for water quality classification problem. The system mainly consists of three stages: 1) preparation of data stream input, 2) processing with Deep GMDH Neural Network by using Polynomial function to generate Partial Descriptions (PDs), estimate the coefficients of the PD and select the PD with the best predictive capability, 3) batch incremental learning by updating weight polynomial matrix for processing the next chunk of data. The results showed that the most accurate models with the maximum of layers had 10 layers and each layer had a maximum of 10 nodes, a batch size 800 and training data set 80% with accuracy 90.71%. Therefore, system can be applied to Data stream mining and monitoring system in the real environment.**

Keywords—Increment learning, GMDH, Data stream Mining, Deep learning

I. INTRODUCTION

Nowadays, the machine learning techniques such as neural networks provide acceptable accuracy and are widely used for solving a lot of Data Stream Mining tasks[1]. However, the performance of conventional algorithms (such as ANN, SVM) depends on the design of network and features [2]. In addition, it is time-consuming when operated with high dimensional data or Big Data, and it is difficult to guarantee the global convergence. Thus, the shallow neural neurons are not suitable for data streams processing. [1] , Therefore, the deep network was introduced by Hinton [3].

In recent years, Computational Intelligence researcher are interested in deep neural networks (DNN), and it became practically feasible to solve this problem. Among a great number of possible deep neural networks' architectures, the deep networks based on GMDH are one of the most effective networks[4]. The networks are based on the group method of data handling [5] ,which automatically increases a number of layers for information processing to achieve the required accuracy of results.

The Group Method of Data Handling (GMDH) is also known as Polynomial Neural Networks, Abductive and Statistical Learning Networks. The GMDH was applied in a great variety of areas for deep learning and knowledge discovery, forecasting and data mining, optimization and pattern recognition. Inductive GMDH algorithms give possibility to find automatic interrelations in data, to select an optimal structure of model or network and to increase the accuracy of existing algorithms.

However, one relevant problem is "catastrophic forgetting" [6] that may occur when a network, trained with a large set of patterns, has to learn new input patterns, or has to be adapted to a different environment. The risk of catastrophic forgetting is particularly high when a network is adapted with new data that do not adequately represent the knowledge included in the original training data.[5] The solution for this problem is adding a new ability to classifiers. Having the incremental learning can be of great benefit by automatically including the newly presented patterns in the training dataset without affecting class integrity of the previously trained system. [7]

II. METHODOLOGY

A. Collection of Samples

Efficiency of the proposed algorithm was examined based on solving water quality classification problems. The samples of water were collected from Din Daeng water quality control plants. The data were collected by using e-nose and e-tongue consisting of sensors: MQ2, MQ3, MQ4, MQ5,MQ6, MQ7, MQ8, MQ9, MQ135, pH, EC, TDS, salinity, DO, temperature and turbidity. The samples were collected before water flew (inlet area) to the water quality control plants, and the data were collected at the area where water flew out (outlet area) by using a 800cc bottle to collect the data. In addition, a headspace was placed above the water surface, approximately 5 centimeters, to measure a response of the sensor to a smell of water sample.

B. Data preparation

Data preparation: the data were prepared by cleaning and removing noisy data , outlier and normalize data to 0-1.

C. Framework

The conceptual framework of algorithm is shown in Fig. 1.

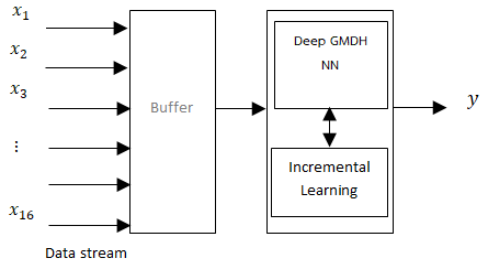

Fig. 1. The architecture of the proposed Incremental Learning with Deep GMDH Neural Network

Algorithm Incremental learning with Deep GMDH Neural Network.

Input:load datastream($x_1, \ldots, x_n$)

1. Begin:

2. chunk_size ← 200,400,600 and 800(records)

3. for i←1 to chunk_size

4. set data.Inputs

5. set data.Targets

6. Innitial Train Data

7. Innitial Test Data

8. Calculate GMDH Network

9. Calculate FitPolynomial

10. Update weights(W)

11. for j←1 to numOf chunk_size

12.

$$W_{x_i,L_{n(j)}new} \leftarrow \frac{(W_{x_i,L_{n(j)}old}(C_{J,new} - 1) + W_{x_i,L_{n(j)}new})}{C_{J,new}}$$

13. endfor j

14. endfor i

15. end

III. RESULTS

The performance of different models is evaluated by accuracy, RMSE and time. The results are shown in the Table 1

TABLE I. RESULTS OF MODEL PERFORMANCE TESTING

Maximum Number of Neurons in a Layer = 5 Maximum Number of Layers = 5				
Chunk Size	**Train(%)**	**Accuracy(%)**	**RMSE**	**Time(sec.)**
200	50	83.51	0.36	23.56
	70	85.32	0.32	27.13
	80	85.64	0.31	26.49
400	50	83.51	0.34	11.69
	70	86.71	0.32	12.37
	80	84.96	0.33	12.39
600	50	87.53	0.33	8.15
	70	85.22	0.33	8.32
	80	84.33	0.33	7.54
800	50	83.67	0.34	4.99
	70	84.22	0.36	5.68
	80	85.75	0.33	5.35
Maximum Number of Neurons in a Layer = 10 Maximum Number of Layers = 10				
Chunk Size	**Train(%)**	**Accuracy(%)**	**RMSE**	**Time(sec.)**
200	50	89.48	0.30	28.50
	70	88.63	0.30	32.63
	80	89.73	0.30	33.94
400	50	88.90	0.30	14.26
	70	89.08	0.30	17.59
	80	90.03	0.30	19.04
600	50	88.43	0.31	9.4
	70	88.00	0.29	10.69
	80	87.62	0.31	10.44
800	50	88.98	0.30	6.26
	70	89.37	0.29	7.60
	80	**90.71**	0.29	7.42
Maximum Number of Neurons in a Layer = 12				
Maximum Number of Layers = 12				
Chunk Size	**Train(%)**	**Accuracy(%)**	**RMSE**	**Time(sec.)**
200	50	89.33	0.29	38.12
	70	90.05	0.30	41.84
	80	88.78	0.29	41.00
400	50	89.07	0.29	20.20
	70	89.16	0.29	20.51
	80	90.23	0.29	25.55
600	50	89.25	0.28	12.10
	70	88.22	0.29	13.97
	80	89.83	0.29	14.34
800	50	89.28	0.30	8.68
	70	88.06	0.27	9.46
	80	88.57	0.30	9.35

According to the Table 1, it showed that the model with batch size was 800, the training dataset was 80%, the maximum number of neurons in a layer was 10, the maximum number of layer was 10, and the confirmed accuracy was up to 90.71%

It is noted that the number of neurons, number of layers and size of training dataset will affect the accuracy. An increase to these numbers will affect accuracy more. On the other hand, if there are too many or too few, it may cause a decrease in accuracy due to overfitting and underfitting, respectively.

According to the results, it can be concluded that the system can be applied to data stream mining and monitoring system in the real environment since the system has the ability to learn and is adaptive when faced with unseen data.

REFERENCES

[1] Bodyanskiy, Y., et al. Fast learning algorithm for deep evolving GMDH-SVM neural network in data stream mining tasks. in 2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP). 2016.

[2] Liu, W., et al., *A survey of deep neural network architectures and their applications.* Neurocomputing, 2017. **234**: p. 11-26.

[3] Hinton, G.E., S. Osindero, and Y.-W. Teh, *A fast learning algorithm for deep belief nets.* Neural Comput., 2006. **18**(7): p. 1527-1554.

[4] Schmidhuber, J., *Deep learning in neural networks: An overview.* Neural Networks, 2015. **61**: p. 85-117.

[5] Ivakhnenko, A.G., *Polynomial Theory of Complex Systems.* IEEE Transactions on Systems, Man, and Cybernetics, 1971. **SMC-1**(4): p. 364-378.

[6] French, R.M., *Catastrophic forgetting in connectionist networks.* Trends in Cognitive Sciences, 1999. **3**(4): p. 128-135.

[7] Tudu, B., et al., *Electronic nose for black tea quality evaluation by an incremental RBF network.* Sensors and Actuators B: Chemical, 2009. **138**(1): p. 90-95.

Database Creation by Natural Language Processing

Chalermpol Tapsai
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
chalermpol.t@email.kmutnb.ac.th

Phayung Meesad
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
phayung.m@it.kmutnb.ac.th

Choochart Haruechaiyasak
National Electronics and Computer Technology Center
Pathumthani, Thailand
Choochart.Haruechaiyasak@nectec.or.th

***Abstract*— The aim of this research is to present a new model of database creation by natural language. This will allow users, who lack the technical knowledge and skills, to be able to create their own databases without having to practice or learn additional languages. By using a variety of techniques, including the analysis of natural language sentences at the level of words, phrases, and sentences. In addition, semantic patterns and ontology are also used to analyze and specify the structure of the data according to the user requirements. Evaluation of the model is conducted by the 30 samples including students and lecturers by inputting natural language description of data into the model to command the computer for database creation. The results showed that this model can support a variety of sentence syntaxes, and create databases that meet the requirements of users with very high accuracy.**

Keywords*— Database, Creation, Natural language processing, Semantic patterns, Ontology.

I. Introduction

A database is an important source of information that helps organizations manage their work efficiently. However, the creation and management of the database require technical knowledge and skills. Moreover, the specific language, SQL is needed for database administration [1], therefore unskilled users cannot create or access the database. Though numerous studies related to Natural Language Processing have been conducted to allow users interfaced with computers by human languages in various topics e.g., text summary [2], document analysis [3], [4], language translation [5], and interface with database[6]. However, in the case of interface with database, most of these studies focus on data retrieval as in [7] without any study directly related to database creation. For this reason, the researchers are interested in developing the new model called Database Creation by Natural Language Processing (DCNLP). This will help users who lack technical knowledge and skills to create their own database easier. By using many techniques, including lexical analysis, phrase analysis, semantic pattern parsing and ontology, The DCNLP model is designed to be able to support the natural language sentences in a variety of syntax that corresponds to the actual usage.

II. Research Methodology

As shown in Fig. 1, there are 3 steps in this research: data collection, model development, and model evaluation.

A. Data collection

In the first step, the researcher collected 100 natural language descriptions of data which are required to store in the computer system by 50 experimental samples including students and lecturers.

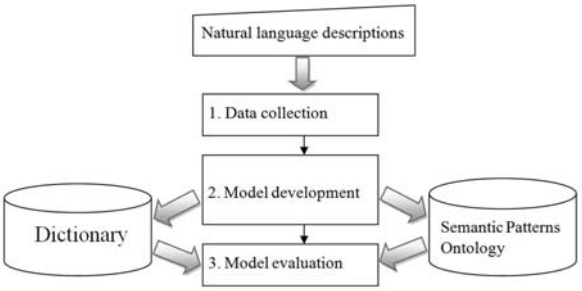

Fig. 1. Example of a figure caption.

B. Model development

The second step, all descriptions were used as a learning dataset which each description was analyzed for keywords, sentence patterns to create a dictionary, Rules and Semantic patterns which are the important parts used by the model to analyzing the sentence meaning and specify the structure of the database according to user specification.

C. Model evaluation

The experimental group consisted of 30 participants input 60 natural language descriptions of data into the model to analyze and create databases. Then the accuracy of the database structure was evaluated according to these descriptions.

III. Processing of the Model

In Fig. 2. There are 6 steps in the DCNLP processing.

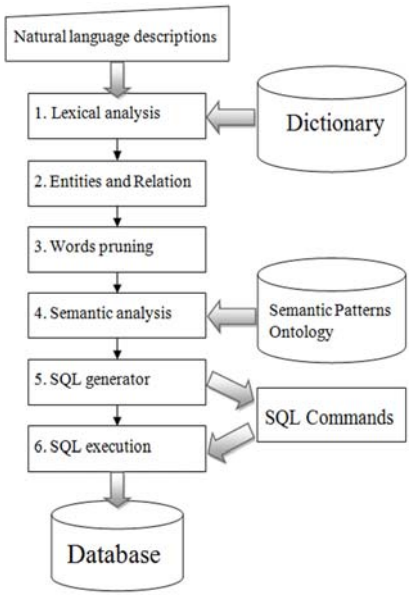

Fig. 2. Processing steps of the DCNLP model

A. *Lexical analysis*

In this step, each natural language description that is inputted by the user is analyzed to separate into words and specify the type of words using the TLS-ART[8].

B. *Entities and Relation analysis*

In this step, all words derived from Step 1 were analyzed to identify the key elements which may be used as Entities or Relation i.e., nouns, noun-phrases, verbs, and verb-phrases. noun-phrases and verb-phrases are words that are made up of several types of words, as shown in Fig. 3.

- รหัสนักศึกษา(Noun) + นักศึกษา(Noun) = รหัสนักศึกษา(Noun)
- คะแนน(Noun) + สอบ(Verb) = คะแนนสอบ(Noun)
- วันที่(noun) + ซื้อ(Verb) + สินค้า(Noun) = วันที่ซื้อสินค้า(Noun)
- ลง(Verb) + ทะเบียน(noun) = ลงทะเบียน(Verb)
- การ(Noun) + เรียน(Verb) = การเรียน(Noun)

Fig. 3. Types of noun-phrases and verb-phrases

C. *Words pruning*

In this step, unnecessary, and redundancy words are eliminated.

D. *Semantic analysis*

In this step, all remaining words are parsed to the semantic pattern in the form of Nondeterministic Finite Automaton with output as shown in Fig. 4 to get the results as the entities, relations, and attributes.

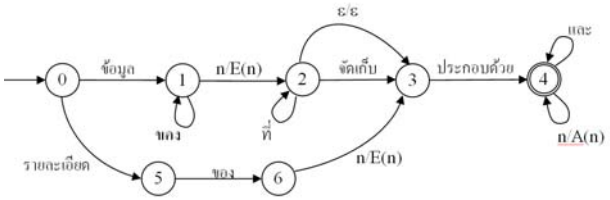

Fig. 4. The semantic pattern in the form of Nondeterministic Finite Automaton with an output

E. *SQL generator*

In this step, the entities, relations, and attributes are mapped to tables and fields and then the SQL commands are created.

F. *SQL execution*

For this step, all SQL commands are executed to create a database.

IV. Experimental result

In the experiment, 60 natural language descriptions are inputted into the model to evaluate the performance. The result showed that the DCNLP model was able to analyze natural language descriptions and create databases with a very high accuracy of 91.67%.

V. Conclusion and discussion

Despite a very high accuracy in the model evaluation, the errors remain a serious problem in database creation. There is 3 main causes of errors: (1) typo error, (2) used of unknown words, and (3) used of unknown sentence syntaxes. Therefore, to improve the efficiency of the model, the number of samples tested should be increased and analyzed for unknown words and more sentence syntaxes to be added into the Model.

References

[1] C. J. Date, An Introduction to Database Systems, 7th ed. Addison-Wesley, Massachusetts: USA, 2000, pp. 83–98.

[2] O. M. Foong, S. P. Yong, and F.A. Jaid, “Text Summarization using Latent Semantic Analysis Model in Mobile Android Platform,” in 9th Modelling symposium, 2015, pp.35-39.

[3] F. Agung, “Software Requirements Specification Analysis Using Natural Language Processing Technique,” IEEE Quality in Research, 2013.

[4] A. S. Hussein, “Visualizing Document Similarity Using N-Grams and Latent Semantic Analysis,” in SAI Computing Conference, London, UK, 2016, pp. 269-279.

[5] D. Moussallem, M. Wauer, A. N. Ngomo, Machine Translation using Semantic Web Technologies: A Survey, Journal of Web Semantics, Volume 51, 2018, pp. 1-19.

[6] M. Llopis, and A. Ferrández, “How to make a natural language interface to query databases accessible to everyone: An example. Computer Standards & Interfaces,”, 2013, pp. 470-481. doi: http://dx.doi.org/10.1016/j.csi.2012.09.005.

[7] A. Shah, J. Pareek, H. Patel, and N. Panchal, “NLKBIDB - Natural language and keyword-based interface to database,” Paper presented at the Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on, pp. 1569-1576.

[8] C. Tapsai, P. Meesad, and C. Haruechaiyasak, “Thai Language Segmentation by Automatic Ranking Trie, ” in Proceedings of 9th International Conference Autonomous Systems (AutoSys 2016), Spain.

The Fatigue Monitoring System using the EEG Signals

Worawut Yimyam
Department of Information Technology Management
King Mongkut's University of Technology North Bangkok, Thailand
Bangkok, Thailand
Worawut_yimyam@hotmail.com

Mahasak Ketcham
Department of Information Technology Management
King Mongkut's University of Technology North Bangkok, Thailand
Bangkok, Thailand
mahasak.k@it.kmutnb.ac.th

***Abstract*—there are many military missions such as the security of the border and the surveillance of Interception drug trafficking. These missions require a lot of military force and take several days to perform, so the soldiers were fatigue from the mission. Therefore, the fatigue effected to work performance and wrong decision making. This paper proposes to develop an algorithm used for monitoring fatigue of the soldiers while they perform their duty. Electroencephalography (EEG) signals are analyzed by an Artificial Neural Networks (ANN) technique and compared with other techniques. The experimental results show that the ANN provides more accurate results than Bayesnet, Support Vector Machines (SVM), and Naïve Bayes techniques. The result of the ANN technique provides the accuracy, recall and precision values at 83.77, 0.838 and 0.838 respectively.**

Keywords—Fatigue, EEG, ANN

I. INTRODUCTION

Nowadays, there are many important duties of solders in the mission of the military to perform against the law breaking such as the maintenance of the independence, sovereignty, national security, border security, and forest patrol. There are also the offenses that affect the national security such as drug and weapon trade, and deforestation. Each mission need lots of times to achieve on it. From that, soldiers may feel fatigued because of their mission. [1] When fatigue occurs, it causes accidents during the mission and also reduce work performance of the solders. For example, there was an accident occurred to US military called M985 in truck drive. The investigation found that the accident happened because of the driver's fatigue which leads him to death because of the lack of sleeping. [2] This type of problem has affected to the wrong decision, communication errors, and risk assessment. All these consequences have systematic relationships. [3]

Military's mission may face the risk any time. One mission needs many hours for working and uses a lot of military personnel. However, a number of soldiers are not enough in some military base. [6-7] Hence, these actions cause fatigue because the soldiers might work several tasks. For example, the mission of the military is to train the pilots to fly helicopter. In facts, the pilots should sleep appropriately because they have to work all day based on FFA rule. They have to wake up at 5.00 am, monitor the helicopter's engine before flying at 4 pm, take off at 5.30 pm, land at 10.30 pm, and store the helicopter at 12.30 am. Then, they move to another base to prepare helicopter at 2.30 am., recheck again at 6.00 am. until military mission is completed. As from their daily routine duties, it can be seen that the pilot's problem is the lack of sleep. Thus, the system is developed to monitor fatigue from the eyes [4-5]. There are many researches working in this area. However, the measurement of fatigue is not certainly accurate and effective.

This paper proposes the development of an algorithm used for monitoring fatigue of the soldiers while they perform their duty. Electroencephalography (EEG) signals are analyzed by an Artificial Neural Networks (ANN) technique and compared with other techniques.

II. SYSTEM DESIGN

This research focuses on the fatigue monitoring system of the military mission. The system is implemented by the EEG sensor in which it can send signal to smartphone via ZigBee and the frequency radio wave, and can collect EEG data via smartphone program. The system is able to analyze the soldier's fatigue conditions and sends the alert to the admin. The Fig. 1 shows the overview of system.

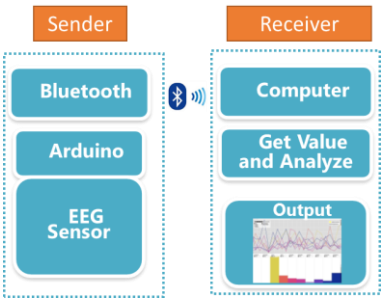

Fig. 1. Overview of system

A. *EEG Sensor*

This part is the analysis of the EEG signal received from EEG sensor. The fatigue from researcher's EEG signal is tested as shown in Fig.2.

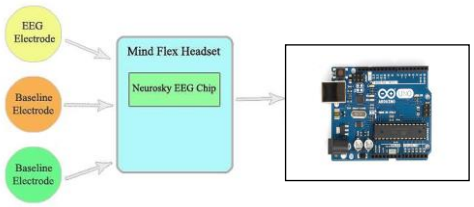

Fig. 2 Neurosky EEG Chip

B. *Communication Link Arduino*

EEG monitoring data is received from MindFlex Headset device. Arduino board is also used to convert value received from sensor to different types of signals which can be divided into 8 ranges: Delta 1-3 Hz, Theta 4-7 Hz, Low Alpha 8-9 Hz, High Beta 18-20 Hz, Low Gamma 31-40 Hz, and High Gamma 41-50 Hz. These frequency waves are transformed in ASCII coding.

C. *Data reception*

In data reception, the system connects to the Bluetooth module in order to send signal to computer for analyzing the fatigue in the next step.

D. Receive value and analyze with ANN

Artificial Neural Network technique (ANN) is used to nalyze data receiving from the EEG signals. The processing applies with neural network of the human brain. The EEG signal is an input of the ANN technique. The EEG data composes of Delta, Theta, Low Alpha, High Alpha, Low Beta, High Beta, Low Gamma, and High Gamma signals. All inputs are multiplied with weight which is represented as w1, w2, w3, w4, w5, w6, w7, and w8. Each neuron is a bias adjustment with the weighting. It has been sent to the transfer function in order to calculate the result as shown in Figure 3.

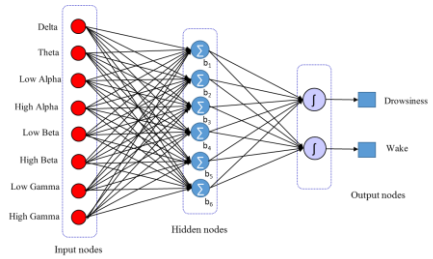

Fig.3 Example of Artificial Neural Network Technique

The Equation is shown as below:

$$a^m = f^{m+1}(w^{m+1}x^m + b^{m+1})$$

Where

a^m means Output Node

f^m means Transfer Function

w^m= 0.2

b^m = 0.1

x^m =means Delta, Theta , Low Alpha, High Alpha, Low Beta ,High Beta ,Low Gamma , High Gamma

III. Experimental Results

The performance of classification is conducted by the ANN technique. It considers the accuracy of precision and recall as shown equation (2), (3), and (4). Table 1 shows the experimental results of the performance of predicted class. Table 2 shows the comparison of the performance of classification.

$$Precision(p) = \frac{TP}{TP+FP} \quad (2)$$

$$Recall(r) = \frac{TP}{TP+FN} \quad (3)$$

$$Accuracy(A) = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

TP = True Positive, FP = False Positive, FT = True Negative และ FN = False Negative

Table 1. The result of data experimental classification

		PREDICTED CLASS	
		Drowsiness	Wake
ACTUAL CLASS	Drowsiness	3546	595
	Wake	667	2993

Table 2. The comparison of the performance of classification

Model	*10-fold cross validation*		
	Accuracy	Recall	Precision
ANN	83.77	0.838	0.838
Bayesnet	77.07	0.771	0.777
SVM	75.40	0.754	0.759
NaiveBaye	62.10	0.621	0.712

IV. Conclusion

Researcher proposes the development of algorithm for monitoring fatigue in military mission based on brain signals. The ANN was applied to analyze the data. As a result, ANN performs higher performance than Bayesnet, Support Vector Machines (SVM), and Naïve Bayes. The experimental result of ANN technique showed the percentage of its accuracy, recall, and precision values at 83.77, 0.838, and 0.838, respectively.

References

[1] B. Sicard, E. Jouve, and O. Blin, "Risk propensity assessment in military special operations," Military medicine, 166(10), 871-874, 2001. (1)

[2] U. S. Army, "Leaders' manual for combat stress control," Field Manual 22 51 (1994).

[3] J. M. How, S. C. Foo, E. Low, T. M. Wong, A. Vijayan, M. G. Siew and R. Kanapathy, "Effects of sleep deprivation on performance of Naval seamen: I. Total sleep deprivation on performance," Annals of the academy of medicine, Singapore, 23(5), 669-675, 1994.

[4] T. Brandt, R. Stemmer, and A. Rakotonirainy, "Affordable visual driver monitoring system for fatigue and monotony." In Systems, Man and Cybernetics, 2004 IEEE International Conference on (Vol. 7, pp. 6451-6456). IEEE.

[5] T. Von Jan, T. Karnahl, K. Seifert, J. Hilgenstock, and R. Zobel, "Don't sleep and drive-VW's fatigue detection technology," In Proceedings of 19th International Conference on Enhanced Safety of Vehicles, Washington, DC (Vol. 168). 2005.

[6] U.S. Army Safety Center. "Sustaining Performance in Combat," Flight fax, (31)5:9-11. 2003.

[7] U.S. Army Safety Center. "Fatigue", Countermeasure, (23)3:4-5. 2002.

Author Index

Keyword Index

The 5th Joint Symposium on Computational Intelligence, Bangkok, Thailand

Title: Proceedings of the 5th Joint Symposium on Computational Intelligence (JSCI 5)
Editors: Phayung Meesad and Kitsuchart Pasupa
Published by: IEEE Computational Intelligence Society Thailand Chapter
Printing House: King Mongkut's University of Technology North Bangkok
Edition: 1
Month and Year: September 2018
ISBN: 978-616-368-065-5 ©2018 JSCI e-book