MEDINFO 2021: One World, One Health – Global Partnership for Digital Innovation
P. Otero et al. (Eds.)

doi:10.3233/SHTI220153

Protected Health Information Recognition of Unstructured Code-Mixed Electronic Health Records in Taiwan

You-Qian Lee[a], Bo-Hong Wang[a], Chu-Hsien Su[b], Pei-Tsz Chen[c], Wu-Qing Lin[a], Chi-Shin Wu[b], Hong-Jie Dai[a,d,e]

[a]*Intelligent System Lab, College of Electrical Engineering and Computer Science, Department of Electrical Engineering, National Kaohsiung University Science and Technology, Kaohsiung, Taiwan R.O.C.*
[b]*Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan R.O.C.,*
[c]*Department of Chemical Engineering, Feng Chia University, Taichung, Taiwan R.O.C.*
[d]*National Institute of Cancer Research, National Health Research Institutes, Tainan, Taiwan, R.O.C.*
[e]*School of Post-Baccalaureate Medicine, Kaohsiung Medical University, Kaohsiung, Taiwan R.O.C.*

Abstract

Electronic health records (EHRs) at medical institutions provide valuable sources for research in both clinical and biomedical domains. However, before such records can be used for research purposes, protected health information (PHI) mentioned in the unstructured text must be removed. In Taiwan's EHR systems the unstructured EHR texts are usually represented in the mixing of English and Chinese languages, which brings challenges for de-identification. This paper presented the first study, to the best of our knowledge, of the construction of a code-mixed EHR de-identification corpus and the evaluation of different mature entity recognition methods applied for the code-mixed PHI recognition task.

Keywords:

Electronic Health Record; Data Anonymization; Code-Mixing.

Introduction

In recent years, the growing request of electronic health record (EHR) systems led health-care providers to adopt rapidly various solutions [9]. In particular, EHRs in unstructured format are valuable sources for research in both clinical and biomedical domains. To protect the privacy of patients whose data were secondary used for other purposes, regulations or laws such as General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) requires that protected health information (PHI) be removed from records before they can be disseminated. In scenarios like researches, obtaining explicit consent may be impractical or impossible. If one can de-identify personal data to a level of full anonymization, the data is no longer personal data, and subsequent uses of the data are no longer regulated. However, manual de-identification of large volume of EHRs is prohibitively expensive, time-consuming and prone to error, necessitating automatic methods for large-scale, automated de-identification.

On the other hand, physicians in Taiwan usually write notes in the mixing of English and Chinese languages referring to as the problem of code mixing [4]. Code-mixing causes problems for many language processing systems that are based on a particular language model. In this work, we present the construction of a de-identification corpus and investigate the effectiveness of the application of state-of-the-art entity recognition methods to de-identify PHIs mentioned in unstructured medical records in the manner of Chinese-English code-mixing.

Methods

In this subsection, we first describe the source of our dataset, the annotation guideline and the annotation process. We then analyze the level of code-mixing of our dataset. Finally, we describe the implementation of several methods whose performance will be reported in the Results section.

Code-mixing De-identification Corpus and Its Annotations

Data Source

With the approval of the research ethics committee of the National Taiwan University Hospital (NTUH) the EHR data sampled from the psychiatric unit of the NTUH-Integrated Medical Database were used in this study. The collected corpus contains 4,737 unstructured discharge summaries. The dataset has been previously used in our previous works [1; 13]. We randomly sub-sampled 800 summaries as the final dataset used in this study.

Annotation Guideline and Annotation Process

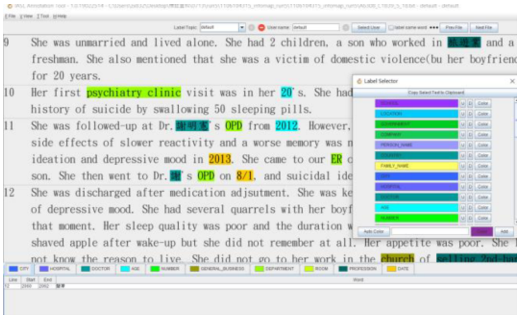
We extended the privacy rules defined by HIPAA to define the following types of PHIs represented in either English or Chinese, which are required to be removed from patients' medical records to protect patient privacy.

- Patient/Doctor/Person/Family: Name of the patient, medical staffs, other persons or any person who has relationship with the patient.
- Date: Any calendar date, including years, seasons, months, and holidays.
- Age: Age of any person.
- Hospital: Health care institutions providing patient treatment with specialized medical and nursing staff and medical equipment (e.g. National Taiwan University Hospital, 署北醫院).
- Department/Room/Number: Designated section/hospital rooms/bed numbers in hospital where the patient was being treated (e.g. emergency department, rehabilitation ward, 08診, 10 號).
- Location: State/country names as well as addresses and cities. Each part of an address should annotate with its own tag (e.g. street, city, country, region, etc.)

Table 1: The PHI distributions in the training and test sets. (means the sentence is code-mixed)*

PHI Type	ENG		ENG-CHI*		All	
	Train	Test	Train	Test	Train	Test
Person	0	4	47	29	47	33
Doctor	44	26	802	308	846	334
Patient	3	6	86	36	89	42
Family	0	0	0	1	0	1
Date	18,625	5,794	3,761	1,365	22,386	7,159
Age	1,307	450	377	152	1684	602
Location	12	9	156	37	168	46
Nationality	5	3	4	2	9	5
Region	5	0	3	1	8	1
Country	219	72	105	53	324	125
City	91	49	174	46	265	95
Hospital	867	286	1,227	343	2,094	629
Department	1,617	488	2,052	612	3,669	1,100
Room	498	285	854	262	1,352	547
Number	0	0	691	203	691	203
School	11	17	364	140	375	157
General Business	99	58	420	124	519	182
Profession	245	89	773	211	1,018	300
ID Number	12	1	98	137	110	138
Medical Record	0	6	61	29	61	35
Phone	0	0	0	1	0	1
Numbers of PHI Tokens	23,660	7,643	12,055	4,092	35,715	11,735
Number of Non-PHI token	391,758	127,579	155,900	52,604	547,658	180,183
Number of Sentences	34,672	11,318	6,514	2,001	41,286	13,319

- Nationality: Country of citizenship (*e.g.* Taiwanese, Vietnamese)
- School: Educational institution designed to provide learning spaces and learning environments (*e.g.* 格玫高中, NKUST).
- General Business: other generic locations like "KTV" or named organizations by types.
- Profession: Any job which is not held by someone on the medical staff.
- Medical Record: Any medical record IDs or numbers.
- Phone: Phone number or fax.
- Id Number: Not sure what type of an ID is, annotate it as with this type (*e.g.* "32001CXM").



9 She was unmarried and lived alone. She had 2 children, a son who worked in [illegible] and a
freshman. She also mentioned that she was a victim of domestic violence(bu her boyfrienc
for 20 years.
10 Her first psychiatry clinic visit was in her 20's. She had
history of suicide by swallowing 50 sleeping pills.
11 She was followed-up at Dr. [illegible]'s OPD from 2012. However,
side effects of slower reactivity and a worse memory was n
ideation and depressive mood in 2013. She came to our ER c
son. She then went to Dr. [illegible]'s OPD on 8/1. and suicidal ide
12 She was discharged after medication adjsutment. She was ke
of depressive mood. She had several quarrels with her boyf
that moment. Her sleep quality was poor and the duration w
shaved apple after wake-up but she did not remember at all. Her appetite was poor. She
not know the reason to live. She did not go to her work in the church of selling 2nd-han

Figure 1: Annotation tool developed for the annotators.

To annotate the data, we developed an annotation tool shown in Figure 1, which can load the textual content of an EHR with an editing interface for annotators to label one or many continuous words as a PHI and assign it with the corresponding PHI type. Four annotators who know both English and Chinese were recruited to use the tool to annotate the dataset. They followed the above guideline to annotate an identical set of 200 randomly sampled records. Afterwards a meeting was organized to discuss issues and concerns encountered during the annotation process and the annotation guideline was adjusted according to the conclusion of the meeting. The above process was conducted iterative until the annotators achieved an agreement above substantial. The final Kappa inter-annotation agreement [12] was 0.85. The remaining unlabeled 600 EHRs were evenly distributed to all annotators for labeling.

Code Mixing Level Analysis

Code-Mixing Index

We used the following code-mixing index (CMI) formula defined by [4] to measure the level of mixing between English and Chinese in our corpus.

$$\text{CMI} = \begin{cases} 100 \times \left[1 - \dfrac{max\{W_i\}}{n-u}\right] : n > u \\ 0 : n = u \end{cases}$$

where $max\{W_i\}$ is the number of words of the most frequent language (English in our case); n is the total number of mixed words; u indicates the number of language independent words such as numeric values and punctuations. We calculated CMI for the most frequent language while the sentence is mixed with English and Chinese. If the sentence is monolingual, CMI is 0. We show the example in Figure 2 illustrates an example for the calculation of CMI.

For the sentence containing 77 words, we calculated the number of codes as follows.

ZH: 49, EN: 20, UNIV: 4

The corresponding CMI given the above values is 32.8.

De-identification Methods

We formulated the task as a sequential labeling task and applied the BILOU encoding schema.

- The U-label refers to a single word PHI.
- The B-label refers to the beginning of a multi-word PHI.
- The I-label refers to the inside of a multi-word PHI.
- The L-label refers to the last word of a multi-word PHI.
- The O label refers to the outside of PHIs.

He|**en** was|**en** presented|**en** with|**en** dyspnea|**en** for|**en** days|**en** and|**en** treated|**en** at|**en** 高|**zh** 雄|**zh** 基|**zh** 督|**zh** 教|**zh** 醫|**zh** 院|**zh** initially|**en** …|**univ** We|**en** informed|**en** family|**en** critical|**en** condition|**en** they|**en** signed|**en** DNR|**en** (|**univ** 家|**zh** 屬|**zh** 主|**zh** 動|**zh** 表|**zh** 示|**zh** 已|**zh** 填|**zh** 寫|**zh** DNR|**en** 並|**zh** 在|**zh** 此|**zh** 宣|**zh** 死|**zh** 亡|**zh** 因|**zh** 為|**zh** 病|**zh** 患|**zh** 年|**zh** 歲|**zh** 已|**zh** 大|**zh** 和|**zh** 已|**zh** 植|**zh** 物|**zh** 人|**zh** 狀|**zh** 態|**zh** 臥|**zh** 床|**zh** 35|**univ** 年|**zh** 不|**zh** 想|**zh** 再|**zh** 看|**zh** 到|**zh** 病|**zh** 患|**zh** 受|**zh** 苦|**zh** |**univ**

Figure 2: An example sentence contains 77 words. Each word is tagged with one of the following codes: English (en), Chinese (zh), and language-independent symbols (univ).

We implemented the following methods to examine the effectiveness of different de-identification methods on the compiled corpus of code-mixing text.

- **Dictionary-based Approach**: The algorithm assigns each word with the most prevailing label observed in the training set. In case the word is out-of-vocabulary, the O label is assigned.
- **Conditional Random Field (CRF)**: The CRF algorithm [6] was implemented and trained with the following features.
  - Word features: A context window of three was used to extract the word features. Note that all extracted words were lowercased.
  - Part-of-Speech (PoS): the PoS information within the context window of three was extracted by using our clinical natural language processing tool [2].
  - Orthographical features: the orthographical features proposed in our previous work [10].
- **Bidirectional Long Short-Term Memory (BiLSTM) + CRF [5]:** The neural network-based method includes an embedding layer, followed by a BI-LSTM layer to capture the long-term dependency and a CRF layer to determine the best labeling sequence.
- **BERT-based Models:** BERT [3] is a language model based on the bidirectional encoder of Transformer [11], which broke several records of language-based tasks. We followed the suggestion of Devlin et al. [3] by adding a linear layer on top of the used BERT models to transform the output of BERT to meet the number of expected number of BILOU tags of our PHI types. The following three BERT pretrained models were considered and fine-tuned on our code mixing de-identification corpus: BERT-base cased, BERT-base-Chinese and BERT-base-multilingual-cased. The reason why we consider the above models is that our dataset contains Chinese-English mixed text. So we would like to see the effect of transferring the pre-trained multilingual BERT on our task. We didn't distinguish the Chinese and English tokens and didn't apply Chinese word segmentation because the applied BERT models were pre-trained without applying whole word masking.

Experimental Settings and Evaluation

The python implementation of CRF (sklearn-crfsuite) was used to develop the CRF model. The coefficients for both the L1 and L2 regularizations were set to 0.1. For BiLSTM+CRF, The embedding layer was initialized with the pretrained GloVe [8] vectors with the dimension of 300. In total of 256 × 2 hidden units was used in the BiLSTM network.

For the BERT-based Models, we fine-tuned the whole layers and parameters of the pretrained BERT. The optimizer for the neural network-based models was AdamW. The max iteration was set to 20.

Results are reported as F-scores defined as follows by using a sequence labeling evaluation script developed for named entity recognition [7]. *nb_correct* indicates the number of correctly recognized PHI mentions whose categories and spans exactly matched with the ground truths. *nb_pred* refers to the number of predicted PHI mentions. *nb_true* is the number of manually annotated PHIs.

$$\text{precision} = \frac{nb\_correct}{nb\_pred}; \text{recall} = \frac{nb\_correct}{nb\_true}$$

$$\text{F}_1\text{-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Results

PHI Distribution of the Code-Mixing Deidentification Corpus

We considered the 200 co-annotated EHRs as the test set and the remaining 600 EHRs as the training set. The training and test sets containing 41,286 and 13,319 sentences, respectively. The statistics information of the 22 PHI types is listed in Table 1. ENG and CHI means English and Chinese respectively.

From Table 1, it is worth noting that the total number of PHI tokens in the training set is 35,715, whereas the non-PHI tokens add up to 547,658. This means that only 6% of the tokens of the training set are PHI-related. Likewise, the ratio of tokens for the test set is 6.1%, this skewed distribution poses a challenge considering that the datasets are further separated by 22 PHI types. We can also observe that except for "Date" and "Age", most PHI types tend to be mentioned more frequently in the code-mixed sentences than sentences in English.

Table 2 presents the results of code-mixing analysis. The calculated CMIs of the training and test sets are around 10%. Comparing with the CMI values of English–Bengali (~5.15%), and Dutch–Turkish (~4.13%) in chat posts [4], our mixed level is larger, which should lead to more serious challenges in the task of de-identification.

Table 2: The CMI for training and test set

	Avg. CMI(%)	Number of sentence
Training	10.84%	41,286
Test	9.56%	13,319
Overall	10.2%	54,605

Table 3: F-scores of the developed methods on the test set. The values in boldface means the highest score for the PHI type.

PHI Type	Dictionary	CRF	BiLSTM+ CRF	BERT-base-cased	BERT-base-Chinese	BERT-multilingual
Person	0.049	0.000	0.154	0.100	0.273	**0.304**
Doctor	0.119	0.836	0.884	0.692	**0.897**	0.883
Patient	0.000	0.467	0.627	0.548	0.747	**0.780**
Family Name	0.000	0.000	0.000	0.000	0.000	0.000
Date	0.125	0.963	**0.983**	0.974	0.976	0.976
Age	0.234	0.878	**0.942**	0.881	0.894	0.888
Location	0.000	0.197	0.310	0.111	**0.370**	0.262
Nationality	0.444	0.000	0.000	0.000	**0.667**	0.286
Region	0.000	0.000	0.000	0.000	0.000	0.000
Country	0.551	0.793	0.827	0.773	0.865	**0.919**
City	0.443	0.684	0.683	0.696	0.835	**0.857**
Hospital	0.376	0.903	0.908	0.840	**0.922**	0.902
Department	0.507	0.837	**0.865**	0.796	0.853	0.847
Room	0.080	0.871	**0.874**	0.827	0.815	0.831
Number	0.038	0.990	0.990	0.978	0.990	0.990
School	0.013	0.680	0.694	0.442	**0.750**	0.745
General Business	0.056	0.544	0.548	0.254	0.590	**0.602**
Profession	0.088	0.545	0.581	0.379	0.628	**0.650**
ID Number	0.611	0.871	**0.871**	0.793	0.684	0.802
Medical Record	0.000	0.848	0.906	0.871	0.906	0.906
Phone	0.000	0.000	0.000	0.000	0.000	0.000
Micro-F	0.177	0.909	**0.929**	0.886	0.921	0.922
Macro-F	0.178	0.582	0.617	0.538	**0.663**	0.652

Code-Mixed De-identification Performance

Performance Comparison

Table 3 shows the results of the implemented methods on the test set. The F-scores of the dictionary-based method is lower than 0.5, which may be owing to the reason that we didn't compile comprehensive PHI dictionaries for all PHI types; we only collected the dictionaries from the training set. All the machine learning-based models achieved satisfied micro-F-scores over 0.9. Surprisingly BERT-base-cased had the lowest micro-/macro-F-scores while the CRF model without transferred embedding achieved a comparable performance. The BiLSTM+CRF model with the pretrained GloVe embedding achieved the highest micro-F-score owing to it performed better than others in PHI types like "Date", "Age", and "Department", which occupies more than half of the annotations in the test set. The BERT models pre-trained on the Chinese and multilingual corpora achieved better macro-F-scores. With considering Table 1, we can observe that the two BERT-based models performed better in PHI types like "Person", "Patient", "Nationality" and "School", which appeared more frequently in code-mixed sentences. The results suggest that the inclusion of the pretrained word embedding or pretrained language models could enhance the capability of the models to recognize PHIs in code-mixed sentences.

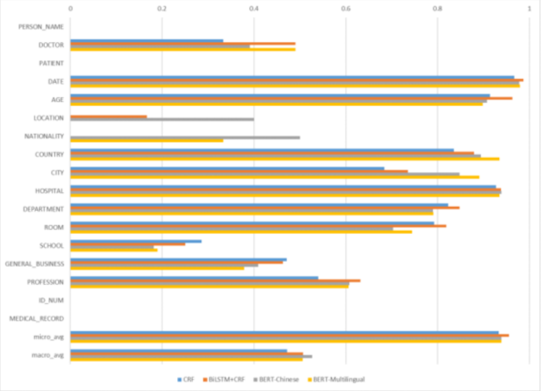


Figure 3: F-score Comparison among sentences in English

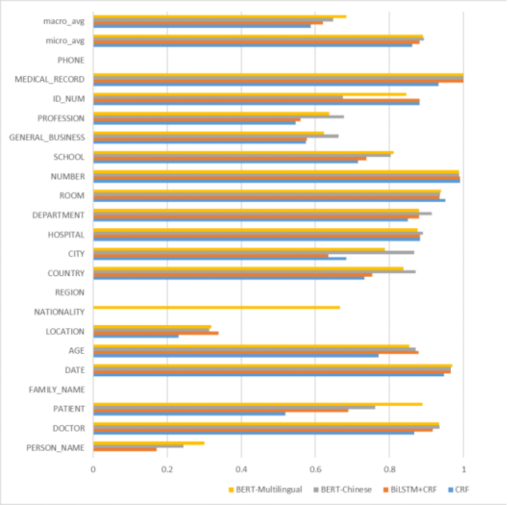


Figure 4: F-score Comparison among sentences contained English-Chinese mixed text.

Discussion

As illustrated in Table 2, the mixed level of our corpus is high, we therefore separated the sentences in the test set into sentences without code mixing and sentences containing code-mixing words to examine the influence of code-mixing for the task of PHI recognition. Figure 3 and 4 shows the results of the performance comparison. Note that in our corpus, we didn't have any sentences containing only words in Chinese.

All supervised learning methods performed better in the set of English sentences, but their F-scores dropped by ~0.07 for CRF and BiLSTM+CRF and ~0.05 for BERT-based models on the mixed sentences. In general, we observed that the BERT-based methods tend to perform slightly better in PHI types described in the manner of code-mixed. For instance, Figure 5 shows the distribution of the code-mixing for "Patient" and "ID Number" in the test set. We can see that for the "Patient" PHI type, 86% PHIs are described in Chinese words only. In contrast, 99% PHIs are described in English words only for the "ID Number" type. By contribution to the pretrained language models, the BERT-Chinese and BERT-multilingual models respectively outperformed CRF by 0.242 and 0.37 on "Patient". For PHIs in "ID Number" which were mainly described in English, BERT-multilingual achieved a similar F-score (0.845) with CRF (0.881), but the F-score of BERT-Chinese dropped to 0.675.

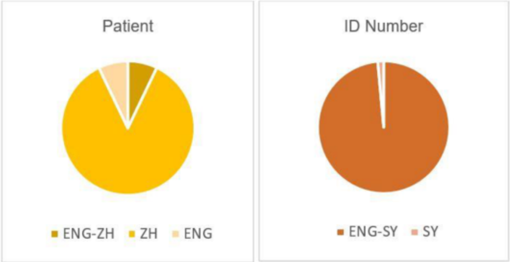


Figure 5:Code mixing level for "Patient" and "ID Number"

Conclusions

In this work we have presented the construction of a unique code mixing dataset for the task of de-identification of EHRs in Taiwan. The experiment results exhibited that supervised learning methods can reliably recognize most of PHI types but for PHI types like "Person" and "Patient", which tend to be described in Chinese or mixed language, the state-of-the-art pretrained language models still have room for improvement. This serves as a strong evidence that we still require the development of more robust approaches that can recognize PHIs in the code-mixed manner.

Acknowledgements

This study was supported by the grant from the Ministry of Science and Technology, Taiwan (MOST 109-2221-E-992-074 - MY3). We thank the staff of the Department of Medical Research, National Taiwan University Hospital, for access to the Integrated Medical Database.

References

[1] H.-J. Dai, C.-H. Su, Y.-Q. Lee, Y.-C. Zhang, C.-K. Wang, C.-J. Kuo, and C.-S. Wu, Deep Learning-Based Natural Language Processing for Screening Psychiatric Patients, *Frontiers in Psychiatry* **11** (2021).

[2] H.-J. Dai, C.-H. Su, and C.-S. Wu, Adverse Drug Event and Medication Extraction in Electronic Health Records via a Cascading Architecture with Different Sequence Labeling Models and Word Embeddings, *Journal of the American Medical Informatics Association* (2019).

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of NAACL-HLT 2019*, Minneapolis, Minnesota, 2019, pp. 4171-4186.

[4] B. Gambäck and A. Das, On measuring the complexity of code-mixing, in: *Proceedings of the 11th International Conference on Natural Language Processing, Goa, India*, 2014, pp. 1-7.

[5] Z. Huang, W. Xu, and K. Yu, Bidirectional LSTM-CRF models for sequence tagging, *Computing Research Repository* (2015).

[6] J. Lafferty, A. McCallum, and F. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in: *Proceedings of the 18th International Conference on Machine Learning (ICML)*, 2001, pp. 282–289.

[7] H. Nakayama, seqeval: A python framework for sequence labeling evaluation, *Software available from https://github.com/chakki-works/seqeval* (2018).

[8] J. Pennington, R. Socher, and C.D. Manning, Glove: Global vectors for word representation, *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* **12** (2014), 1532-1543.

[9] S. Silvestri, A. Esposito, F. Gargiulo, M. Sicuranza, M. Ciampi, and G. De Pietro, A big data architecture for the extraction and analysis of EHR data, in: *2019 IEEE World Congress on Services (SERVICES)*, IEEE, 2019, pp. 283-288.

[10] R.T.-H. Tsai, C.-L. Sung, H.-J. Dai, H.-C. Hung, T.-Y. Sung, and W.-L. Hsu, NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition, *BMC Bioinformatics* **7** (2006), S11.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in: *Advances in neural information processing systems*, 2017, pp. 5998-6008.

[12] A.J. Viera and J.M. Garrett, Understanding interobserver agreement: the kappa statistic, *Fam med* **37** (2005), 360-363.

[13] C.-S. Wu, C.-J. Kuo, C.-H. Su, L.-X. Wei, W.-H. Lu, S.H. Wang, and H.-J. Dai, Text mining approach to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records, *Journal of Affective Disorders* **260** (2020), 617-623.

Address for correspondence

Hong-Jie Dai, hjdai@nkust.edu.tw, Intelligent System Lab,
Department of Electrical Engineering, College of Electrical
Engineering and Computer Science, National Kaohsiung University
of Science and Technology, No. 415, Jiangong Rd., Sanmin Dist.,
Kaohsiung City 807618, Taiwan R.O.C.