IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 25, 2021, accepted January 21, 2022, date of publication February 1, 2022, date of current version March 4, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3148396

Principle-Based Approach for the De-Identification of Code-Mixed Electronic Health Records

CHEN-KAI WANG[1,2], FENG-DUO WANG[3], YOU-QIAN LEE[4], PEI-TSZ CHEN[5], BO-HONG WANG[4,6], CHU-HSIEN SU[7], JOSEPH CHIN-CHI KUO[8], CHI-SHIN WU[7,9], YI-LING CHIEN[10], HONG-JIE DAI[4,11,12], (Member, IEEE), VINCENT S. TSENG[1], (Fellow, IEEE), AND WEN-LIAN HSU[13,14], (Life Fellow, IEEE)

[1]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan
[2]Advanced Technology Laboratory, Chunghwa Telecom Laboratories, Taoyuan 326402, Taiwan
[3]Department of Computer Science, National Tsing Hua University, Hsinchu 300044, Taiwan
[4]Intelligent System Laboratory, Department of Electrical Engineering, College of Electrical Engineering and Computer Science, National Kaohsiung University of Science and Technology, Kaohsiung 807618, Taiwan
[5]Department of Chemical Engineering, Feng Chia University, Taichung 407802, Taiwan
[6]Department of Electrical Engineering, National Sun Yat-sen University, Kaohsiung 804201, Taiwan
[7]National Center for Geriatrics and Welfare Research, National Health Research Institutes, Miaoli 35053, Taiwan
[8]Big Data Center, China Medical University Hospital, China Medical University, Taichung 404332, Taiwan
[9]Department of Psychiatry, National Taiwan University Hospital, Yunlin Branch, Yunlin 632007, Taiwan
[10]Department of Psychiatry, College of Medicine, National Taiwan University Hospital, Taipei 100229, Taiwan
[11]School of Post-Baccalaureate Medicine, Kaohsiung Medical University, Kaohsiung 807378, Taiwan
[12]National Institute of Cancer Research, National Health Research Institutes, Tainan 70456, Taiwan
[13]Department of Computer Science and Information Engineering, College of Information and Electrical Engineering, Asia University, Taichung 413305, Taiwan
[14]Pervasive AI Research Labs, Ministry of Science and Technology, Hsinchu 300093, Taiwan

Corresponding authors: Hong-Jie Dai (hjdai@nkust.edu.tw) and Vincent S. Tseng (vtseng@cs.nctu.edu.tw)

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant MOST109-2221-E-992-074-MY3 and Grant MOST 110-2314-B-400-053-MY3.

ABSTRACT Code-mixing is a phenomenon where at least two languages are combined in a hybrid manner in the context of a single conversation. The use of mixed language is widespread in multilingual and multicultural countries and poses significant challenges for the development of automated language processing tools. In Taiwan's electronic health record (EHR) systems, unstructured EHR texts are usually represented in a mixture of English and Chinese which increases the difficulty for de-identification and synthetization of protected health information (PHI). We explored this problem by applying several state-of-the-art pre-trained mono- and multilingual language models and propose to exploit the principle-based approach (PBA) for the tasks of PHI recognition and resynthesis on a code-mixed EHR corpus annotated with 6 main categories and 25 subcategories of PHIs. A hierarchical principle slot schema is defined in the PBA to encode knowledge of code-mixed PHIs and utilize slots to learn from the training set to assemble principles for recognizing PHI mentions and synthesizing surrogates simultaneously. In addition, a semantic disambiguation process is implemented to disambiguate ambiguous PHI categories in the de-identification process and to dynamically extend the knowledge encoded in PBA during the knowledge augmentation process. The experiment results demonstrate that the proposed method can achieve the best micro- and macro-F-scores in comparison to the other mono- and multilingual language models fine-tuned on our code-mixed corpus.

INDEX TERMS Electronic health record, data anonymization, code-mixing, principle, named entity recognition, deep learning.

I. INTRODUCTION

The Health Information Technology for Economic and Clinical Health Act signed in 2009 expedited the growth of the use of electronic health record (EHR) systems in USA and subsequently many countries worldwide [1]. A survey conducted in 2016 revealed that almost 95% of all hospitals in USA are eligible for using EHR systems [2]. The development of EHR systems in Taiwan was initiated by the National Health Informatics Project in 2004. According to the statistics

The associate editor coordinating the review of this manuscript and approving it for publication was Gina Tourassi.

[張三]PATIENT is a **[25]**AGE-year-old male with history of …**[張三]**PATIENT was brought to **[中國醫]**HOSPITAL for help in **[2095/8]**DATE where he visited medical physician for his multiple somatic complaints such as muscle soreness and generalized weakness ... education: studied at **[中國醫]**SCHOOL 獸醫三年級，因發病休學一年 …

Hospitalization:
Admission date R VS
[2105.12.15]DATE **[王五]**DOCTOR [趙六]DOCTOR

FIGURE 1. A snippet of code-mixed de-identified electronic health record.

of the Electronic Medical Record Exchange Center in 2016, 411 of 496 hospitals (80.4%) and about 5,244 of 9,782 private clinics (53.6%) in Taiwan were certified as having interoperable EHRs [3]. Technology advancements have resulted in a tremendous volume of EHRs containing significant health care information, prompting healthcare providers to adopt various approaches to extract and analyze EHR data for secondary use, particular data within unstructured texts [4].

To protect the privacy of patients whose data was used for secondary purposes, laws and regulations such as the General Data Protection Regulation (GDPR) [5] and the Health Insurance Portability and Accountability Act (HIPAA) [6] request protected health information (PHI) to be removed from records through a process called de-identification before they can be disseminated. However, manual de-identification of a large volume of EHRs is prohibitively expensive, time-consuming and prone to error, necessitating the development of methods for automated large-scale de-identification.

Figure 1 provides a snippet of a EHR in Taiwan illustrating that physicians in Taiwan usually write notes in a mixture of English and Chinese which leads to the problem of code-mixing [7]. This linguistic phenomenon occurs when two or more languages are alternatively used in the context of a single conversation or situation. It is widespread in multilingual and multicultural countries [8]–[10] and prevalent in the physician-patient relationship [11]. Code-mixing is a major problem for most language processing systems that were established based on a particular language model. For example, the Chinese knowledge and information processing (CKIP)[1] word segmentation system cannot correctly segment the code-mixed descriptions shown in Figure 1. Furthermore, the code-mixed writing style also hinders the recognition and synthesis of PHIs. For instance, dates can be described in various code-mixed forms such as "184/08" in the Minguo calendar, which is an equivalent of "2095/8" in the Christian calendar in Figure 1.

Without access to EHRs containing identifiable information, researchers relied on "resynthesized" EHRs, such as the n2c2 de-identification corpus [12], to design their de-identification tools. In such corpora, the actual PHI of patients is replaced with dummy identifiers to simulate natural language [13]. In this work, we propose a principle-based approach (PBA) which performs PHI recognition and resynthesis simultaneously. This approach has been successfully employed in several areas including sentimental analysis [14] and biomedical named entity recognition and normalization [15], but is being applied to EHR de-identification for the first time.

For ambiguous PHIs such as "中國醫" in Figure 1, we extended the PBA by using contextual embedding generated by the bidirectional encoder representations from transformers (BERT) [16] to disambiguate the ambiguous mentions. The same framework is further implemented to continuously enrich the coverage of the proposed PBA. The performance of the developed method is assessed on our code-mixed EHR de-identification corpus, which is an extended version of our previous work [17], and compare with that of the state-of-the-art methods based on pre-trained language models.

II. MATERIALS AND METHODS

A. DATA SOURCE AND ANNOTATION GUIDELINE

With the approval of the research ethics committee of the National Health Research Institutes (EC1090212-E) and National Taiwan University Hospital (NTUH-201610072RINA), the EHR data collected from the NTUH-Integrated Medical Database was used in this study. We extended the annotation guideline presented in our previous work [17] to define 6 main categories and 25 subcategories of PHIs listed in Table 1 that need to be de-identified. Four annotators (P.C., B.W., C.S., and W.L.) followed the guideline to annotate the PHIs and compile a code-mixed de-identification dataset containing a total of 1,700 EHRs. We randomly sub-sampled 1,500 summaries as the training set, and the remaining 200 summaries were used as the test set.

B. PRINCIPLE-BASED DE-IDENTIFICATION METHOD

The proposed PBA provides an integrated solution for PHI recognition and resynthesis. This method relies on supervised learning approaches to induce principles from our training corpus and uses the compiled principles to recognize and resynthesize PHI mentions.

In the training phase of the PBA, we start by constructing the principle knowledge for the task of de-identification. A hierarchical slot-based schema was used to express the

[1] Available at https://ckip.iis.sinica.edu.tw/demo.

compiled knowledge. Following the construction, during the principle generation step defined slots were labeled on the training set and automatically assembled and summarized into principles by observing the arrangement of slots that can be used to recognize PHIs. Principles generated during the training phase are represented in a human-interpretable manner that can be further updated by annotators or domain experts.

In the test phase, a given input text is labeled with the defined slots, and PBA utilizes a fuzzy matching algorithm along with the learned principles to recognize and resynthesize PHIs. In the following subsections, we elaborate the proposed PBA for the de-identification task.

C. DEFINITIONS AND STRUCTURE OF THE PRINCIPLE-BASED APPROACH

We define the term ''principle'' in terms of the de-identification task as follows.

A ''principle'' refers to an organized semantic pattern of a type of PHI.

Each principle consists of a collection of ''slots'' and ''relations''. A slot serves as the basic component that holds a piece of information in a particular principle which may contain a set of words, phrases, semantic categories, regular expressions, or other slots. One can specify the ordering and compatibility among slots as a relation in a principle. Any combination of slots and relations is referred to as ''knowledge'' in the PBA.

Figure 2 exhibits a snapshot of the compiled knowledge which was constructed and visualized in Information Map (InfoMap) [18]. Figure 2.a illustrates how the knowledge is constructed for representing the PHI type of ''Date''. The principle slot scheme in InfoMap is represented in a tree structure which covers the knowledge required for de-identification. The entire hierarchy includes a bunch of subtrees each of which defines the slots and the relations among them for the corresponding PHI type.

The nodes in the first level of the hierarchy are referred to as the concept slots. A concept slot is responsible for recognizing and resynthesizing a PHI type. Each concept slot defines its fundamental components under the ''HAS-PART'' node which contains definitions of sub-slots for capturing characteristics shared among different PHI mentions. For instance, the term ''2021年11月16日'' consists of slots including year (2021), month (11), date (16), and unit ((年, 月and 日). The co-occurrence of these sub-slots in a text indicates the appearance of a ''Date'' mention. Hence, we can define sub-slots like ''[unit]'', ''[Date]'', ''[Month]'', and ''[Year]'' which can be used to develop indicating words or regular expression patterns that match a given text to capture the occurrences of these slots. In our implementation, the regular expressions or terms that could serve as indicating words were compiled based on the training set.

D. PRINCIPLE GENERATION FOR PHI RECOGNITION AND RESYNTHESIS

The INSTANCE node shown in Figure 2.a contains definitions of principle instances composed of one to multiple slots arranged in a certain order that can be used by a matcher (described in the next subsection) to recognize PHIs. Take the last instance of Date in Figure 2.a as an example. The principle is represented as ''[Month_num]:[unit]:[Date]:[unit]:[Year]'', indicating that a Date mention can be made up of the sub-slots in this arrangement.

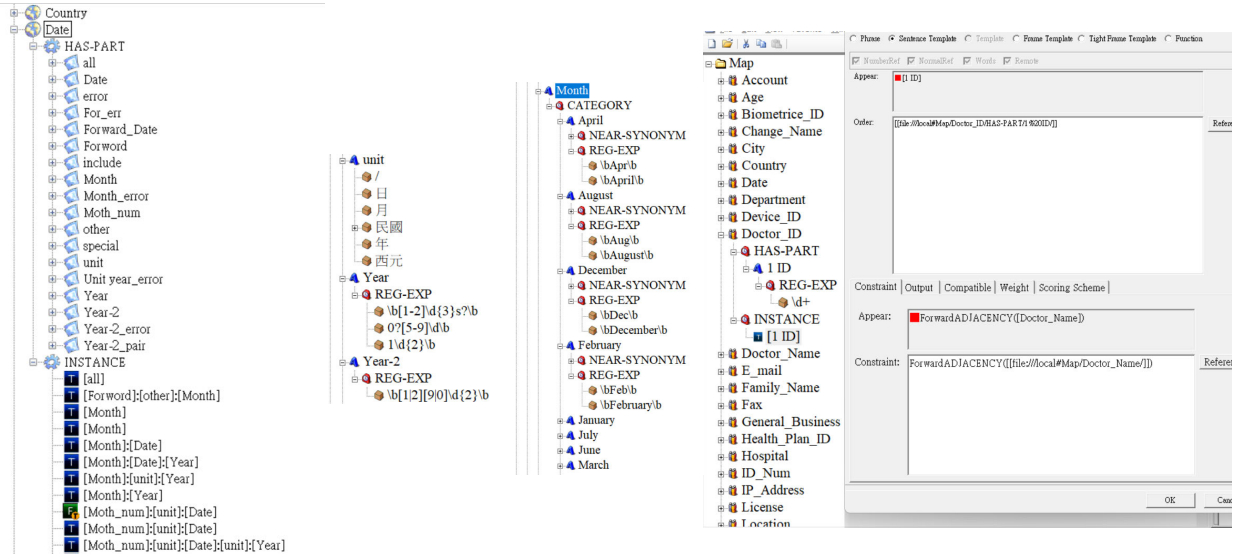
The principle instances for each PHI type can be manually defined by knowledge engineers or semi-automatically learned from a given training dataset. To generate PHI principles from our dataset, split sentences were first labeled with the defined slots. Unlabeled words in a sentence were considered as insertions, and the remainder of the labeled sequence was regarded as a candidate principle for representing PHIs. Subsequently, several instances of slot combinations were generated. The dominating set-based algorithm developed in our previous study [15] was then used to summarize all candidate principles into representative instances in InfoMap. The algorithm considers all slot definitions including concept slots and sub-slots. Consequently, the created principle instances could include other slot definitions as shown in Figure 2.d, where the ''Doctor_ID'' concept slot includes the ''Doctor Name'' concept slot as its sub-slot which is used as a constraint for recognizing the PHI type. The hierarchical slot schema used in InfoMap is visualized by our knowledge editor, COMPASS [19], which enables knowledge engineers to easily update the compiled knowledge and interpret the prediction results.

E. PRINCIPLE MATCHING FOR PHI RECOGNITION AND RESYNTHESIS

In the principle matching stage, we applied the proposed method to recognize PHIs and generate the corresponding surrogates for resynthesizing the content. Our system accepts a batch of EHRs for a patient arranged in a specific order as its input. This arrangement is required to ensure the temporality of the synthesized results for a patient. The output is a list of surrogates and their corresponding offsets in the synthesized text.

1) PHI RECOGNITION

During the process for identifying PHIs mentioned in the discharge summary, the given sequence of words was first labeled with the compiled slots. We then employed an alignment-like algorithm proposed in our previous study [15] to determine the span of words that matches the principles defined in InfoMap. Take the first sentence ''he was brought to中國醫 for help in 2095/8…'' as an example. Sub-slots like ''中國醫'', ''2095'', ''/'', and''8'' were labeled for matching with the principles defined for each PHI type. Among them, ''2095'', ''8'', and ''/'' were mapped to [Year], [Month], and



(a) (b) (c) (d)

FIGURE 2. Knowledge represented for Date in InfoMap. Figure 2.a displays the principle slot schema hierarchy defined for Date. Figure 2.b and 2.c show the slots defined for Date described in Chinese and English, respectively. Figure 2.d illustrates a slot definition which includes another concept slot as a constraint.

TABLE 1. Categories and subcategories of de-identified PHI annotations.

PHI category	Subcategory	Type	Example
Date	None	Numerical	2090/04/17, 7/02, 1190220, MK120
Age	None	Numerical	13 years old, 38 歲
Names	Patient, Person, Family, Doctor	Non-numerical	Ho, 黃大銘, 葉醫師
Locations	Location, Nationality, Region, State, Country, City, Street, Hospital, Department, Room, Number, School, General Business, Market	Non-numerical	華僑, east-south asia, Utah, 台灣, 高雄, 忠孝東路, CMUH, 腸胃科, 12 診, 27 號, 高雄科技大學, 711
Profession	None	Non-numerical	editor, 技師
IDs	ID Number, Medical Record, Phone, NIC	Non-numerical	03w20702, 32001CXM, 分機 **2234**

[Unit], respectively. However, "中國醫" displays a semantic ambiguity because it could be mapped to either [Hospital] or [School] in the PBA.

To address the ambiguity problem, we first estimated the representative vector for each PHI type on the training set by using a pre-trained language model. When the system encounters an ambiguous PHI during the recognition process, we used the same pre-trained model to generate a representative vector for the ambiguous term and compared the vector with the estimated representative vector for each PHI type by the cosine similarity defined in Equation 1. The PHI type of the closest representative vector is then selected as the type of the ambiguous term to solve the semantic ambiguity.

$$Similarity\,(A,B) = \frac{A \cdot B}{\|A\|\,\|B\|} = \frac{\sum_{gg=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \times \sqrt{\sum_{i=1}^{n}(B_i)^2}} \quad (1)$$

In our implementation, the representative vector is constructed by concatenating the text representations of the ambiguous term and its surrounding context. A text representation is calculated by the average of the word representative vectors generated by the pre-trained language model as defined in Equation 2, where $wr_{i,d=1}$ indicates the value of the first dimension of the ith word representation, and n refers to the number of words of the ambiguous term or context.

$$\text{TextRepresentation}([wr_1, wr_2, \ldots, wr_n]) = \frac{1}{n}\left[\sum_{i=1}^{n} wr_{i,d=1}, \ldots, \sum_{i=1}^{n} we_{i,d=j}\right] \quad (2)$$

2) PHI RESYNTHESIS

As shown in Table 1, the PHI categories can be classified as a numerical or non-numerical type. Both types of PHIs can be described in Chinese or English, which can be determined by the matched slots. In the propose PBA, the corresponding

code-matched surrogates are synthesized for the recognized PHIs with the assistance of the compiled principle knowledge to maintain the original code-mixing grammatical structure of the original sentence.

For numerical types of PHIs including dates and ages, we used the offset displacement to generate the surrogates. The offset can be manually specified or randomly generated. The dates and ages in the EHRs of the same patient were shifted with the same offsets to maintain temporality. However, we came across two major challenges when generating the surrogates for both categories. First of all, the numeric values of both categories could be in either Chinese or English with various forms including different calendars. For example, "2021/12/25" could appear as "2021 Dec. 25", "110.12.25", or even "MK110十二月25日" in the Minguo calendar. The ambiguity of date strings caused by different date notation styles leads to the second challenge of value interpretation in various date formats. For example, a date string "23/04/05" could be in the format of "YY [both Minguo or Christian calendar]/MM/DD", or "MM/DD/YY [both Minguo or Christian calendar]". Our system tracks all matched slots for a patient and applies the majority rule to determine the format in cases of ambiguities. The parsed dates and ages with a known format are then shifted by adding the specified time shifts.

When resynthesizing non-numerical types of PHIs, we randomly generate surrogates from the INSTANCE node defined for the same PHI category in the defined principle knowledge. The principle instance with the largest number of matched slots is used for the generation. In addition, if the same PHI instance occurs in the following EHRs of the same patient, the PHI will be replaced with the same surrogate to preserve the integrity of the context. Take the patient name "張三" in Figure 1 as an example. The same surrogate such as "李四" will be synthesized for all occurrences of the patient name. Furthermore, the compiled principle knowledge was designed to preserve the gender information by defining separate slots which include the first names and surnames for males and females in Chinese or English.

F. KNOWLEDGE AUGMENTATION

In order to continuously expand the knowledge encoded in the PBA, an augmentation method is developed to distill the knowledge earned from processing texts. We trained a bidirectional long short-term memory conditional random fields (BiLSTM-CRF) model [17], [20] on our training set. The same pre-trained language model used for disambiguation was applied to represent words as contextual embeddings. The contextual features went through the BiLSTM layer and were finally decoded by the CRF layer for recognizing PHIs.

During the de-identification process, the PHIs recognized by the BiLSTM-CRF model but not recognized by the PBA were extracted. The disambiguation method mentioned in Section II.E.1 was employed to confirm the type of each PHI, which is added to the corresponding concept slot of InfoMap if the similarity score is higher than a threshold estimated in the training phase of the augmentation step. The knowledge engineers (F. W. in this study) can further review the acquired slots to determine whether to accept the update.

G. EXPERIMENTAL SETTINGS AND MODEL CONFIGURATIONS

Three pre-trained BERT-based models were used in our experiments, including the original BERT (BERT-base-cased), the Chinese BERT (BERT-base-Chinese), and the multilingual BERT (BERT-base-multilingual-cased). For the pre-trained BERT model used by the proposed PBA, the multilingual BERT was used for the estimation of the representative vector for each token in the disambiguation process. The model was also used as the feature extractor for representing the given text in the knowledge augmentation process.

We followed our previous work [17] to establish the PHI recognition baseline systems, in which the task was formulated as a sequential labeling task with the adoption of the BILOU encoding schema. A linear layer was added to transform the output of the adapted BERT model to meet the expected numbers of BILOU tags. We did not apply Chinese word segmentation because the adapted models were pre-trained without applying whole word masking. The whole layers and parameters of the pre-trained multilingual BERT were fine-tuned with the AdamW optimizer by setting the number of iterations and maximum sequence length to 20 and 512, respectively. For the proposed PBA, we adopted a naïve disambiguation method for comparison. The disambiguation baseline was implemented by sampling from the distribution of the number of PHI subcategories.

PHI recognition results are reported in terms of micro- and macro-F-scores defined as follows. TP (true positive) indicates the number of correctly recognized PHI mentions whose category and span exactly match the ground truth. FP (false positive) refers to the number of predicted PHI mentions with category or span mismatches against the ground truth. FN (false negative) is the number of manually annotated PHIs that cannot be recognized.

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}; \quad \text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

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

We used the accuracy defined below as the performance indicator for the evaluation of the proposed disambiguation method. The calculations of TP/TN/FP/FN only take the category information of each PHI mention into consideration.

$$\text{Accuracy} = \frac{TP + TN}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

III. RESULTS

A. STATISTICS OF THE CODE-MIXING DE-IDENTIFICATION CORPUS

We randomly sub-sampled 1,500 summaries from 1,700 EHRs as the training set, with the remaining 200 summaries as the test set. The training and test sets contained 297,621

TABLE 2. Distribution of the PHI types in the code-mixed training and test sets. (* indicates the sentence is code-mixed; ENG and CHI means the sentence is monolingual in English and Chinese, respectively.)

PHI Category	ENG		CHI		ENG-CHI*		ALL	
	Train	Test	Train	Test	Train	Test	Train	Test
Date	**32,195**	**6,059**	**2,138**	**457**	**4,649**	**932**	**38,982**	**7,448**
Age	**2,442**	**532**	**19**	**5**	**641**	**126**	**3,102**	**663**
Names	**100**	**18**	**270**	**44**	**1,335**	**243**	**1,705**	**305**
Patient	9	9	106	12	40	14	155	35
Person	8	3	6	1	122	32	136	36
Family	0	0	0	0	1	0	1	0
Doctor	83	6	158	31	1,172	197	1,413	234
Locations	**6,021**	**976**	**4,790**	**1,106**	**5,719**	**1,092**	**16,430**	**3,172**
Location	17	9	11	2	271	67	299	76
Nationality	6	4	1	2	8	1	15	7
Region	8	10	0	0	11	1	19	11
State	2	2	0	0	1	0	3	2
Country	360	53	8	5	184	54	552	112
City	179	28	20	5	430	77	629	110
Street	0	0	0	0	2	0	2	0
Hospital	1,533	260	71	10	1,916	397	3,520	667
Department	2,706	390	2,276	542	1,285	259	6,267	1,501
Room	955	172	1,199	267	239	51	2,293	490
Number	0	0	1,161	256	26	1	1,187	257
School	31	6	14	1	636	151	681	158
General Business	217	39	29	16	692	161	938	216
Market	7	3	0	0	18	19	25	22
Profession	**524**	**86**	**66**	**9**	**1,364**	**397**	**1,954**	**492**
IDs	**60**	**0**	**192**	**25**	**197**	**5**	**449**	**30**
ID Number	50	0	98	11	170	2	318	13
Medical Record	1	0	92	13	19	3	112	16
Phone	9	0	2	1	7	0	18	1
NIC	0	0	0	0	1	0	1	0
Numbers of PHIs	41,343	7,671	7,439	1,644	13,905	2,795	62,722	12,110

and 60,633 sentences, respectively. Table 2 displays the statistics of the 25 PHI types in the compiled datasets. The distribution of PHI types in the training set and the test set are similar and some categories have highly imbalanced annotations in their subcategories. The top 3 PHI categories are ''Date'', ''Locations'', and ''Age'', which account for approximately 62.15% (38,982/62,722), 26.19% (16,430/62,722), and 4.95% (3,102/62,722) of all annotations, respectively. Some subcategories such as State, Family, Street, and NIC have less than 20 annotations. Since the last three PHI types have no annotations in the test set, we excluded them in the following performance evaluations. As one can observe in Table 2, the PHIs in our corpus could be described in monolingual (English or Chinese) and bilingual (mixture of English and Chinese) sentences. Therefore, we applied the code-mixing index (CMI) defined in a previous study [4] to measure the mixing level of our corpus. We calculated the CMIs for English and Chinese separately with sentences dominated by each language and reported the results in Table 3. The code-mixed sentences occupy 61.5% of the data in our corpus, and the average CMIs for the training and test sets are 3.22% and 18.98% for English and Chinese, respectively. Compared with the CMI values in the previous work [4], our average CMI of 11.1 is higher and brings forth additional challenges in the task of de-identification.

B. PRINCIPLE-BASED DE-IDENTIFICATION METHOD

In this experiment, we evaluated the effectiveness of the proposed semantic disambiguation process (SDP) and knowledge augmentation process (KAP). For the SDP described in Section II.E.1, we inspected its capability of disambiguation with a 3-fold cross validation (CV) on the training set. In the training set of each fold, we collected the sentences containing PHIs. For each PHI mentioned in a sentence, the representative vector was estimated by using Equation 2 along with the multilingual BERT for updating the representation of the corresponding PHI type. The model was selected as Lee, *et al.* [17] has demonstrated that it outperformed the original or Chinese BERTs when processing texts containing mixtures of Chinese-English descriptions. After compiling the representations for all PHI subcategories, we applied Equation 1 to disambiguate the PHIs mentioned in each sentence of the test set in the corresponding fold and compared the disambiguated results against the gold annotations.

TABLE 3. The code-mixing index for the training and test sets. ENG indicates that the English words dominate these sentences, while CHI indicates that Chinese words are more frequently observed in these sentences.

	ENG. CMI	CHI. CMI	Number of ENG Sentences	Number of CHI Sentences
Training	3.25%	18.95%	182,649	44,859
Test	3.18%	19%	37,699	9,112
Overall	3.22%	18.98%	220,348	53,971

TABLE 4. The effectiveness of the semantic disambiguation process (SDP) and knowledge augmentation process (KAP) for the PHI recognition task on the test set. Note that the macro- and micro-avg. scores were estimated by considering all PHI categories.

Ambiguous PHI Type	Baseline			+SDP			+KAP			+SDP+KAP		
	P	R	F	P	R	F	P	R	F	P	R	F
Names	**0.97**	0.37	0.52	-0.32	+0.18	+0.07	-0.07	+0.18	+0.16	-0.23	**+0.26**	**+0.16**
Patient	**1.00**	0.23	0.37	±0	±0	±0	±0	**+0.17**	**+0.20**	±0	**+0.17**	**+0.20**
Person	**0.91**	0.28	0.43	-0.57	+0.19	-0.03	-0.02	+0.16	**+0.14**	-0.42	**+0.36**	+0.12
Doctor	**0.99**	0.61	0.76	-0.06	+0.17	+0.09	-0.12	+0.22	+0.12	-0.04	**+0.26**	**+0.15**
Locations	**0.81**	0.59	0.63	±0	±0	±0	-0.06	+0.15	**+0.10**	-0.06	**+0.15**	+0.09
Location	**0.94**	0.20	0.33	±0	±0	±0	-0.01	**+0.31**	**+0.33**	-0.01	**+0.31**	**+0.33**
Nationality	**1.00**	0.29	**0.44**	±0	±0	±0	-0.57	**+0.14**	-0.01	-0.57	**+0.14**	-0.01
City	0.68	0.77	0.72	±0	±0	±0	**+0.13**	+0.04	**+0.09**	+0.11	**+0.05**	+0.08
Hospital	0.92	0.81	0.86	±0	±0	±0	**±0.01**	**±0.05**	**±0.03**	**+0.01**	**+0.05**	**+0.03**
Department	0.90	0.80	**0.85**	±0	±0	±0	**-0.04**	**+0.04**	±0	-0.04	**+0.04**	+0.00
Room	**0.92**	0.72	0.81	±0	±0	±0	-0.06	**+0.16**	**+0.06**	-0.06	**+0.16**	**+0.06**
School	**0.84**	0.58	0.69	±0	±0	±0	-0.05	**±0.22**	**±0.11**	-0.05	**+0.22**	**+0.11**
General Business	0.55	0.67	0.60	±0	+0.02	+0.01	**+0.09**	+0.15	+0.12	**+0.09**	**+0.17**	**+0.13**
Market	0.89	0.77	0.83	±0	±0	±0	**+0.02**	**+0.18**	**+0.10**	**+0.02**	**+0.18**	**+0.10**
Profession	0.73	0.63	0.67	±0	±0	±0	**+0.05**	**+0.12**	**+0.10**	**+0.05**	**+0.12**	**+0.10**
Micro-avg.	0.91	0.86	0.88	+0.01	+0.01	+0.02	**+0.01**	**+0.04**	**+0.03**	**+0.02**	**+0.04**	**+0.04**
Macro-avg.	0.72	0.57	0.58	**+0.08**	+0.09	+0.10	+0.07	+0.17	+0.16	+0.07	**+0.18**	**+0.16**

The accuracy of SDP estimation on the CV and test set is 0.940 and 0.883, respectively.

When more than one principle is matched during PHI recognition, the SDP is applied to disambiguate the PHI candidate by comparing its representation to the pre-calculated representations of matched subcategories under the same PHI category. Although there are only three PHI categories, ambiguity has proven to be a major issue for ''Locations'', ''Profession'', and ''Names'' in the current training set. Nevertheless, we still generated the representative vectors for all subcategories except ''Age'', ''Date'', ''Medical Record'', and ''NIC'' in order to support the KAP for the extraction of augmented candidates. The threshold for accepting an augmentation is set to 0.55, which was estimated on the same CV set during the development of the proposed method.

Table 4 presents an ablation study of the PHI recognition performance on the test set, where only PHI categories with improvement were reported. One can observe that the KAP made a greater contribution to the overall performance improvement than the SDP. Sole application of SDP only enhanced the recall of the ''Names'' category. However, as described in Section II.F, the KAP relied on the SDP to determine the target concept slot for augmentation, so the SDP is an indispensable component to facilitate the KAP in the proposed method. Furthermore, combination of the SDP and KAP achieved the highest recall for the ''Names'' category which is the most critical PHI type that must be de-identified in the real world hospital setting.

C. DE-IDENITFICATION PERFORMANCE COMPARISON

We compared the performance of the proposed method (PBA+SDP+KAP) with that of the BERT-based models in Table 5. The configuration settings described in Section II.G were applied. We can see that the original BERT model have comparable F-scores on PHIs with numerical type such as ''Date'' and ''Age''. However, it underperformed on

TABLE 5. Performance of the developed methods on the test set. The values in bold indicate the highest score for the PHI type.

PHI Category	BERT-base			BERT-Chinese			BERT-Multilingual			PBA+SDP+KAP		
	P	R	F	P	R	F	P	R	F	P	R	F
Date	0.98	**0.98**	**0.98**	0.97	**0.98**	0.98	0.97	**0.98**	0.98	**0.99**	0.94	0.96
Age	**0.95**	0.97	0.96	0.94	**0.98**	**0.96**	0.94	**0.98**	**0.96**	0.83	0.90	0.87
Names	0.59	0.44	0.50	**0.82**	**0.65**	**0.73**	0.82	0.63	0.71	0.81	0.63	0.71
Patient	0.76	0.37	0.5	**1.00**	**0.51**	**0.68**	0.94	0.46	0.62	**1.00**	0.40	0.57
Person	0.14	0.06	0.08	**0.50**	0.53	0.51	0.50	0.44	0.47	0.49	**0.64**	**0.55**
Doctor	0.88	0.88	0.88	0.95	**0.92**	**0.94**	0.94	**0.92**	0.93	**0.95**	0.87	0.91
Locations	0.46	0.45	0.45	0.64	0.62	0.63	0.59	0.58	0.58	**0.71**	**0.64**	**0.67**
Location	0.19	0.16	0.18	0.52	0.57	0.54	0.49	0.51	0.50	**0.93**	**0.51**	**0.66**
Nationality	0.00	0.00	0.00	0.33	0.14	0.20	0.50	0.14	0.22	**0.43**	**0.43**	**0.43**
Region	0.00	0.00	0.00	0.50	0.36	0.42	0.20	**0.09**	0.13	**1.00**	**0.09**	**0.17**
State	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Country	0.73	0.69	0.71	**0.92**	0.94	0.93	0.85	0.94	0.89	0.85	**0.96**	**0.90**
City	0.52	0.51	0.52	**0.78**	0.65	0.71	0.75	0.67	0.71	0.79	**0.82**	**0.80**
Hospital	0.81	0.82	0.81	0.89	0.90	0.90	0.90	**0.91**	**0.90**	**0.93**	0.86	0.89
Department	**0.90**	0.87	**0.88**	0.62	0.88	0.73	0.63	**0.89**	0.74	0.86	0.84	0.85
Room	0.83	0.91	0.87	0.82	**0.94**	0.88	0.83	0.93	**0.88**	**0.86**	0.88	0.87
Number	**1.00**	**1.00**	**1.00**	1.00	0.99	1.00	0.99	**1.00**	1.00	**1.00**	0.99	1.00
School	0.60	0.64	0.62	0.83	0.87	0.85	**0.84**	**0.90**	**0.87**	0.79	0.80	0.80
General Business	0.38	0.30	0.33	0.65	0.73	0.69	0.60	0.69	0.64	**0.64**	**0.84**	**0.73**
Market	0.00	0.00	0.00	**0.5**	0.14	0.21	0.75	0.41	0.23	0.91	**0.95**	**0.93**
Profession	0.37	0.39	0.38	0.64	0.73	0.68	0.61	0.73	0.67	**0.78**	**0.75**	**0.77**
IDs	0.54	0.95	0.69	0.54	0.97	0.69	0.52	0.95	0.67	**0.81**	**1.00**	**0.90**
ID Number	0.55	0.85	0.67	0.60	0.92	0.73	0.63	0.92	0.75	**1.00**	**1.00**	**1.00**
Medical Record	**0.94**	**1.00**	**0.97**	**0.94**	**1.00**	**0.97**	**0.94**	**1.00**	**0.97**	**0.94**	**1.00**	**0.97**
Phone	0.12	**1.00**	0.22	0.09	**1.00**	0.17	0.11	**1.00**	0.20	**0.50**	**1.00**	**0.67**
Micro-avg.	0.90	0.90	0.90	0.89	**0.94**	0.91	0.89	0.94	**0.91**	**0.93**	0.90	0.92

non-numerical types such as ''Profession'' and ''Names''. In particular, if we perform a comprehensive review of the results in Table 2 and 5, it is notable that the model performed significantly worse than others for PHIs under the subcategories that were frequently mentioned in the code-mixed sentences and with less than one thousand corresponding training instances. For example, the F-scores of the original BERT model for ''School'' and ''City'', which were more frequently mentioned in code-mixed sentences as shown in Table 2, are apparently lower than that of the other two BERT models. On the other hand, ''Doctor'' and ''Number'' also frequently existed with code-mixed context, but both have comparable F-scores which may be owing to the quantity of the training instances and the characteristics of these two PHI types.

By contrast, the micro-avg. R of the proposed PBA is slightly lower than that of BERT-Chinese and BERT-multilingual, but it achieved the best micro-avg. P and out-performed the others in terms of the macro-avg. PRF. Furthermore, the PBA obtained F-scores that are at least 10% higher than that of the others on three out of six PHI categories with comparable performance on the ''Names'' and ''Date'' categories, demonstrating the effectiveness of the proposed method.

TABLE 6. Types of medical records processed by the proposed system in the preliminary results.

Types of Medical Records	Training	Test
Colonoscope Ultrasound	70	40
Kidney Ultrasound	60	40
Thoracic Ultrasound	60	40
Chest X-ray	120	40
Thoracic MRI	50	40
Chest CT	90	40
Nuclear Medicine	50	40
Total	500	280

IV. DISCUSSION

A. EFFECTIVENESS OF THE PROPOSED DISAMBIGUATION METHOD

The proposed disambiguation method is an important factor not only for semantic disambiguation but also the knowledge augmentation processes. Therefore, we took a step further to analyze the methodology and results of disambiguation. First, we investigated the importance of the surrounding context for an ambiguous PHI. In our proposed implementation, the representative vector is constructed by concatenating the text representations of the ambiguous term and its surrounding context. We observed that if only the word representations of the ambiguous PHI were used for disambiguation, it is difficult to disambiguate cases such as ''Informant: Discharge note from 八里'', where ''八里'' refers to the ''八里 (Bali) psychiatric center''. Since ''八里(Bali)'' is always referred to as ''Bali district'' in our training set, the representation vector would likely be categorized as ''City'' if we exclude the surrounding context.

Although our investigation demonstrated that the proposed method can successfully exploit the contextual information in the given sentence to disambiguate a target PHI, we noticed that the method still failed in disambiguating PHIs represented in a manner of very limited local surrounding context without further considering the global context. This issue can be illustrated by the two doctor names in Figure 1. After examining the entire record, a human can interpret the last three lines of the description in a semi-structured format to understand that the first name (王五) refers to a resident doctor (represented by R), while the second person (趙六) is the attending doctor represented by VS (Visiting Staff). However, the input sentence for our de-identification system is ''2105.12.15 王五 趙六'' due to the separated line, which leads the proposed method to disambiguate both PHIs to ''Person''.

B. PRELIMINARY RESULTS WHEN THE DEVELOPED SYSTEM IS APPLIED IN THE HOSPITAL SETTING

The developed system was introduced to the China Medical University Hospital (CMUH) for the de-identification of seven types of records listed in Table 6. Different from our corpus, only six PHI types including ''City'', ''Doctor'', ''Phone'', ''Hospital'', ''ID Number'', and ''Medical Record ID'' were found in 780 records. We extended the knowledge defined in the PBA by training it on the training set of 500 records. In addition, we directly included the doctor and staff names listed on the CMUH web page as an additional INSTANCE node to enhance the capability to identify CMUH medical staffs. In comparison to applying the pre-trained language model, implementing the proposed PBA demonstrated high customizability and achieved satisfactory micro-PRF-scores of 1.00, 0.986, and 0.993, respectively.

V. CONCLUSION

In this work, we presented our PBA for the tasks of PHI recognition and resynthesis. Unlike previous studies which mainly focused on de-identifying PHIs in monolingual corpora, we applied our method on a code-mixed EHR corpus and compared its performance with that of the state-of-the-art pre-trained mono- and multilingual language models. The experiment results show that the proposed PBA along with the SDP and KAP can achieve decent micro- and macro- F-scores of 0.92 and 0.74, respectively. When compared to the pre-trained BERT-based models that can only be applied for the PHI recognition task, the proposed method can further generate surrogates for the recognized PHIs based on the same framework. Preliminary results of the developed system working in a cross-hospital setting demonstrate the robustness and customizability of the proposed method. Although the code-mixed descriptions seem to have less impact on the recognition performance of the proposed PBA, BERT-Chinese and BERT-Multilingual models, all of them still performed apparently worse on the recognition of the most critical PHI category ''Names''. Moving forward, we will study the re-identification risk for our PHI resynthesis method continue to investigate the challenge the recognition of code-mixed PHIs mentioned in the code mixed context and improve the proposed PBA in the real world hospital setting environment.

ACKNOWLEDGMENT

The authors would like to thank the staff of the Department of Medical Research, National Taiwan University Hospital for the Integrated Medical Database (NTUH-IMD). *(Chen-Kai Wang and Feng-Duo Wang contributed equally to this work.)*

REFERENCES

[1] *Health Information Technology for Economic and Clinical Health Act*, United States Congr., United States Dept. Health Hum. Services, Washington, DC, USA, 2009.

[2] F. Toscano, E. O'Donnell, M. Unruh, D. Golinelli, G. Carullo, G. Messina, and L. Casalino, ''Electronic health records implementation: Can the European Union learn from the United States?'' *Eur. J. Public Health*, vol. 28, no. 4, p. 2, Nov. 2018.

[3] H.-C. Wen, W.-P. Chang, M.-H. Hsu, C.-H. Ho, and C.-M. Chu, ''An assessment of the interoperability of electronic health record exchanges among hospitals and clinics in Taiwan,'' *JMIR Med. Informat.*, vol. 7, no. 1, Mar. 2019, Art. no. e12630.

[4] 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 *Proc. IEEE World Congr. Services (SERVICES)*, vol. 2642, Jul. 2019, pp. 283–288.

[5] P. Voigt and A. Von dem Bussche, *The EU General Data Protection Regulation (GDPR): A Practical Guide*, vol. 10, 1st ed. Cham, Switzerland: Springer, 2017, Art. no. 3152676.
[6] *Health Insurance Portability and Accountability Act of 1996, Accountability Act, Public Law 104-191*, United States Dept. Health Hum. Services, Washington, DC, USA, 1996.
[7] B. Gambäck and A. Das, "On measuring the complexity of code-mixing," in *Proc. 11th Int. Conf. Natural Lang. Process.*, Goa, India, 2014, pp. 1–7.
[8] H. Ndebele, "A socio-cultural approach to code-switching and code-mixing among speakers of IsiZulu in KwaZulu–Natal: A contribution to spoken language corpora," M.S. thesis (in African), Howard College Campus, Univ. Kwazulu-Natal, Berea, South Africa, 2012.
[9] M. Obrocka, C. Copley, T. Gqaza, and E. Grant, "Prevalence of code mixing in semi-formal patient communication in low resource languages of South Africa," 2019, *arXiv:1911.05636*.
[10] L. Nurhasanah, "Analysis of code mixing in A1 class of English education study program 2016," *Project, Prof. J. English Educ.*, vol. 3, no. 6, pp. 697–702, 2020.
[11] N. I. Wood, "Departing from doctor-speak: A perspective on code-switching in the medical setting," *J. Gen. Internal Med.*, vol. 34, no. 3, pp. 464–466, Mar. 2019.
[12] Ö. Uzuner and A. Stubbs, "Practical applications for natural language processing in clinical research: The 2014 i2b2/UTHealth shared tasks," *J. Biomed. Informat.*, vol. 58, pp. S1–S5, Dec. 2015.
[13] R. Yeniterzi, J. Aberdeen, S. Bayer, B. Wellner, L. Hirschman, and B. Malin, "Effects of personal identifier resynthesis on clinical text de-identification," *J. Amer. Med. Inform. Assoc.*, vol. 17, no. 2, pp. 159–168, Mar. 2010.
[14] Y.-C. Chang, C.-H. Chu, C. C. Chen, and W.-L. Hsu, "Linguistic template extraction for recognizing reader-emotion," *Int. J. Comput. Linguistics Chin. Lang. Process.*, vol. 21, no. 1, pp. 29–50, Jun. 2016.
[15] N.-W. Chang, H.-J. Dai, Y.-L. Hsieh, and W.-L. Hsu, "Statistical principle-based approach for detecting miRNA-target gene interaction articles," presented at the IEEE 16th Int. Conf. Bioinf. Bioeng. (BIBE), Taichung, Taiwan, Nov. 2016.
[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[17] Y.-Q. Lee *et al.*, "Protected health information recognition of unstructured code-mixed electronic health records in Taiwan," in *Proc. MedInfo*, 2021. [Online]. Available: https://dblp.org/db/conf/medinfo/medinfo2019.html
[18] W.-L. Hsu, S.-H. Wu, and Y.-S. Chen, "Event identification based on the information map-INFOMAP," in *Proc. IEEE Int. Conf. Syst., Man Cybern. e-Syst. e-Man Cybern. Cyberspace*, vol. 3, Oct. 2001, pp. 1661–1666.
[19] S.-H. Wu, M.-Y. Day, T.-H. Tsai, and W.-L. Hsu, "FAQ-centered organizational memory," in *Knowledge Management and Organizational Memories*. Boston, MA, USA: Springer, 2002, pp. 103–112.
[20] Z. Dai, X. Wang, P. Ni, Y. Li, G. Li, and X. Bai, "Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2019, pp. 1–5.

CHEN-KAI WANG received the M.Sc. degree in biomedical informatics from Taipei Medical University, in 2018. He is currently pursuing the Ph.D. degree in computer science with the National Yang Ming Chiao Tung University (NYCU), Taiwan.

From 2016 to 2018, he was an Intern at the Institute of Information Science, Academia Sinica, Taiwan. He was a Research and Development Substitute Service (RDSS) at the Big Data Laboratory, Chunghwa Telecom Laboratories (CHTTL), Taiwan, from 2018 to 2021. He is also an Associate Researcher with the Advanced Technology Laboratory, CHTTL. His current research interests include natural language processing, artificial intelligence, bioinformatics, medical informatics, and machine learning.

Mr. Wang was a recipient of the Outstanding RDSS Award in 2020.

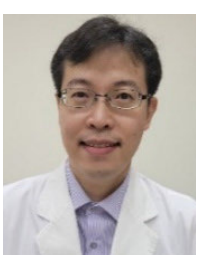
FENG-DUO WANG is currently pursuing the degree with the Department of Computer Science and Information Engineering, National Tsing Hua University, Hsinchu, Taiwan. He worked as an Artificial Intelligence Developer of the de-identification of electronic medical records project.

YOU-QIAN LEE was born in New Taipei City, Taiwan, in July 1998. He received the B.S. degree in engineering from the Department of Electrical Engineering, National Kaohsiung University of Science and Technology (NKUST), Taiwan, where he is currently pursuing the master's degree in electrical engineering. His research interests include electronic health records, code-switch analysis, and de-identification.

PEI-TSZ CHEN was born in Kaohsiung City, Taiwan, in May 2001. She is currently pursuing the bachelor's degree with the Department of Chemical Engineering, Feng Chia University, Taiwan. She is also working as a part-time Annotator with the Intelligent System Laboratory, National Kaohsiung University of Science and Technology, Kaohsiung.

BO-HONG WANG was born in Kaohsiung City, Taiwan, in June 1999. He received the B.S. degree from the Department of Electrical Engineering, National Kaohsiung University of Science and Technology, Taiwan. He is currently pursuing the master's degree in electrical engineering with National Sun Yat-sen University. His research interests include deep learning, natural language processing, image processing, 5G, and the AIoT communications networks.

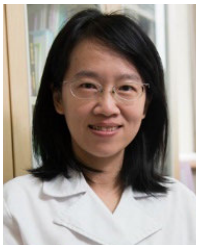
CHU-HSIEN SU received the master's degree from the Institute of Biomedical Informatics, National Yang Ming University, in 2014. From 2014 to 2017, he worked as a Research Assistant with Academic Sinica and the National Taiwan University Hospital. He is currently a Research Assistant with the National Center for Geriatrics and Welfare Research, National Health Research Institutes, Miaoli, Taiwan. His current research interests include using data analysis and natural language processing to handle geriatrics and psychiatry issues.



JOSEPH CHIN-CHI KUO received the M.D. degree (Hons.) from China Medical University, Taiwan, in 2003, the degree (Hons.) in nephrology from the National Taiwan University Hospital, in 2009, and the Ph.D. degree (Hons.) in epidemiology from Johns Hopkins University, in 2014. Since 2016, he has been leading the Big Data Center, China Medical University Hospital, Taiwan, and has built a smart data ecosystem and platform, the iHi (ignite Hyper-intelligence) platform, driving innovation in personalize medicine and digital

public health. He is currently a Physician-Epidemiologist. His current projects are focused on multi-domain data integration to gain mechanistic and translational data insight into clinical practice and drug discovery. His research investigates the effects of environment exposures and genetic variations on the development of kidney diseases using diverse analytical techniques. He is a member of Phi Beta Kappa and Sigma Xi Honor Societies. Since 2022, he serves as an Editorial Board Member for the *American Journal of Kidney Disease*.

CHI-SHIN WU is currently an Associate Investigator with the National Center for Geriatrics and Welfare Research, National Health Research Institutes, Miaoli, Taiwan. He is also an Attending Psychiatrist with the Department of Psychiatry, National Taiwan University Hospital, Yunlin Branch, Taiwan. He focuses on geriatric mental health and the risks and benefits of psychopharmacological treatment. He uses pharmacoepidemiologic approaches to investigate the therapeutic and adverse effects based on the administration database of the Taiwan's National Health Insurance Program, nationwide surveys, and electronic health records. His current research investigates the precision medicine of psychiatric treatment using machine learning. He also expands to natural language processing and text mining to integrate information from electronic health records for epidemiological studies.



YI-LING CHIEN received the Doctor of Medicine degree from Chang Gung University, in 2001, and the Ph.D. degree from the Graduate Institute of Clinical Medicine, National Taiwan University College of Medicine, in 2015.

She is currently a Clinical Associate Professor at the National Taiwan University School of Medicine. She also works as an Attending Psychiatrist of the Department of Psychiatry, National Taiwan University Hospital, Taipei, Taiwan. She devotes to the research of sensory characteristics, neurocognitive function, and intervention in children and adults with autism spectrum disorder and schizophrenia. She has also conducted intervention studies in group-based social skill training for autistic adults.

HONG-JIE DAI (Member, IEEE) received the Ph.D. degree in computer science from the National Tsing Hua University, in 2012.

From 2013 to 2015, he was an Assistant Professor at the Graduate Institute of Biomedical Informatics, Taipei Medical University (TMU). He was a Visiting Scholar at the Institute of Information Science, Academia Sinica, Taiwan, from 2013 to 2018. He worked as the Director of the Application Development Section, National Taitung University (NTTU) Library and Information Centre. From 2016 to 2017, he was an Associate Professor with the Department of Computer Science and Engineering, NTTU. Since 2019, he has been an Adjunct Associate Research Fellow with the National Institute of Cancer Research, National Health Research Insurance, Taiwan, and a Jointly Appointed Associate Professor with Kaohsiung Medical University, Taiwan. He became a Professor with the Department of Electrical Engineering, National Kaohsiung University of Science and Technology (NKUST), Taiwan, in 2021. He is currently the Director of the Department of Electrical Engineering, NKUST. He has authored book chapters for biomedical and clinical text mining and has published research articles in peer-reviewed journals on fields including natural language processing, bioinformatics, and clinical informatics. His current research interests include natural language processing, artificial intelligence, bioinformatics, medical informatics, machine learning, and software engineering.

Dr. Dai is a member of the Taiwan Association for Medical Informatics, the Taiwanese Association for Artificial Intelligence, and the Association for Computational Linguistics and Chinese Language Processing. He was a recipient of the Best Doctoral Dissertation Award for the Association for Computational Linguistics and Chinese Language Processing in 2012, the TMU Newly-Appointed National Science Council Research Fellow Award in 2013 and 2014, the NTTU Outstanding Research Merit Award in 2016 to 2018, and the NKUST Outstanding Research Merit Award in 2020 and 2021.

VINCENT S. TSENG (Fellow, IEEE) received the Ph.D. degree with major in computer science from, Hsinchu, Taiwan, in 1997. He is currently a Chair Professor with the Department of Computer Science, National Yang Ming Chiao Tung University (NYCU). After that, he was a Postdoctoral Research Fellow with the Computer Science Division, Electrical Engineering and Computer Science Department, University of California at Berkeley, Berkeley, CA, USA, during 1998–1999. He was the Founding Director of the Institute of Data Science and Engineering, NYCU (2017–2020), the Chair of the IEEE CIS Tainan Chapter (2013–2015), and the President of the Taiwanese Association for Artificial Intelligence (2011–2012). He also acted as the Director of the Institute of Medical Informatics, National Cheng Kung University, during 2008–2011. He has authored more than 380 research papers in refereed journals and conferences as well as 15 patents (held and filed). By Google Scholar, his publications have been cited by more than 12,000 times with H-index 57. His research interests include data mining, big data analytics, machine learning, biomedical informatics, and mobile and web technologies. He was a recipient of a number of prestigious awards, including ACM Distinguished Scientist Member (2019), Outstanding Research Award (2015 and 2019) by the Ministry of Science and Technology Taiwan, Outstanding I. T. Elite Award (2018), FutureTech Breakthrough Award (2018), and the K. T. Li Breakthrough Award (2014). He served as the chair/program committee member for a number of premier international conferences/institutions and has been the Steering Committee Chair for PAKDD, since 2020. He has been on the Editorial Board of a number of top journals, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, *IEEE Computational Intelligence Magazine*, and *ACM Transactions on Knowledge Discovery from Data*.

WEN-LIAN HSU (Life Fellow, IEEE) received the Ph.D. degree in operations research from Cornell University, Ithaca, NY, USA, in 1980.

He was a Tenured Associate Professor with Northwestern University, before joining the IIS, Academia Sinica, Taipei, Taiwan, as a Research Fellow, in 1989. He is currently a Chair Professor with the Department of Computer Science and Information Engineering, College of Information and Electrical Engineering, Asia University, Taichung, Taiwan, an Adjunct Research Fellow of the Institute of Information Science (IIS), Academia Sinica, and a Jointly Appointed Professor with the National Tsing Hua University, Hsinchu, Taiwan. He was involved in graph algorithms and he has applied similar techniques to tackle computational problems in biology and natural language. In 1993, he developed a Chinese input software, GOING, which has since revolutionized Chinese input on computer. He was the Director of the Bioinformatics Program, Taiwan International Graduate Program (TIGP), Academia Sinica, from 2003 to 2018, the Director of the IIS, Academia Sinica, from 2012 to 2018, and a Distinguished Research Fellow of the IIS, Academia Sinica, from 2008 to 2021. His current research interests include applying natural language processing techniques to understanding DNA sequences as well as protein sequences, structures, and functions and also to biological literature mining.

Dr. Hsu received the Research Initiation Award of the National Science Foundation in 1981, the Top ten Most Distinguished Chinese computer products in Taiwan of Intelligent Chinese Input Software in 1993, the Outstanding Research Award from the National Science Council in 1991, 1994, and 1996, the first K. T. Li Research Breakthrough Award in 1999, the IEEE Fellow in 2006, and the Teco Award in 2008. He was the President of the Artificial Intelligence Society in Taiwan, from 2001 to 2002, and the President of the Computational Linguistic Society of Taiwan, from 2011 to 2012.