Hindawi
Disease Markers
Volume 2023, Article ID 9836340, 1 page
https://doi.org/10.1155/2023/9836340

Retraction

Retracted: BERT-Based Clinical Name Entity Reorganization Model for Health Diagnosis

Disease Markers

Received 20 June 2023; Accepted 20 June 2023; Published 21 June 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope

(2) Discrepancies in the description of the research reported

(3) Discrepancies between the availability of data and the research described

(4) Inappropriate citations

(5) Incoherent, meaningless and/or irrelevant content included in the article

(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. D. Borole, A. V. Agrawal, M. Tesfayohanis, D. Thakkar, M. R. Abonazel, and F. A. Awwad, "BERT-Based Clinical Name Entity Reorganization Model for Health Diagnosis," *Disease Markers*, vol. 2022, Article ID 2297063, 12 pages, 2022.

Hindawi
Disease Markers
Volume 2022, Article ID 2297063, 12 pages
https://doi.org/10.1155/2022/2297063

Research Article

BERT-Based Clinical Name Entity Reorganization Model for Health Diagnosis

Yogini Dilip Borole,[1] **Anurag Vijay Agrawal**,[2] **Miretab Tesfayohanis**,[3] **Dhruv Thakkar**,[4] **Mohamed R. Abonazel**,[5] **and Fuad A. Awwad**[6]

[1]*G H Raisoni College of Engineering and Management, Wagholi, Pune, India*
[2]*E & ICT Academy, Department of Electronics and Communication Engineering, Indian Institute of Technology Roorkee, PIN-2476672, Roorkee, Uttarakhand, India*
[3]*Department of Information Technology, Ethiopia*
[4]*Department of Computer & Information Sciences, Temple University, USA*
[5]*Department of Applied Statistics and Econometrics, Faculty of Graduate Studies for Statistical Research, Cairo University, Giza, Egypt*
[6]*Department of Quantitative Analysis, College of Business Administration, King Saud University, Riyadh, Saudi Arabia*

Correspondence should be addressed to Miretab Tesfayohanis; miretab@dadu.edu.et

Received 2 August 2022; Revised 28 August 2022; Accepted 5 September 2022; Published 11 October 2022

Academic Editor: Vijay Kumar

The National Health and Family Planning Commission requires medical institutions to use the International Classification of Diseases (ICD) codes. However, due to many commonly used words in clinical disease descriptions, the direct mapping matching rate between the diagnosis names entered in the electronic medical records and the ICD codes is low. In this paper, based on the actual diagnostic data on the regional health platform, a disease term map incorporating standard terms was constructed. Specifically, based on the rule algorithm based on the components of the disease, a data-enhanced BERT (bidirectional encoder representation from transformers) upper and lower relationship recognition algorithm is proposed. Synonymous upper and lower relationships identify diseases, and the hierarchical structure is further integrated. In addition, a task assignment based on the association map of disease departments is also proposed. Methods were used for manual verification, and finally, 94,478 disease entities formed a large-scale disease term map, including 1,460 synonymous relationships and 46,508 hyponymous relationships. Evaluation experiments show that, based on the disease term map and clinical diagnosis, the coverage rate of diagnostic data is 75.31% higher than direct mapping coding based on ICD. In addition, using the disease term map to code diseases automatically will shorten the coding time by about 59.75% compared with manual coding by doctors, and the correct rate is 85%.

1. Introduction

With the continuous improvement of informatization in the medical field, medical research institutions in Europe and the United States have established a series of medical terminology databases, such as the systematized nomenclature of medicine-clinical terms (SNOMED-CT) [1], the unified medical language system (UMLS) [2], ICD-10 (international classification of diseases 10th revision) [3], and ICD-11 (international classification of diseases 11th revision) [4]. Among them, the National Health Commission of the People's Republic of China clearly requires all medical institutions to uniformly use the Chinese version of ICD-10 (ICD10) in the writing of medical records, which greatly promotes the standardization and standardized management of medical services.

However, when ICD10 is actually applied to clinical data, less than 20% can directly establish the mapping. There are two main problems: first, the diversity of disease name descriptions. For example, "urinary tract infection" is a common term in clinical diagnosis, but it is not included in the ICD10. The word is a synonym for "urinary tract infection"; the corresponding code in ICD10 is N39.0. Second, the granularity of common disease terms is finer. For example, "diabetes with ocular changes" cannot find a matching synonym in ICD10, only its hyponym "diabetes" can be found, and the corresponding code of diabetes in ICD10 is E14.900. Therefore, using ICD10 as the standard to construct a disease term map that integrates common terms, and incorporating common terms as synonyms or hyponyms into ICD10, can effectively establish the mapping relationship between disease names and ICD10, which will facilitate doctors to find disease names or machine ICD automatic coding. However, the fusion of common terms requires a lot of medical knowledge, and the manual mapping is time-consuming and labor-intensive, and the accuracy of automatic machine mapping is relatively low. In addition, the classification system of ICD10 continues the traditional list structure, which is too flat and inconvenient to browse and search.

In view of the above problems and difficulties, this paper proposes a large-scale disease terminology map construction scheme that integrates common terms. Specifically, this paper screened out the common terms in the disease data of the Shanghai regional medical and health platform (which contains the clinical diagnosis and treatment information of 38 tertiary hospitals in the city) and integrated the common terms with ICD10. In addition, in order to facilitate doctors' search, the category layer of ICD10 and the hierarchical structure of the Chinese version of ICD-11 (abbreviated as ICD11) were further integrated to form a large-scale disease term map fused with common terms. The construction of the disease term map combines the advantages of machines and humans. In the proposed scheme, firstly, the components of disease words are analyzed, and the synonymous relationship between diseases is identified by the rule algorithm based on disease components, and the upper and lower relationship between diseases is found through the data-enhanced BERT (bidirectional encoder representation from transformers) upper and lower relationship identification algorithm. Then, using the characteristics of the ICD system itself, according to the type of disease, the disease data is verified based on the subspecialty grouping. The main contribution of the paper includes the following aspects:

(1) Constructing a large-scale disease term map fused with common terms for clinical diagnosis data, the map can represent the hyponymous relationship and synonymous relationship between medical terms and fuse common terms with standard terms. In the end, 1460 synonymous relations and 46508 upper and lower relations were found

(2) Designing a task assignment method based on the association map of disease departments, which is convenient for proofreaders to verify medical data, so as to ensure the accuracy of the relationship between disease medical entities

(3) Experimental studies reveal that the disease term map constructed in this paper is efficient in terms of the coding coverage, coding efficiency, and coding accuracy when compared with the manual coding and ICD10 system

2. Related Work

There are abundant researches on the construction of terminology system at home and abroad. A large number of biomedical classification systems have been presented in the literature. In addition to the general classification systems such as UMLS [1] and SNOMED-CT, there are also subdivisions such as the drug-oriented naming system RxNorm [1], the inspection-oriented coding system LOINC [2], and the widely used International Classification of diseases system. The domestic medical terminology system is constantly in line with international standards, such as ICD10. The construction of the early terminology system is purely manual, such as the semantic-oriented English dictionary WordNet [2] and the common knowledge graph CYC [3], in which CYC consists of 500,000 entities and 7 million assertions.

In recent years, the use of automatic methods to construct terminology systems has been widely used. The construction process involves the problem of automatic classification and induction; that is, it can effectively expand the entire knowledge structure. A large number of works have studied methods based on language model matching to solve the problem of terminology and its relationship with the problem of automatic classification and induction of relations between hypernyms. For example, Demir et al. [4] described a method to automatically obtain hyponyms from unrestricted text, and determined a set of lexical-syntactic patterns that were easy to identify. Reference [5] proposed a graph-based approach aimed at automatically learning lexical taxonomies starting from domain corpora and the Web. Experiments show that high-quality results can be obtained both when constructing a completely new taxonomy and when reconstructing the WordNet subhierarchy. Reference [6] proposed a new algorithm to automatically learn the upper and lower (isGa) relations from text to solve the problem of automatically constructing and extending semantic taxonomies such as WordNet. Reference [7] proposed a new metric-based framework for the task of automatic classification and induction. In recent years, the use of word embedding-based methods to identify relations to reconstruct taxonomy is also very popular [8–11].

New information such as common terms is added to the existing taxonomy, mainly focusing on enhancing the WordNet taxonomy [12] enriched WordNet with 310,742 named entities and 381,043 "relationship instances." Reference [13] created Medical-WordNet, which is not only a lexical expansion of medical terms in the original WordNet, but a new type of repository. Reference [14] studied the knowledge structure expansion problem, that is, how to add a large number of new concepts to the existing knowledge structure.

There are dual challenges to this problem, how to detect unknown entities or concepts and how to insert new concepts into existing knowledge structures without destroying the semantic integrity of newly created relationships. They propose a framework for ETFs to enrich large-scale general taxonomies with new concepts from sources such as news and research publications, linking new concepts to existing concepts and gaining potential parent-child relationships. However, the manual construction method requires a lot of manpower and material resources, and only the automatic construction method cannot guarantee the correct rate of the machine. Therefore, this paper adopts a method combining manual and automatic construction.

3. Construction of Disease Terminology Atlas

3.1. Problem Definition. This paper refers to and expands the classification hierarchy of ICD10 and ICD11 and defines the relationship between disease medical entities as follows:

Definition 1. Mapping relationship $R(E_i, E_j)$ between different disease medical entities. Among them, E_i and E_j are disease medical entities, and R is the mapping relationship. There are two types of mapping relationships:

(1) is_hypernym: relation is_hypernym (E_i, E_j) represents the upper and lower relationship between entities E_i and E_j. In particular, the is_hypernym relation is inversely functional: is_hypernym (E_i, E_j) $\Leftrightarrow$ is_hyponymy (E_i, E_j), that is, E_i is the hypernym of E_j, which is equivalent to E_j being the hyponym of E_i. For the sake of convenience, unless otherwise specified, the upper and lower relations in this paper refer exclusively to the upper and lower relations

(2) is_same: Relation is_same(E_i, E_j) represents the synonymous relationship between entities E_i and E_j. The synonymous relationship includes two parts: one is the medical synonymous relationship, similar to the synonymous relationship between "insulin-dependent diabetes mellitus" and "type 1 diabetes mellitus" The second is the synonymous relationship caused by the different writing habits of doctors, similar to the synonymous relationship between "type 1 diabetes" and "diabetes (type 1)".

The main task of this paper is to link common terms to ICD10 according to the relationship of disease medical entities and to fuse the category layer in ICD10 with the hierarchical structure of ICD11, so as to construct a large-scale disease term map that integrates common terms. Among them, common terms are defined as the names of diseases that appear more than 5 times in the clinically diagnosed disease data on the regional platform.

3.2. Overall Framework. The overall framework of this paper is shown in Figure 1. ICD10 fuses common terms, then adds ICD11 hierarchical structure information, and finally forms a disease term map fused with common terms. The left side of Figure 1 shows the basic framework of the disease term map. The fusion process is to determine whether the disease pairs with the standard disease terms in ICD10, and whether the common terms have a hyponymous relationship or a synonymous relationship. According to the disease medical entity relationship of the disease pair, the commonly used words are linked to each layer of the ICD10 to realize the classification of the commonly used words. The right side of Figure 1, respectively, shows the use of the disease component rule algorithm to identify whether the disease pair has a synonymous relationship, and the combination of BERT to identify whether the disease pair has an upper and lower relationship on the basis of the disease component based rule algorithm. Secondly, according to the mapping rules, the category layer in ICD10 is linked to the hierarchical structure of ICD11. Finally, in order to ensure the correctness of the fusion results, a task assignment method based on the association map of disease departments is introduced, which is convenient for verifiers to correct the relationship between disease medical entities contained in the disease term map.

3.3. ICD10 Fusion Phrases. For the task of identifying the relationship between disease terms, this paper defines it as the identification of synonyms and hyponyms, and the focus is on the identification of hyponyms. Reference [15] proposed a rule-based upper and lower identification algorithm, which is driven by knowledge and used for relationship judgment by preconstructing a dictionary containing a large number of fine-grained clinical entities and a set of upper and lower relations between entities. The rule-based method can identify the upper and lower relations with high quality, but limited by the size of the dictionary, its recall rate is very low. Therefore, on the basis of using the pretraining model combined with the reference results provided by the rules to provide auxiliary information, this paper proposes a data augmentation-based BERT upper and lower relationship recognition algorithm.

Given a disease pair (X_1, X_2), X_1 is the standard disease term in ICD10, and X_2 is the common term. Firstly, X_2 is sent to the rule algorithm based on disease components, and the optimal matching word X_3 of X_2 in the ICD10 corpus is obtained and the optimal matching pair (X_2, X_3). Next is the reference pair (X_3, X_1). Then, the disease pair (X_1, X_2) and the reference pair (X_3, X_1), respectively, go through BERT [16] to obtain the correlation representation [12] and [17] of the two elements in the word pair. Finally, concatenate [12] and [18], and use the feed forward neural network (FNN) to predict the upper and lower relationship.

In this paper, common terms and all the standard disease terms in ICD10 are formed into disease pairs, and the data-enhanced BERT epigenetic relationship recognition model is used to predict the hypostatic relationship of disease pairs, and all the prediction results are predicted as hyponymic relationships according to the model. The probability is sorted, and the highest probability (X_1, X_2) is taken as the final output result.

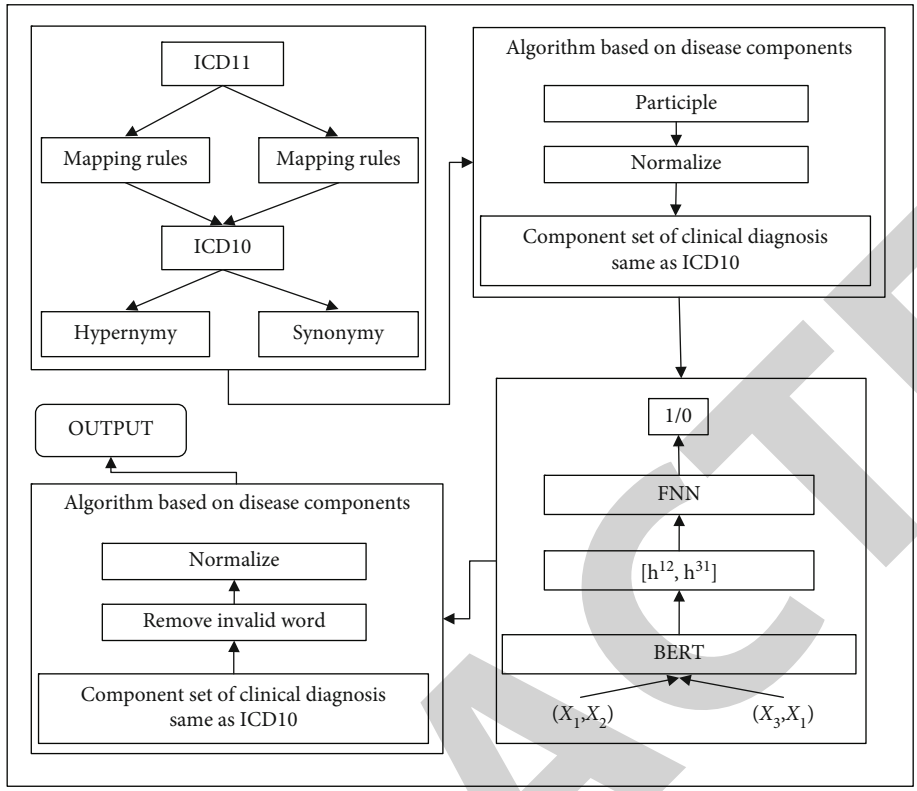

Figure 1: An overall framework for large-scale disease terminology map incorporating common terms.

3.3.1. Reference Pair Construction. The goal of constructing the reference pair (X_3, X_1) is to provide assistance for the identification of the upper and lower relationship between the disease pair (X_1, X_2), by judging the correlation between the reference results (X_3, X_1) predicted by the rule algorithm based on disease components information. Therefore, this paper defines the disease components to obtain the corresponding dictionary and uses the rule algorithm based on the disease components to obtain X_3.

(1) Definition of disease components: based on the analysis of the clinical diagnosis data of 38 hospitals included in the ICD and regional medical platform, this paper summarizes the disease words into atomic disease words, causal words, pathological words, and parts. Part words and clinical expression words are composed of five major components. Table 1 gives the specific meanings

(2) The rule algorithm based on disease components gives the disease name set $D = \{D_1, D_2, D_n\}$ of ICD10, where n is the total number of disease names. The rule algorithm based on disease components firstly segmented D_i and X_2 based on the bidirectional maximum matching algorithm of disease components and eliminated the invalid words "accompanied by" etc. Then, replace the remaining words with their corresponding standard names, thereby obtaining valid element sets set$D$ and set$X$, respectively. For the elements in the valid element sets set$D$ and set$X$ of D_iX_2, this paper iteratively replaces the hyponymous disease components with their hypernyms to detect the epistasis relationship, until the following situations occur

If set$X$ contains set$D$, then D_i is the hypernym of X_2 and returns the number of subsituations; otherwise, continue to perform hypernym substitution until there is no hypernym to replace. Finally, set X_3 as D_j satisfies the hypernym condition and has the least number of substitutions. The pseudocode of the algorithm is shown in Algorithm 1.

The problem of identifying the semantic relationship of disease medical entities based on the BERT hypernym relationship recognition algorithm based on data enhancement can be regarded as a classification task, that is, whether the standard disease term X_1 in the ICD10 is a hypernym of the common term X_2. The model architecture is shown in Figure 2.

This paper uses the pretrained language model BERT to encode disease pairs X_1, X_2 and reference pairs X_3, X_1, respectively. Taking the disease pair X_1 and X_2 as an example, the [SEP] tag is used to identify the segmentation information of the two disease words, and a special tag [CLS] is added at the beginning of the input sequence to form "$[CLS]X_1\ [SEP]\ X_2\ [SEP]$" as input. The model first calculates the input embedding, which includes the sum of word embedding, sentence embedding, and position embedding. Then, the input embedding is sent to the bidirectional

Table 1: Examples of disease components.

Disease components meaning	Disease components meaning
Atomic disease words	Atomic disease words: a part of a disease name, but not divide into finer grained words, such as diabetes
Causal words	Including the cause of disease and conditions. The cause of disease refers to those factors that can cause the disease and give the disease specificity, for example, hereditary
Pathological words	Modifying words such as severity, nature, and period of onset. For example, pregnancy is the pathological word of "gestational hypertension"
Part words	Indicating the location of disease in disease name. For example, stomach is the part word of "gastric ulcer"
Clinical expression words	A series of abnormal changes in a patient's body after he has a certain disease, such as "fever"

Input: Standard disease terms X_1 in ICD10, common terms X_2 in clinically diagnosed disease data, synonymous relation set R in the dictionary of disease components, stop word set $S = \{S_1, S_2, ..S_n\}$, disease components HypernymMap in the lexicon;
Output: The relationship of disease to (X_1, X_2)Perform word segmentation on (X_1, X_2) according to the bidirectional maximum matching algorithm, and obtain the components of $X_1 = \{X_{11}, X_{12}, \cdots X_{1m}$, $X_2 = \{X_{21}, X_{22}, \cdots X_{2n}\}$
for $X_{2i} \in X_2$ *do*
if $X_{2i} = S_i$ *then*
Move X_{2i} out of X_2
elseif X_{2i} in R *then*
Replace X_{2i} *with* the *standard synonym* in R ;
endif
endfor
Do the *same steps* 2 to 8 for X_1 ;
Obtain the *effective component set setX* of X_2 and the \ *effective component set setD* of 1 *respectively* ;
if $setX - setD = \varnothing$ *then*
return synonymous relation ;
else if $setD \in setX$ *then*
return upper − lower relationship ;
else while X_{2i} in *setX has hypernym* in *HypernymMap do*
Replace X_{2i} *with its hypernym counterpart* X_{2i} ;
if $setX - setD = \varnothing$ *then*
return synonymous relation ;
break ;
else if $setD \in setX$ *then*
return upper and *lower relationship* ;
break ;
endif
end while
return irrelevant
end if

Algorithm 1: Rules of algorithm based on disease components.

transformer model, and the output [CLS] contains the information about whether the two disease words are related. The final output [18] of the labeled [CLS] is used as the correlation representation in the classification task vector. Similarly, the reference pair X_3, X_1 is sent to BERT to get [19]. Finally, [20], [21] are concatenated and sent to the feed forward neural network, and the output result is 0/1 (0 means there is no relationship between the two, and 1 means that X_1 is the upper and lower relationship of X_2).

3.3.2. Comparison Experiment of Term Graph Relationship Recognition Algorithm. This paper verifies the effectiveness of the algorithm used in constructing a disease term graph that integrates common terms. We use the disease data in the regional medical platform as the experimental data set. In particular, there are few synonymous relationships between disease names in this dataset, so the rule algorithm based on disease components is directly used to judge the synonymous relationship, so this paper only conducts comparative experiments on the upper and lower relationship.

This paper selects four relationship recognition algorithms for comparison:

(1) String similarity algorithm: first, find out the Levenshtein distance (X_1, X_2) between the standard disease term X_1 and the common term X2 in

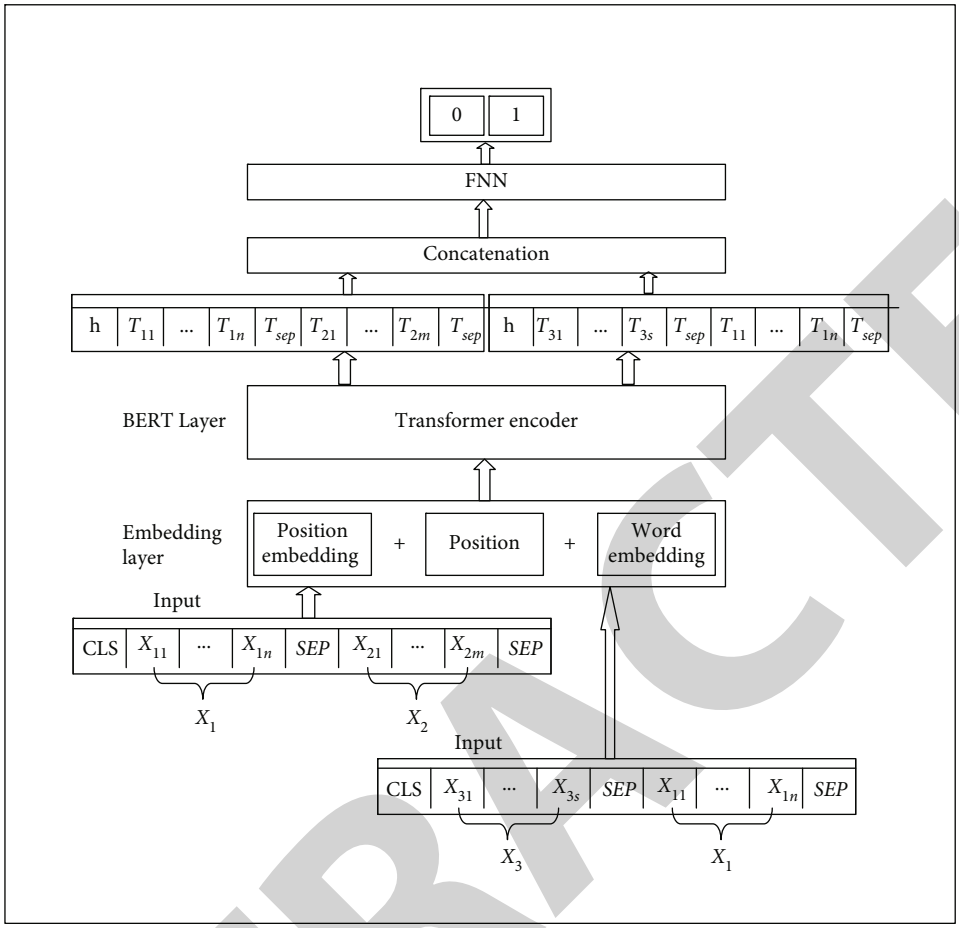

Figure 2: Algorithm model of BERT upper and lower relationship recognition based on data enhancement.

ICD10 for each disease pair. The Levenshtein distance refers to the minimum editing operation required to convert two strings from one to the other frequency. If the result of distance (X_1, X_2) exceeds the threshold, it is considered that X_1, X_2 have an upper-lower relationship; otherwise, there is no relationship. The threshold set in this paper is 0.8

(2) Dynamic distance loss model: Reference [11] trains a hyponym vector O_{X_2} and a hypernym vector E_{X_2} for each common phrase X_2. Whenever X_2 appears as a hyponym, use O_{X_2}; whenever it appears as a hypernym candidate, use E_{X_2}. Then, use the supervised corpus to train the SVM model, and use the trained model to judge whether the input disease pair (X_1, X_2) is a hypernym pair

(3) Rule algorithm based on disease components: according to Reference [15], the disease pairs (X_1, X_2) are firstly segmented according to the dictionary, and the elements after word segmentation are subjected to stop words and standardization operations. If the elements of X_1 are included in the elements of X_2, then X_1 is the hypernym of X_2, otherwise, iteratively replaces that element of X_2 with its hypernym

(4) BERT reference [16]: input the disease pairs (X_1, X_2) in the form of $"[CLS]\, X_1\, [SEP]\, X_2\, [SEP]"$ into the pretraining model BERT, followed by a feed-forward neural network for binary classification

For the relationship identification results, the evaluation indicators in this paper use the most commonly used Precision, Recall, and $F1$_score as the evaluation criteria. The calculation formula of the evaluation results is

$$\begin{aligned} \text{Precision} &= \frac{\text{Number of right relationships}}{\text{Total number of relationships}} \times 100\%, \\ \text{Recall} &= \frac{\text{Number of right relationships}}{\text{Total number of relationships in standard results}} \times 100\%, \\ F1_{\text{score}} &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 100\%. \end{aligned} \tag{1}$$

Table 2 and Figure 3 show the Precision, Recall, and F1_score of the five comparison algorithms. Compared with the existing algorithms, the proposed algorithm obtains the

Table 2: Comparative experimental results.

Algorithm	Precision	Recall	F1_score
String similarity algorithm	96.56	80.03	72.97
Dynamic distance loss model	72.32	88.73	80.36
Rule algorithm based on disease components	95.52	22.12	36.23
BERT	94.63	91.84	93.52
Proposed method	96.89	92.28	94.70

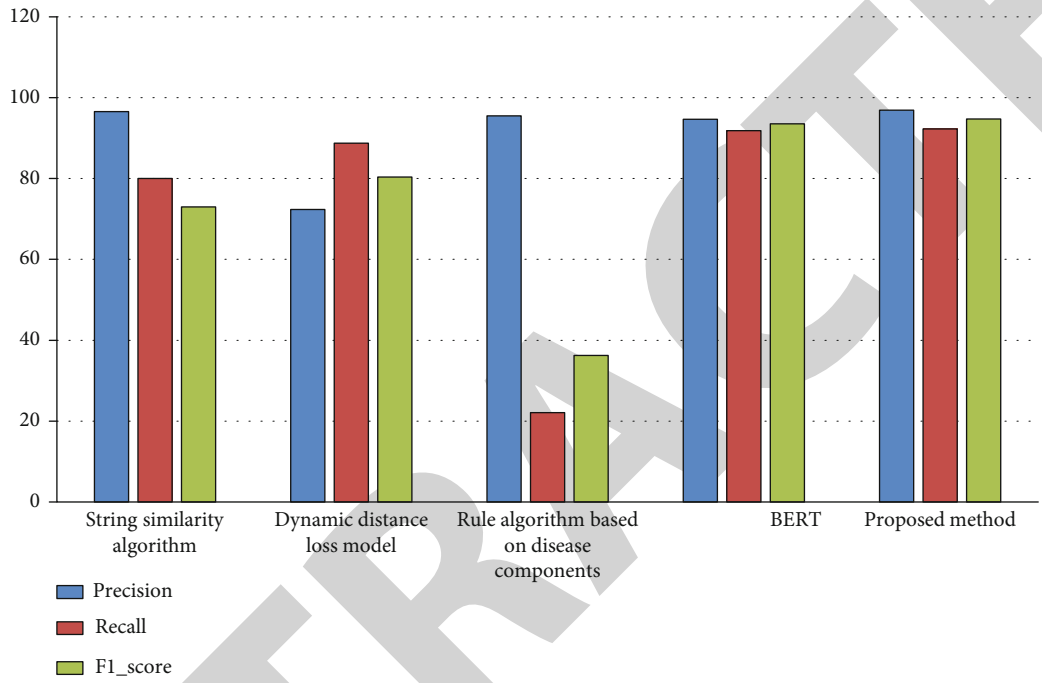

Figure 3: Comparative experimental results.

best F1_score value, and Precision, Recall, and *F*1_score are 96.89%, 92.28%, and 94.70%, respectively. For the rule-based relationship recognition method, its Precision reaches 100%, but the Recall is very low, because the algorithm is limited by the size of the dictionary and does not cover completely, but its prediction results have high confidence. This is also the reason why this paper integrates the algorithm to provide auxiliary information. In addition, we found that the *F*1_score value of the proposed algorithm is 0.92% higher than that of using BERT alone, which proves the effectiveness of the data augmentation-based BERT subordinate relationship recognition algorithm.

3.4. Add ICD11 Hierarchy Information. With the help of the mapping table published on the official website of ICD10 and ICD11, this paper links all category-level diseases in the ICD10 structure that incorporates common terms to the ICD11 hierarchy to add ICD11 hierarchy information to obtain a more fine-grained disease hierarchy. The structure is convenient for doctors to view and screen diseases. The reason for adding ICD11 hierarchy information is as follows:

(1) The hierarchical structure of the 3-digit category code of ICD10 is too flat and fails to reflect the hierarchical structure of diseases. "diabetes" and "endocrine diseases" are on the same level in ICD10, and "diabetes" should belong to "endocrine diseases"; that is, "diabetes" should be located at the lower level of "endocrine diseases"

(2) The classification of diseases is becoming more and more refined. ICD11 adjusts the classification axis, changes the classification level, adds or refines taxonomic units, and revises and improves the original classification structure and classification knowledge of ICD10. However, in view of the fact that medical institutions have used ICD10 as a disease code in the past 10 years, therefore, it is necessary to use ICD10 to fuse with common terms first and then add the hierarchical structure information of ICD11

The classification code of the ICD10 standard is firstly the category, which is divided into suborders with a total of three levels. In this paper, ICD10 category layer diseases are mapped to ICD11 diseases at any layer, and it is found

that ICD10 category layer can map 90.34% of the diseases in ICD11, so the diseases in the category layer in ICD10 (2047 in total) are mapped to the ICD10 category layer diseases. The results of each layer node of ICD11 are shown in Table 3 and Figure 4. A total of 2521 items are mapped in Table 3, while there are 2047 items in the ICD10 category level diseases, and the reason for the extra 474 items is that 213 items are not uniquely mapped. For example, "Other bacterial enteric infections" (code A04) in the ICD10 category layer is further split into "Other Vibrio enteric infections" (code 1A01) and "Escherichia coli" in ICD11. "Intestinal infection" (code 1A03), "bacterial intestinal infection, unspecified" (code 1A02) result in nonunique mappings. Therefore, it is necessary to further align the ICD10 suborder and detail layers with the ICD11 multiple mapping, and this requires the intervention of professional medical staff, so this paper uses the task assignment method based on the association map of disease departments to perform knowledge verification on the large-scale disease term map constructed with the fusion of common terms.

3.5. Knowledge Verification. Even after data enhancement, the upper and lower relationship recognition algorithms based on the above still cannot guarantee that the predicted upper and lower relationships are all correct, and two types of errors may occur:

(1) Common terms are related to wrong ICD10 names. For example, "type 2 diabetic neuropathy" has an upper and lower relationship with "type 2 diabetic neuritis" through an algorithm, and the correct one should be "type 2 diabetes with neurological complications"

(2) The name of ICD10 is not a direct hypernym of common expressions. In this paper, the most adjacent hypernyms in the hierarchy of common terms are called direct hypernyms. The hypernym relationship is transitive; that is, X is the direct hypernym of Y, Y is the direct hypernym of Z, and X is the hypernym (nondirect hypernym) of Z. For example, "type 2 diabetic macro albuminuria" has an upper and lower relationship with "type 2 diabetes" through an algorithm, and "type 2 diabetic nephropathy" is the direct hypernym of "type 2 diabetic macroa lbuminuria". The judgment and correction of the above situations depend on deeper domain knowledge, and in order to ensure the medical correctness of the disease terminology map, manual work is needed

The departments corresponding to the standard disease terms in the disease pair are divided into multiple department-based term subsets to be verified. The same subset of terms to be verified will be assigned to multiple proofreaders in the same department for verification and modification. After completion, it will be automatically judged by the machine, and the data with the reliability of the verification result higher than 0.5 will be classified. For the correct term set, the rest will be checked by experts.

Table 3: ICD10 category layer mapped to the mapping of each layer of ICD11.

ICD11 hierarchy	Mapping number	Mapping percentage (%)
3rd	688	28.11
4th	1049	45.75
5th	582	22.80
6th	115	5.06
7th	17	0.50

(1) Assignment of tasks based on the department where the disease is located. For the standard disease terms and common terms in the disease pair ICD10, firstly, according to the added hierarchical structure information of ICD11, the standard disease terms in the disease pair are roughly classified according to chapters, and then use the disease department knowledge map we constructed previously. The standard disease terms under each chapter are subdivided by departments, and the disease pairs that are finally classified into the same department will be filled in the same knowledge verification form, and the hierarchical structure of their standard disease terms will be expanded

(2) Manual proofreading: the same task will be assigned to n ($n \geq 3$) medical staff for verification, in order to reduce the randomness and chance of verification results. In response to the wrong ICD10 name on the link of common words, the medical staff modified the knowledge verification table (the hierarchy of the term base where the common term is located), it is judged whether the common term is a direct hypernym, and the corresponding modification is made. In the manual proofreading process, if all proofreaders have not modified a certain piece of data, the piece of data will be directly added to the correct term set

(3) Proofreading consistency judgment: when multiple people proofread the same piece of data, there will be a variety of modification situations. In view of the inconsistency of multiperson proofreading results, it is necessary to evaluate the quality of proofreading results. For the results of manual proofreading, the specific quality assessment is as follows: each piece of data to be proofread is regarded as a proofreading task T_i, and each proofreading task T_i is guaranteed to have n ($n \geq 3$) proofreaders to check. Each proofreader may have m kinds of proofreading results in a verification task d_i. Therefore, the confidence level of each proofreading result is calculated as

$$td_j = \frac{n_j}{n} (j = 1, 2, \cdots \cdots, m) \tag{2}$$

Among them, n_j represents the number of people who

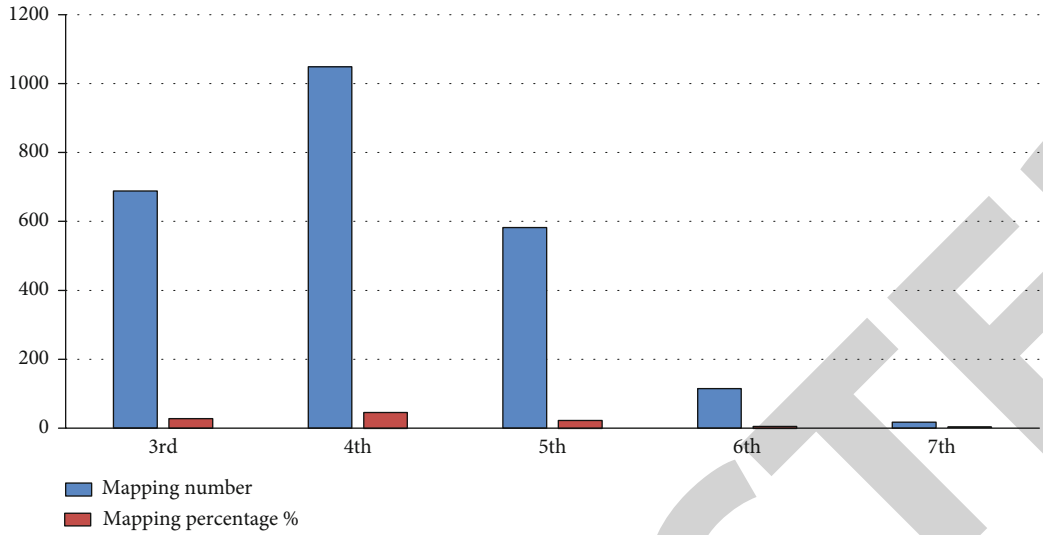

Figure 4: ICD10 category layer mapped to the mapping of each layer of ICD11.

Table 4: Disease name code mapping (%).

Coding mode	Mapping rate of data coding	
	The 1st group	The 2nd group
Coding based on ICD10	11.96	12.01
Coding based on our method	95.69	74.37

choose the jth proofreading result. If $td_j > 0.5$, the result of the jth proofreading task of this proofreading task T_i is correct, and the correct term set will be output directly; otherwise, T_i will be checked by medical experts.

(4) The man-machine combination method saves labor costs. First, for each common term, the algorithm predicts its position in the ICD10. Although this may not be exact, the assignment of the subtree of terms in which it resides is generally accurate. For example, "type 2 diabetic neuropathy" and "type 2 diabetic macro albuminuria", are both subtrees of the term "type 2 diabetes mellitus". This ensures that the data search space is reduced from ICD ensemble search to subtree search. Moreover, diabetes belongs to the department of endocrinology as a whole, and the assignment of specialist proofreading personnel is also correct, which ensures that personnel can check familiar diseases

4. Disease Term Coding Assessment

4.1. Assess Coding Coverage. In order to verify that the disease term map constructed in this paper can effectively cover more clinical diagnostic data, we extracted 10,038 data from the electronic medical record (EMR) discharge summary table as the first group of evaluation data and from the follow-up data. 9426 pieces of data were extracted as the second group of evaluation data, and the number of successful mapping of disease coding based on ICD10 and the disease term map constructed in this paper was counted. The mapping results are shown in Table 4 and Figure 5.

It can be seen from Table 4 that using the disease term map constructed in this paper can increase the coding coverage rate by 74.37% on average compared to the one based on ICD10, which proves that more disease-corresponding codes can be found using the disease term map. However, the disease term map constructed in this paper still fails to find all the medical entity relationships of diseases, and the reasons include two aspects: (1) due to the fact that there are two disease names in the real data. For example, the disease name "neonatal convulsion (epilepsy)", in which "neonatal convulsion" corresponds to P90 in ICD10, the corresponding code of "epilepsy" in ICD10 is G40.901, and "neonatal convulsion" and "epilepsy" are two diseases, and it is difficult for the disease term map to distinguish the disease code according to the algorithm. (2) Data that is not the name of the disease appears in the real data, such as "after autologous stem cell transplantation" and "after posterior urethral valve operation". For the first case, different weights can be set for disease names containing two codes according to the symbols. For the second case, the occurrence of data with nondisease names should not have been linked to the Disease Terminology Atlas.

4.2. Evaluate Coding Efficiency. In order to verify the advantages of the large-scale disease term map constructed in this paper fused with common terms in the medical field when doctors fill in disease codes, we set up two evaluation methods, manual coding and machine-assisted coding, in order to compare the constructed disease term map for doctors. For the effect of coding disease efficiency for manual coding, we recruited 5 medical staff who were familiar with ICD codes, given the ICD10 disease standard classification codes, and counted the completion time of 5 testers to find out the matching codes for 50 randomly sampled disease names.

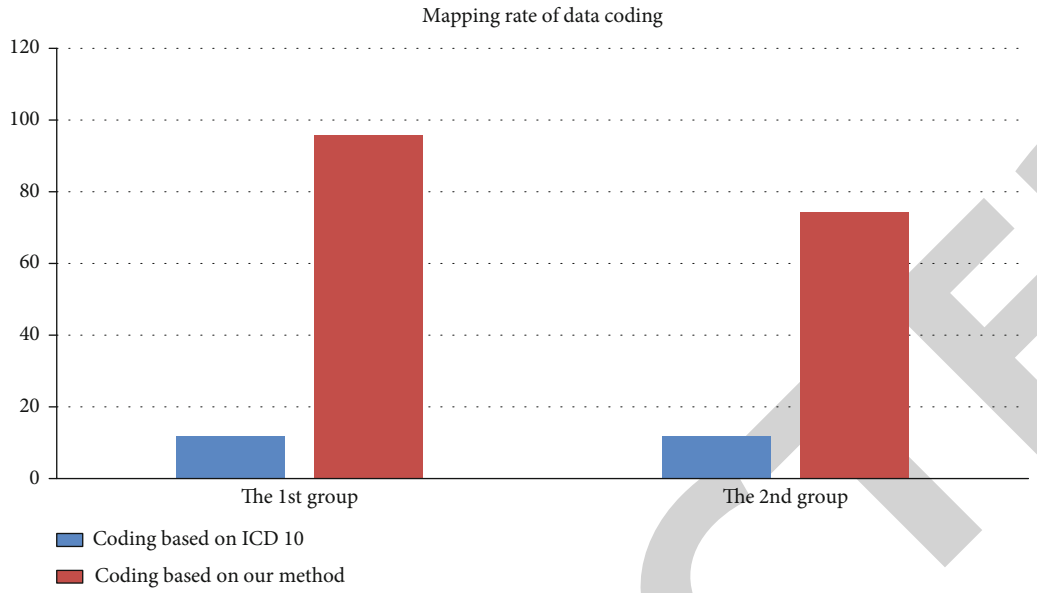

Figure 5: Disease name code mapping (%).

Table 5: Completion time results for human coding and machine-assisted coding.

Proofreader	Manual coding (s)	Machine-aided coding (s)
1	475.017	196.423
2	392.206	162.221
3	412.157	160.871
4	466.403	171.990
5	468.183	179.015
Average	440.102	169.939

For machine-assisted coding, we first used the disease term map constructed in this paper to automatically find the ICD10 codes corresponding to 50 disease names and displayed them in the form of the knowledge verification table. The completion time at this time is defined as the sum of the machine running time and the time spent by the proofreader. The experimental results are shown in Table 5 and Figure 6. The completion speed of the auxiliary coding using the disease term map constructed in this paper is 2.48 times that of manual coding, indicating that using the disease term map constructed in this paper to automatically perform disease coding can shorten the coding time for doctors. In practice, the medical staff of medical institutions are not too familiar with the ICD coding system, which will also affect the coding efficiency, and with the increase in the amount of disease data, the application of the disease term map constructed to the filling process of the first page of medical records have more prominent advantages.

4.3. Evaluate Coding Accuracy. The validity of the disease term map constructed in this paper is verified by using the electronic health record (EHR) data of the regional platform. The data includes the registration data of 38 tertiary hospitals in Shanghai, and the data containing the doctor code accounts for 536,456 pieces, and the data is cleaned. After that, 2 special disease data were randomly selected as evaluation data. The goal of this evaluation is to count the respective accuracy of the doctor's manual coding and the coding using the disease term map constructed in this paper. The results are shown in Table 6. It is worth noting that the standard ICD code of the evaluation data is based on the ICD10 code obtained after knowledge verification.

From the results in Table 6, it can be seen that the correct rate of disease term map coding constructed in this paper is much higher than that of doctors' manual coding, which is increased by 60%. Analyze the reasons for the low accuracy rate of doctors' manual coding: (1) doctors have inconsistent understanding of coding. For the disease name "type 2 diabetic ketosis", the doctor's coding includes E11.103, E11.100, and FFF. The code of the disease term map is E11.100. After verification by the proofreaders, it is synonymous with "type 2 diabetic keto acidosis", and the code should be E11.100.2 some doctors fill in the disease code irregular. For example, the commonly used term "stage IV gastric malignant tumor" should be linked to "stomach malignant tumor" (coded as C16.900) in ICD10, and the doctor code is C16. Another example is the commonly used term "type 2 diabetes", which corresponds to "type 2 diabetes" in ICD10 (coded as E11.900), while the doctor's code is written as E11.90000S.

The above-mentioned experimental study shows that the disease terminology map constructed in this paper not only maintains the existing standard system but also takes into account the convenience of clinical use. The disease term map was evaluated from three aspects: coding coverage, coding efficiency, and coding accuracy. Compared with the ICD10 system, the disease term map constructed can cover 75% more on average. Compared with manual coding, the use of disease term atlas-assisted coding can shorten the time by about 59.75%, and the accuracy rate reaches 85%.

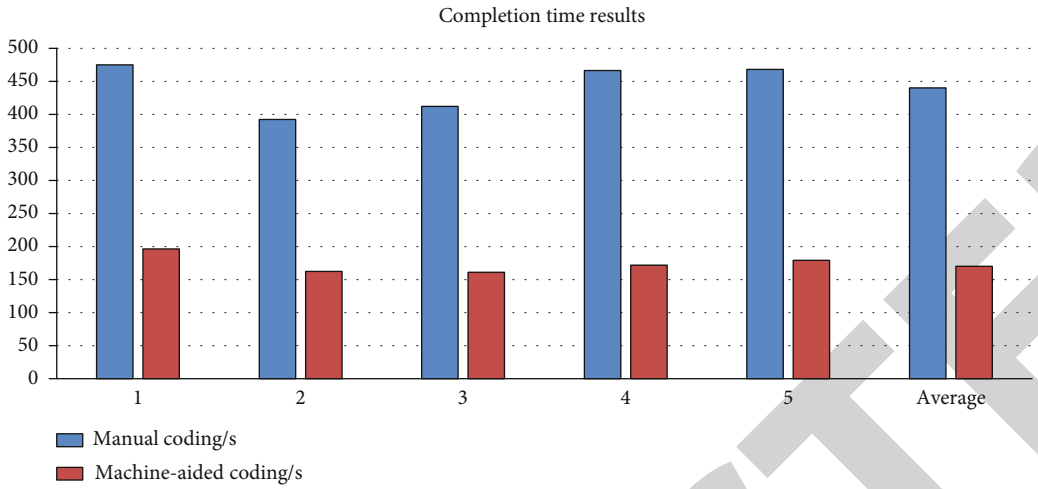

Figure 6: Completion time results for human coding and machine-assisted coding.

Table 6: Manually coded by physicians and automatically coded using the Disease Terminology Atlas.

Coding mode	Accuracy (%)
Doctor manual coding	68.56
Coding based on our method	83.48

5. Conclusion

In this paper, the disease medical entity relationship between common terms and standard disease terms in ICD10 is identified through the rule algorithm based on disease components and the BERT hypernymous relationship recognition algorithm based on data enhancement, and the mapping between common terms and ICD10 codes is realized, and the ICD11 code is added. The hierarchical structure is convenient for doctors to check the ICD10 code corresponding to the disease. Disease coding using the disease term map constructed in this paper has good performance in coding coverage, accuracy, and coding efficiency. In the future, the disease term map can be applied in various medical structures to ensure the coverage, efficiency, and accuracy of disease coding and to promote the standardization process of medical information.

Data Availability

All the data are available on miretab@dadu.edu.et.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The research is self-funded.

References

[1] E. Ennadifi, S. Laraba, D. Vincke, B. Mercatoris, and B. Gosselin, "Wheat diseases classification and localization using convolutional neural networks and GradCAM visualization," in *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pp. 1–5, Fez, Morocco, 2020, June.

[2] B. Lei, P. Yang, T. Wang, S. Chen, and D. Ni, "Relational-regularized discriminative sparse learning for Alzheimer's disease diagnosis," *IEEE Transactions on Cybernetics*, vol. 47, no. 4, pp. 1102–1113, 2017.

[3] D. Mahapatra, A. Poellinger, L. Shao, and M. Reyes, "Interpretability-driven sample selection using self supervised learning for disease classification and segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 10, pp. 2548–2562, 2021.

[4] J. Cai, A. Liu, T. Mi et al., "Dynamic graph theoretical analysis of functional connectivity in Parkinson's disease: the importance of fiedler value," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 4, pp. 1720–1729, 2019.

[5] Z. Lin, S. Mu, F. Huang et al., "A unified matrix-based convolutional neural network for fine-grained image classification of wheat leaf diseases," *IEEE Access*, vol. 7, pp. 11570–11590, 2019.

[6] J. Hariharan, Y. Ampatzidis, and J. Abdulridha, "The basis for development of a foundational biomarker reflectance signature database system for plant cell identification, disease detection, and classification purposes," in *2020 10th Annual Computing and Communication Workshop and Conference (CCWC)*, pp. 0881–0887, Las Vegas, NV, USA, 2020, January.

[7] Y. Yin, X. Yang, J. Xiong, S. I. Lee, P. Chen, and Q. Niu, "Ubiquitous smartphone-based respiration sensing with Wi-Fi signal," *IEEE Internet of Things Journal*, vol. 9, no. 2, pp. 1479–1490, 2022.

[8] M. M. D. Quintana, R. R. D. Infante, J. C. S. Torrano, and M. C. Pacis, "A hybrid solar powered chicken disease monitoring system using decision tree models with visual and acoustic imagery," in *2022 14th International Conference on Computer and Automation Engineering (ICCAE)*, pp. 65–69, Brisbane, Australia, 2022, March.

[9] D. Wang, J. Wang, Z. Ren, and W. Li, "DHBP: a dual-stream hierarchical bilinear pooling model for plant disease multi-task classification," *Computers and Electronics in Agriculture*, vol. 195, p. 106788, 2022.

[10] L. Zhang, G. Zhou, C. Lu et al., "MMDGAN: a fusion data augmentation method for tomato-leaf disease identification," *Applied Soft Computing*, vol. 123, p. 108969, 2022.

[11] R. G. Dawod and C. Dobre, "ResNet interpretation methods applied to the classification of foliar diseases in sunflower," *Journal of Agriculture and Food Research, Volume 9*, vol. 9, p. 100323, 2022.

[12] M. Baas, A. P. Stubbs, D. B. van Zessen et al., "Identification of associated genes and diseases in patients with congenital upper-limb anomalies: a novel application of the OMT classification," *The Journal of Hand Surgery*, vol. 42, no. 7, pp. 533–545.e4, 2017.

[13] A. de Jesús Plasencia Salgueiro, Y. Shichkina, and L. G. Rodríguez, "Parkinson's disease classification and medication adherence monitoring using smartphone-based gait assessment and deep reinforcement learning algorithm," *Procedia Computer Science*, vol. 186, pp. 546–554, 2021.

[14] S. Ma, K. Yang, X. Zhou, X. Xue, and W. Liu, "Similarity-based algorithms for disease terminology mapping," in *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1378–1384, Shenzhen, China, 2016, January.

[15] M. Saeed, M. Ahsan, A. Mehmood, M. H. Saeed, and J. Asad, "Infectious diseases diagnosis and treatment suggestions using complex neutrosophic hypersoft mapping," *IEEE Access*, vol. 9, pp. 146730–146744, 2021.

[16] F. Demir, A. Sengur, A. Ari, K. Siddique, and M. Alswaitti, "Feature mapping and deep long short term memory network-based efficient approach for Parkinson's disease diagnosis," *IEEE Access*, vol. 9, pp. 149456–149464, 2021.

[17] P. Maji and E. Shah, "Significance and functional similarity for identification of disease genes," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 6, pp. 1419–1433, 2017.

[18] W. Haider, A.-U. Rehman, N. M. Durrani, and S. U. Rehman, "A generic approach for wheat disease classification and verification using expert opinion for knowledge-based decisions," *IEEE Access*, vol. 9, pp. 31104–31129, 2021.

[19] S. Jokić, D. Cleres, F. Rassouli et al., "TripletCough: cougher identification and verification from contact-free smartphone-based audio recordings using metric learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2746–2757, 2022.

[20] B. Dan, X. Sun, and L. Liu, "Diseases and pests identification of Lycium barbarum using SE-MobileNet V2 algorithm," in *2019 12th International Symposium on Computational Intelligence and Design (ISCID)*, p. 121, Hangzhou, China, 2019, December.

[21] C. Adak, B. B. Chaudhuri, and M. Blumenstein, "An empirical study on writer identification and verification from <italic>Intra-Variable</italic> individual handwriting," *IEEE Access*, vol. 7, pp. 24738–24758, 2019.