ORIGINAL ARTICLE

Bali Medical Journal (*Bali MedJ)* 2022, Volume 11, Number 3: 1162-1167
P-ISSN.2089-1180, E-ISSN: 2302-2914

Symptoms identification of ICD-11 based on clinical NLP mobile apps for diagnosing the disease (ICD-11)

Rizqi Putri Nourma Budiarti[1*], Sritrusta Sukaridhoto[2], Ilham Achmad Al-Hafidz[3], Naufal Adi Satrio[3]

[1]Department of Information Systems, Faculty of Business Economics and Digital Technology, Universitas Nahdlatul Ulama Surabaya;
[2]Department of Multimedia Creative, Politeknik Elektronika Negeri Surabaya, Indonesia;
[3]Department of Informatics and Computer, Politeknik Elektronika Negeri Surabaya, Indonesia;

*Corresponding author:
Rizqi Putri Nourma Budiarti;
Department of Information Systems, Faculty of Business Economics and Digital Technology, Universitas Nahdlatul Ulama Surabaya;
rizqi.putri.nb@unusa.ac.id

Received: 2022-07-16
Accepted: 2022-08-11
Published: 2022-09-13

ABSTRACT

Introduction: There are still many people in Indonesia who are not aware of the importance of information related to the early symptoms that must be experienced when they become patients. Not infrequently, this lack of information disclosure results in misdiagnosis and even leads to unexpected death. Anamnesis is a process where the doctor or medical record nurse gives several questions about the clinical pathway in the form of a narrative to facilitate early identification of the disease, and the results of this history-taking process are stored in the Electronic Medical Record (EMR). EMR narratives often cannot be processed by computers if language literacy is not standardized or ambiguous, so the need to overcome this problem requires the use of technology to minimize misdiagnosis and facilitate the identification process by developing digitization in the form of mobile applications that are integrated with Natural Language Processing technology and ICD-11 in the symptom identification process. This study aims to identify ICD-11 symptoms based on clinical NLP mobile application to diagnose the disease (ICD-11).

Methods: The applications of Natural language processing includes literature study, Voice Recognition, Tokenization, Stemming, The process of Stopwords Removal, Named Entity Recognition, Data Translation, Access ICD Data, and Mobile User Interfaces.

Results: Named Entity Recognition (NER) is used to identify symptoms of digestive system diseases, with an accuracy rate of 74.3%. In stemming and stopwords processing, the NLP accuracy rates are 95.9% and 97.2%, respectively.

Conclusions: This research focuses on the application mobile and development of the Named Entity Recognition (NER) model. The importance of the NLP process in the development of information, especially for word processing, aims as a device that simplifies speech recognition systems to be more helpful.

Keywords: ICD-11, NLP mobile apps, disease.

Cite This Article: Budiarti, R.P.N., Sukaridhoto, S., Al-Hafidz, I.A., Satrio, N.A. 2022. Symptoms identification of ICD-11 based on clinical NLP mobile apps for diagnosing the disease (ICD-11). *Bali Medical Journal* 11(3): 1162-1167. DOI: 10.15562/bmj.v11i3.3533

INTRODUCTION

The disease is a condition where there is an abnormal body condition, which causes the loss of a healthy normal condition. It is necessary to carry out a diagnostic process to identify the type of disease or health problem in the human body. By carrying out the diagnostic process, a disease can be known or identified earlier so that efforts can be made to control or prevent the spread of the disease.[1]

There are many ways to diagnose a disease. One way is to use the Anamnesis technique, which is a way to find out the patient's health condition through a question and answer process between doctors or other health professionals with the patient directly or indirectly. Anamnesis can be done in two ways. The first is Autoanamnesis, an anamnesis process which is done directly to the patient. The patient himself answers all questions and describes his condition. The second is Alloanamensis, which is an anamnesis process that is carried out with other people, such as the patient's family or friends of the patient, to obtain accurate information about the patient's condition. Usually, in unconscious patients, infants, and children. In this type of history taking, the medical officer must ensure that the source of information comes from the right person. After the diagnosis process has been completed, ICD can be used to classify the patient's disease.[1,2]

The International Classification of Disease or abbreviated as ICD is a disease classification system and various types of symptoms, disorders, complaints, and external causes of disease. Each disease is assigned a category and code. ICD is published by WHO and is widely used for supporting medical decisions. The method to classify diseases on the ICD list is a combination of letters and numeric codes, which aim to classify diseases based on their nature, type, and level of seriousness in disease to facilitate the process of health management. On ICD-11, There are 55,000 disease classification codes based on the index and detailed information, and comparing with ICD-10 there are 14,400 codes different.[1] This study aims to identify ICD-11 symptoms based on clinical NLP mobile application to diagnose disease (ICD-11).

METHODS

Literature Study

a. Speech Recognition

Speech Recognition (SR) is a system developed where a computer can handle the conversion of speech signals in speech and convert them into features that can produce word recognition information. The speech parameters are used to describe words that have been digitized by matching digital signals with specific patterns stored in devices such as mobile devices (for example, using Siri on iPhone, Bing voice search on WinPhone, and Google Assistant on Android).[2] Spoken words are converted into digital signals by converting sound waves into a series of numbers, which are then matched with a specific code to identify the words. The results of the identification of spoken words can be displayed in written form or read by technological devices as a command to do a job, such as pressing a button on a cell phone, which is done automatically by voice command. Voice recognition tools, often called voice recognition, require samples of the user's actual words. The sample words will be digitized, stored on the computer, and used as a database to match the following spoken word.

b. Natural Language Processing

Natural Language Processing (NLP) is a widely used technique that systems can understand instructions to manipulate text or speech.[3] In general, the techniques applied to NLP (Natural Language Processing) are more on the use of speech technology, especially the use of Automatic Speech Recognition and Text-To-Speech synthesis.[4] Processing of input text to be synthesized is one of the important functions in NLP where the natural language words generated by the signal processing module are closely related to the performance of the previous text processing module.[4] NLP makes computers more dynamic and humanist because computers try to convey sentence responses like humans. NLP aims to make humans better understand the specific responses given by the computer system. The implementation of NLP in the health sector allows health workers to recognize and predict a disease based on a patient's electronic medical record and the patient's own words, helping health workers to make decisions.[5]

Implementation

The applications of Natural language processing include literatur study, Voice Recognition, Tokenization, Stemming, The process of Stopwords Removal, Named Entity Recognition, Data Translation, Access ICD Data, and Mobile User Interfaces.

a. Voice Recognition

The initial stage in this system process is to get input from the patient's voice recording. The patient's voice recording contains complaints about digestive diseases experienced by the patient. The bitrate patient recording audio file is 16Kbps. As the research data sample, the patient's voice was recorded 3 times with several types of complaints, as shown in figure 1.

NLP is a linguistic-based technology so the processed voice data must be converted into text. The language in NLP can be adjusted according to the language of each country. The language used in this research is Indonesian. At this stage, all voice data will be converted into text for further processing.

b. Stages of Tokenization

The stage of the tokenization is cutting the input string based on the words it has compiled. In general, tokenization is used to separate a set of patterns or characters contained in text into word units. this is the reason why differences in certain characters can be used as word separators or vice versa. An observable example is the space or tabulation characters such as enter can be categorized as word separators. However, for single quote characters ('), period (.), Semicolon (;), double point (:) or others, can have quite a several roles as word separators. In this research, the function of tokenization is to break down sentences into words.[6]

c. Stages of Stemming

Stemming is the process of mapping and parsing the form of a word into its basic word form, where this part cannot be separated from the IR (Information Retrieval). In the stemming process, there is a minimization of differences in the number of indexations of a document. It functions to group other words that have the same root and meaning but have different primary forms due to different affixes. The use of basic word combinations

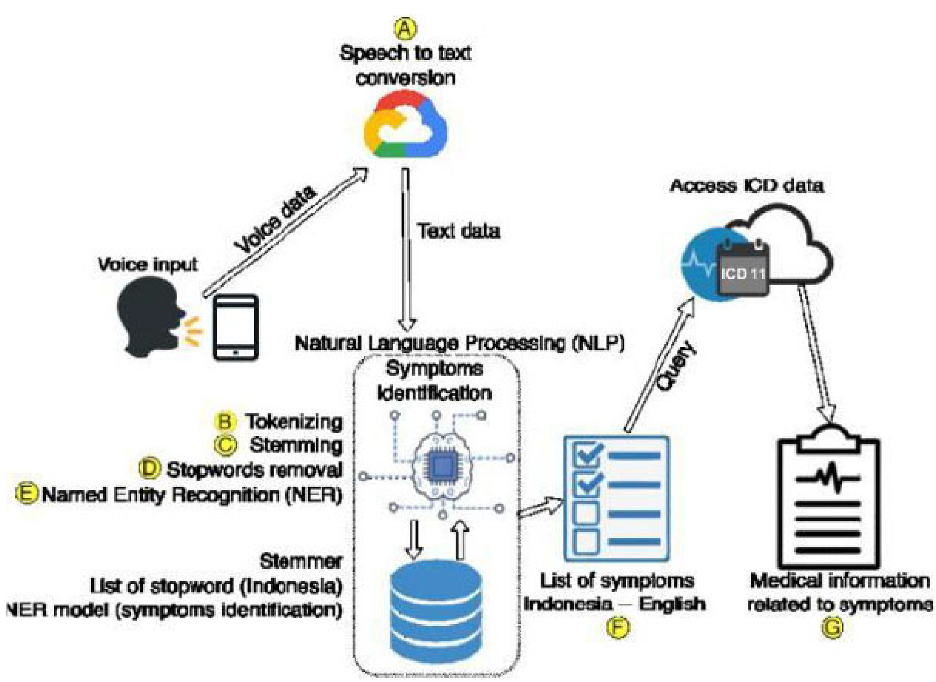

Figure 1. Workflow Diagram

Table 1. **Voice Record Data.**

No.	Speech Input
1	Nafas saya pendek dan dada terasa sakit terkadang diikuti dengan batuk-batuk
2	Tenggorokan terasa gatal dan terkadang hidung saya mampat.
3	Sudah tiga hari ini saya merasa sering mengantuk dan sering letih, kebingungan, atau pusing.

Table 2. **Speech Text Conversion.**

No.	Speech Text Conversion
1	Nafas saya pendek dan dada terasa sakit terkadang diikuti dengan batuk-batuk
2	Tenggorokan terasa gatal dan terkadang hidung saya mampat.
3	Sudah tiga hari ini saya merasa sering mengantuk dan sering letih, kebingungan, atau pusing.

Table 3. **Word Tokenizing.**

No.	Word Tokenizing
1	[‘nafas, ’saya’, ’pendek’, ’dan’, ’dada’, ’terasa’, ‘sakit’, ’terkadang’, ’diikuti’, ’dengan’, ’batuk’, ’batuk’]
2	[’tenggorokan’, ’terasa’, ’gatal’, ’dan’, ’terkadang’, ’hidung’, ’saya’, ’mampat’]
3	[’sudah’, ’tiga’, ’hari’, ’ini’, ’saya’, ’merasa’, ’mengantuk’, ’dan’, ’sering’, ’letih’, ’kebingungan’, ’atau’, ’pusing’]

Table 4. **Stemming Result**s.

No.	Stemming Results
1	[‘nafas, ’saya’, ’pendek’, ’dan’, ’dada’, ’rasa’, ‘sakit’, ’kadang’, ’ikut’, ’dengan’, ’batuk’, ’batuk’]
2	[’tenggorokan’, ’terasa’, ’gatal’, ’dan’, ’kadang’, ’hidung’, ’saya’, ’mampat’]
3	[’sudah’, ’tiga’, ’hari’, ’ini’, ’saya’, ’rasa’, ’ngantuk’, ‘dan’, ’sering’, ’letih’, ’bingung’, ’atau’, ’pusing’]

Table 5. **Stopwords Removal.**

No.	Stopwords Removal
1	[‘nafas, ’pendek’, ’dada’, ’rasa’, ‘sakit’, ’ikut’, ’batuk’, ’batuk’]
2	[’tenggorokan’, ’terasa’, ’gatal’, ’hidung’, ’saya’, ’mampat’]
3	[’tiga’, ’hari’, ’saya’, ’ngantuk’, ’sering’, ’letih’, ’bingung’, ’pusing’]

Table 6. **Bio Dataset Format.**

Word	Label
nafas	B - SYM
saya	I - SYM
pendek	I - SYM
dan	I - SYM
dada	B - SYM
terasa	I - SYM
sakit	I - SYM
terkadang	O
diikuti	O
dengan	O
batuk	O
batuk	O

is often found in Indonesian.[7]

The use of the stemming process in each language is not the same. The stemming process in English text is different from stemming in Indonesian text. In English text, the process of removing the suffix is the most important thing. While the stemming process in Indonesian texts, not only suffixes are removed, but all prefixes, additions or affixes are also removed.

The accuracy value obtained in the stemming process using 30 data is 95.9%.

$$Prefix1 + Prefix2 + Rootword + Suffix3 + Suffix2 + Suffix1$$

d. The process of Stopwords Removal

The process of stopwords removal is intended to determine whether a word is entered into the stop-words or not. Disposing of stopwords in the process of removing meaningless or irrelevant terms. The terms obtained from the tokenization stage are checked in a stopwords list. If a word is included in the stopwords list, the word will not be processed further. Conversely, if a word is not included in the stopwords list, it will enter the next process. This is used to improve system performance so that it can process data efficiently.

The way to find out the symptoms of a disease in this research, where the words are formed from one or several words, a stopword is needed such as a symptom of “tidak pusing”, then the words will be filtered according to the stopwords list and will leave the words. “pusing”. The word “no” in Indonesian is included in the stopwords list so that tokenization can be carried out, and the results of the system managing the desired data are deemed incorrect. Overall, the accuracy of the stopwords removal process was 97.2%.

In this research, the stopwords list is static, which means that the stopwords list is obtained from a third party that provides open-source stopwords. These static stopwords contain common words used in other NLP projects in different languages. Because of this general characteristic, a static stopword list can reduce accuracy in the word filtering process.[8]

e. Named Entity Recognition

Named Entity Recognition (NER) is an algorithm used to extract information. It is part of Natural Language Processing (NLP). NER needs to be trained with datasets, where each dataset must be different. This study applies NER to identify indigestion symptoms by using the BIO (begin, Inside, Other) dataset format.[9] Each sentence in the dataset is processed and labeled - the label “B-SYM” is used to indicate a sign. If any word after the “B-SYM” label corresponds to the predecessor, we mark it with “I-SYM”. The “O” label is used for words that are not hymns and are not related to their predecessors.

Using 2000 lines consists of 100 sentences that have been separated by the label and word. An iteration with 800 epochs variables has been done to complete the training process. It reached a total micro-f1 score of 0.81. The simulation uses the complaints of the disease on input to test the model that has been trained.

The accuracy results obtained from the NER process that identifying symptoms of the disease using 30 input data was 74.3%.

Table 7. Search Results in ICD References Using Word "Asthma".

Search Result on ICD Dataset
'Asthma':
'Allergic asthma',
'Non-allergic asthma',
'Other specified forms of asthma or bronchospasm',
'Unspecified asthma',
' Respiratory tuberculosis, without mention of bacteriological or histological confirmation',
' Left ventricular failure, unspecified',

Table 8. Example of Search Results on ICD Dataset "That Does Not Match".

Search Result on ICD Dataset
'dizzy': [],' choky': []

Figure 2. Speech Recognition Interface.

Figure 3. NLP Process.

Error in word separation is considered wrong or unsuccessful, as in the 24th record. "Abdominal cramps" should be defined as a symptom. Instead, there is a separation between "cramping" and "stomach" meaning the model sees them as two distinct symptoms. Some words such as "cough" and "flu" in the 30th data are not included in the classification of gastrointestinal symptoms. However, the system considers these words to be symptoms of gastrointestinal disease. Therefore the following results are also considered incorrect.

Several factors can affect the identification process's accuracy, including the NER and the training dataset model used. In this research, the NER process uses the Bidirectional LSTM-CRF method because it is commonly used in NER studies that use Indonesian.

Wibawa (2016) reached a total F1-micro score of 0.50 for 15 classes of entities[10], Budi (2015) get a total F1-micro score of 0.67 using three entity classes[11], and Yudi Wibisono (2018) reached a total F1-micro score of 0.73 using four entity class.[12]

f. Data Translation

WHO ICD-11 reference data is only available in English.[13] So, in this paper, all the symptom data gathered in Bahasa are translated into English. However, there are still incorrect data because of the translation process from Bahasa to English.

g. Access ICD Data

The use of ICD data, is often linked to the process of searching for symptom classifications that match the input of disease symptoms. Therefore, access to ICD data is indispensable. Currently, its use refers to ICD-11. Disease symptom strings that have been through the translation process (previous test) are parsed and sent to the ICD. the ICD system will provide data that contain a classification of symptoms or diseases related to input parsed in the form of a JSON file.[14] If the translation data from the previous test are proper, the data obtained is related to the associated symptoms.

If the translation results are not proper, the data sent to ICD are unclear, so the feedback data is not proper. This condition can happen because there are two different languages processing, Bahasa as input while English as a reference on ICD.

h. Mobile UI

We upgraded the application from the web-based interface into a mobile user interface.[15] We divide the mobile design into three sections: the first section is for recording speech and converted it into written form. Users can see real-time voice conversion since speech recognition is also processed in realtime.[16]

The second section is meant to show the process of getting the symptoms. In this section, the word is processed with NLP starting from tokenizing, stemming, stopwords removal, and NER. after that, all that symptoms related to the digestive diseases system are shown.

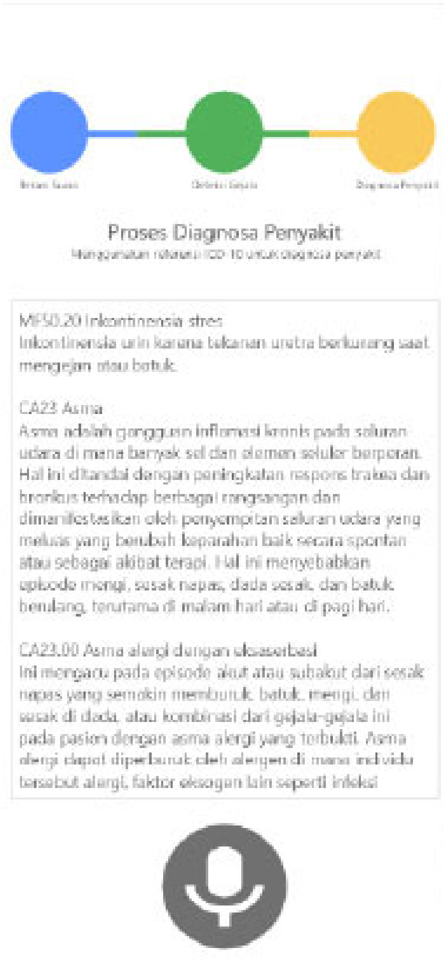

Figure 4. Result Interface.

RESULT

The last section is supposed to contains the diagnosis of disease symptoms based on ICD-11. In this section, the data that has been gathered from the last section is parsed and sent to the ICD. The parsed data is containing a list of symptoms that have passed the NLP process. The ICD database will give feedback data to classify symptoms of related diseases.

DISCUSSION

In the experimental stage of this research, two methods were used, namely, the NER method without using pre-processing for the process of identifying disease symptoms and using deep learning as an end-to-end learning tool. Furthermore, the second method used is NER with pre-processing applied to NLP.

The table above shows that the process of identifying disease symptoms using the NER method can detect words or groups of words that only contain information about disease symptoms.

Pre-processed NLP makes NER performance more efficient because less data is processed. Preprocessing separates the conjunction from the initial input data using stopwords removal and tokenization processes each word in each sentence using NER.

The output on the NLP pre-processing is slightly different. The difference lies in the availability of conjunctions in the two experiments. Data that do not have conjunction tends to be more informative because identifying symptoms requires only the word symptom and does not require conjunction.

Table 9. Data Input.

No.	Input
1	nafas saya terkadang sesak setelah melakukan banyak kegiatan
2	saya sering mengalami pusing, letih dan terkadang merasa kebingungan

Table 10. NER Result.

No	NER Result	
	Text	Type
1	nafas saya sesak	SYM
	melakukan banyak kegiatan	SYM
2	sering pusing	SYM
	letih	SYM
	merasa kebingungan	SYM

Table 11. Pre-process Result.

No.	Pre-process Result
1	nafas sesak karena banyak kegiatan
2	sering pusing dan letih terkadang merasa kebingungan

CONCLUSION

NLP research uses an information technology approach, especially for automatic speech recognition applications. The importance of the NLP process in the development of information, especially for word processing, aims as a device that simplifies speech recognition systems to be more helpful. This research focuses on the application mobile and development of the Named Entity Recognition (NER) model which is used to identify symptoms of digestive system diseases, where the resulting accuracy rate reaches 74.3%. In stemming and stopwords processing, the pre-processed NLP was still unable to filter the data accurately completely, and the accuracy rates were 95.9% and 97.2%, respectively. NER performance is dependent on changes in the data set, and the training process is undertaken. The difference between the data input (Indonesian) and ICD (English) in the reference operation status and translation process will cause incompatible parsed data. Further research with a different study design and a larger sample size is needed to find out more about other factors that affect the identification of ICD-11 symptoms based on clinical NLP mobile application to diagnose disease (ICD-11).

AUTHOR CONTRIBUTION

All authors contributed to this study's conception and design, data analysis and interpretation, article drafting, critical revision of the article, final approval of the article, and data collection.

FUNDING

This research has been funded by Penelitian Internal UNUSA Skema II with contract number: 362.23/ UNUSA/Adm-LPPM/V/2020.

CONFLICT OF INTEREST

There is no conflict of interest in this manuscript.

Table 12. NER Result with Pre-process Result.

No	NER Result	
	Text	Type
1	nafas sesak	SYM
	banyak kegiatan	SYM
2	pusing	SYM
	letih	SYM
	kebingungan	SYM

ETHICAL CONSIDERATION

This study has been declared ethical by the Ethical Commission for Health Research of Universitas Nahdlatul Ulama Surabaya.

REFERENCES

1. Gaebel W, Stricker J, Riesbeck M, Zielasek J, Kerst A, Meisenzahl-Lechner E, et al. Accuracy of diagnostic classification and clinical utility assessment of ICD-11 compared to ICD-10 in 10 mental disorders: findings from a web-based field study. Eur Arch Psychiatry Clin Neurosci. 2020;270(3):281–9.
2. Li J, Deng L, Haeb-Umbach R, Gong Y. Robust automatic speech recognition: A bridge to practical applications. 2015. 1–286 p.
3. Santra S, Bhowmick S, Paul A, Chatterjee P, Deyasi A. Development of GUI for Text-to-Speech Recognition using Natural Language Processing. In: 2018 2nd International Conference on Electronics, Materials Engineering & Nano-Technology (IEMENTech). 2018. p. 1–4.
4. Reshamwala A, Mishra D, Pawar P. IRACST Eng. Sci. Technol. An Int. J. None. 2013;3(1):113–116.
5. Pocai B. The ICD-11 has been adopted by the World Health Assembly. World Psychiatry. 2019;18:371.
6. Solangi YA, Solangi ZA, Aarain S, Abro A, Mallah GA, Shah A. Review on natural language processing (NLP) and its toolkits for opinion mining and sentiment analysis. In: 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS). 2018. p. 1–4.
7. Rizki AS, Tjahyanto A, Trialih R. Comparison of stemming algorithms on Indonesian text processing. Telkomnika. 2019;17(1).
8. Gerlach M, Shi H, Amaral LAN. A universal information theoretic approach to the identification of stopwords. Nat Mach Intell. 2019;1(12):606–12.
9. Lignos C, Kamyab M. If You Build Your Own NER Scorer, Non-replicable Results Will Come. In: Proceedings of the First Workshop on Insights from Negative Results in NLP. 2020. p. 94–9.
10. Wibawa AS, Purwarianti A. Indonesian named-entity recognition for 15 classes using ensemble supervised learning. Procedia Comput Sci. 2016;81:221–8.
11. Silalahi M, Cahyani DE, Sensuse DI, Budi I. Developing indonesian medicinal plant ontology using socio-technical approach. In: 2015 International Conference on Computer, Communications, and Control Technology (I4CT). IEEE; 2015. p. 39–43.
12. Wibisono Y, Khodra ML. Pengenalan entitas bernama otomatis untuk Bahasa Indonesia dengan pendekatan pembelajaran mesin. 2018;
13. Névéol A, Robert A, Anderson R, Cohen KB, Grouin C, Lavergne T, et al. CLEF eHealth 2017 Multilingual Information Extraction task Overview: ICD10 Coding of Death Certificates in English and French. In: CLEF (Working Notes). 2017.
14. Muslim A, Mutiara AB, Suhendra A, Oswari T. Expert mapping development system with disease searching sympthom based on ICD 10. In: 2018 Third International Conference on Informatics and Computing (ICIC). IEEE; 2018. p. 1–4.
15. Putra FB, Yusuf AA, Yulianus H, Pratama YP, Humairra DS, Erifani U, et al. Identification of Symptoms Based on Natural Language Processing (NLP) for Disease Diagnosis Based on International Classification of Diseases and Related Health Problems (ICD-11). In: 2019 International Electronics Symposium (IES). 2019. p. 1–5.
16. He Y, Sainath TN, Prabhavalkar R, McGraw I, Alvarez R, Zhao D, et al. Streaming end-to-end speech recognition for mobile devices. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2019. p. 6381–5.