

Article

Predicting ICD-9 Codes Using Self-Report of Patients

Anandakumar Singaravelan ¹, Chung-Ho Hsieh ², Yi-Kai Liao ³ and Jia-Lien Hsu ^{3,*} 

¹ Graduate Institute of Applied Science and Engineering, Fu Jen Catholic University, New Taipei City 242062, Taiwan; 408068080@mail.fju.edu.tw

² Department of General Surgery, Shin Kong Wu Ho-Su Memorial Hospital, Taipei 111045, Taiwan; M012363@ms.skh.org.tw

³ Department of Computer Science and Information Engineering, Fu Jen Catholic University, New Taipei City 242062, Taiwan; calvinliao@fju.edu.tw

* Correspondence: alien@csie.fju.edu.tw

Abstract: The International Classification of Diseases (ICD) is a globally recognized medical classification system that aids in the identification of diseases and the regulation of health trends. The ICD framework makes it easy to keep track of records and evaluate medical data for evidence-based decision-making. Several methods have predicted ICD-9 codes based on the discharge summary, clinical notes, and nursing notes. In our study, our approach only utilizes the subjective component to predict ICD-9 codes. Data cleaning and segmentation, and Natural Language Processing (NLP) techniques are applied on the subjective component during the pre-processing. Our study builds the Long Short-Term Memory (LSTM) and the Gated Recurrent Unit (GRU) to develop a model for predicting ICD-9 codes. The ICD-9 codes contain different ICD levels such as chapter, block, three-digit code, and full code. The GRU model scores the highest recall of 57.91% in the chapter level and the top-10 experiment has a recall of 67.37%. Based on the subjective component, the model can help patients in the form of a remote assistance tool.



Citation: Singaravelan, A.; Hsieh, C.-H.; Liao, Y.-K.; Hsu, J.-L. Predicting ICD-9 Codes Using Self-Report of Patients. *Appl. Sci.* **2021**, *11*, 10046. <https://dx.doi.org/10.3390/app112110046>

Academic Editor: Keun Ho Ryu

Received: 23 August 2021

Accepted: 24 October 2021

Published: 27 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: ICD-9; medical record; LSTM; GRU

1. Introduction

Deep learning has significantly enhanced computational efficiency in recent years, as a result of continuous scientific research and the upgrading of hardware requirements, and it has emerged as the most vigorous technology in the field of artificial intelligence. Many applications are created using deep learning algorithms with a large amount of data and some medical applications are invaluable towards healthcare system. Our study aims to produce a ICD-9 prediction model by only utilizing subjective component. The study implements two distinct deep learning approaches to produce a prediction model. The prediction model behaves as an assistance tool that helps to know the ICD-9 codes just before approaching the hospital. This model can also act as a clinical support tool to aid medical professionals. Most of the existing medicine recommendation systems aid doctors to choose reliable clinical decisions [1]. Early prediction of any disease can help to move further towards life safety. A recent study develops a computational model to predict sepsis at the early stage [2]. There are various techniques involved in predicting the recurrence of breast cancer to assist the process of decision making [3–7]. EHRs contain massive storage of data that are invaluable for healthcare system [8]. Utilizing the massive amount of data, decision making systems aids to contemporary healthcare system.

The World Health Organization (WHO) categorizes an International classification of diseases (ICD-9) for disease maintenance [9]. ICD-9 is widely used in all localities of healthcare to report diagnosis. The ICD-9 system assigns a specific code to each illness. The physicians assign the codes based on the Electronic Health Records (EHR) for determining severity and tracking diseases. Some countries acquire ICD-9 codes for compute billing and estimating severity [9]. Medical professionals assign the ICD-9 codes in order to support a

doctor's clinical diagnosis. In this technological age, it is essential to be able to predict ICD codes [10,11]. Thousands of ICD codes are used by the international medical community to classify illness, accident, and cause of death [12].

Many countries consider ICD as the tool to follow up death and statistics of death. The ICD system has more than one version, which depends on the situation in different countries and their clinical conditions [13]. Various countries have adopted the ICD-10, including the United States. The United States follows ICD-10-CM and Canada follows the ICD-10-CA in their health care system. In Taiwan, the ICD-9-CM, which has been in use since 1994, has now changed to ICD-10-CM. There are some modifications between ICD-9 to ICD-10, such as attaching laterality and recategorization [14]. However, this study focuses on the ICD-9 code prediction. Since our medical record data still follows ICD-9, in this study we focus on the ICD-9 code prediction. In addition, our proposed method is generic to ICD-9 and ICD-10. The ICD-9 codes contain various levels such as chapter, block, three-digit code, and full code. There are more than 13,000 distinct codes of the ICD-9 system, which are divided into 17 chapters of disease codes and 2 chapters of trauma and supplemental classifications. Further, 17 chapters are divided into 135 blocks which contain the disease codes. Figure 1 demonstrates the different levels of ICD code. The hierarchical table of code 765.02 is mentioned for better understanding in Table 1. ICD-9 format is represented by using five digit code, the first three digits are called category and the last two digits indicate the cause (Etiology), anatomic site (Anatomic site), and symptom sign (Manifestation); Figure 2 clearly demonstrates the format using the same code.

In this study, we collaborate with a medical center to develop our prediction model. Outpatient records, emergency medical records, and inpatient medical records are the three types of medical records that are commonly used. Clinical medical records include information such as hospitalization records, medication descriptions, test reports, analysis records, treatment plans, and derived ICD codes. Whereas admission records, hospitalization course records, appointment records, intrusive diagnosis or procedure records, surgical records, anesthesia records, consent documents, prescription records, antibiotic use, progress notes, and discharge summary are found in a hospitalization medical record. The POMR (Problem-Oriented Medical Record) is a recording approach of medical data; It consist of four parts, which include database, problem list, initial plan, and progress note. The POMR follows the format of SOAP, which consists of subjective component, objective component, assessment, and plan [13,15]. In the SOAP format, the subjective component is said to be a training data, which includes feelings, opinions, and complaints of patients.

The contributions of this study are two things. The major contribution is the construction of deep learning-based approaches to forecast ICD-9 codes by only applying subjective components. The second is that the study predicts 1871 ICD codes and especially scores 57.91% of recall in the chapter level.

Table 1. Hierarchy of ICD code—765.03.

ICD-9 Code—765.03			
Level	Level Definition	Code Range	Description
1	Chapter	760–779	Certain conditions originating in the perinatal period
2	Block	764 –779	Other conditions originating in the perinatal period
3	3-digit category code	765	Disorders relating to short gestation and low birth weight
4	Full code	765.03	Extreme immaturity, 750—999 grams

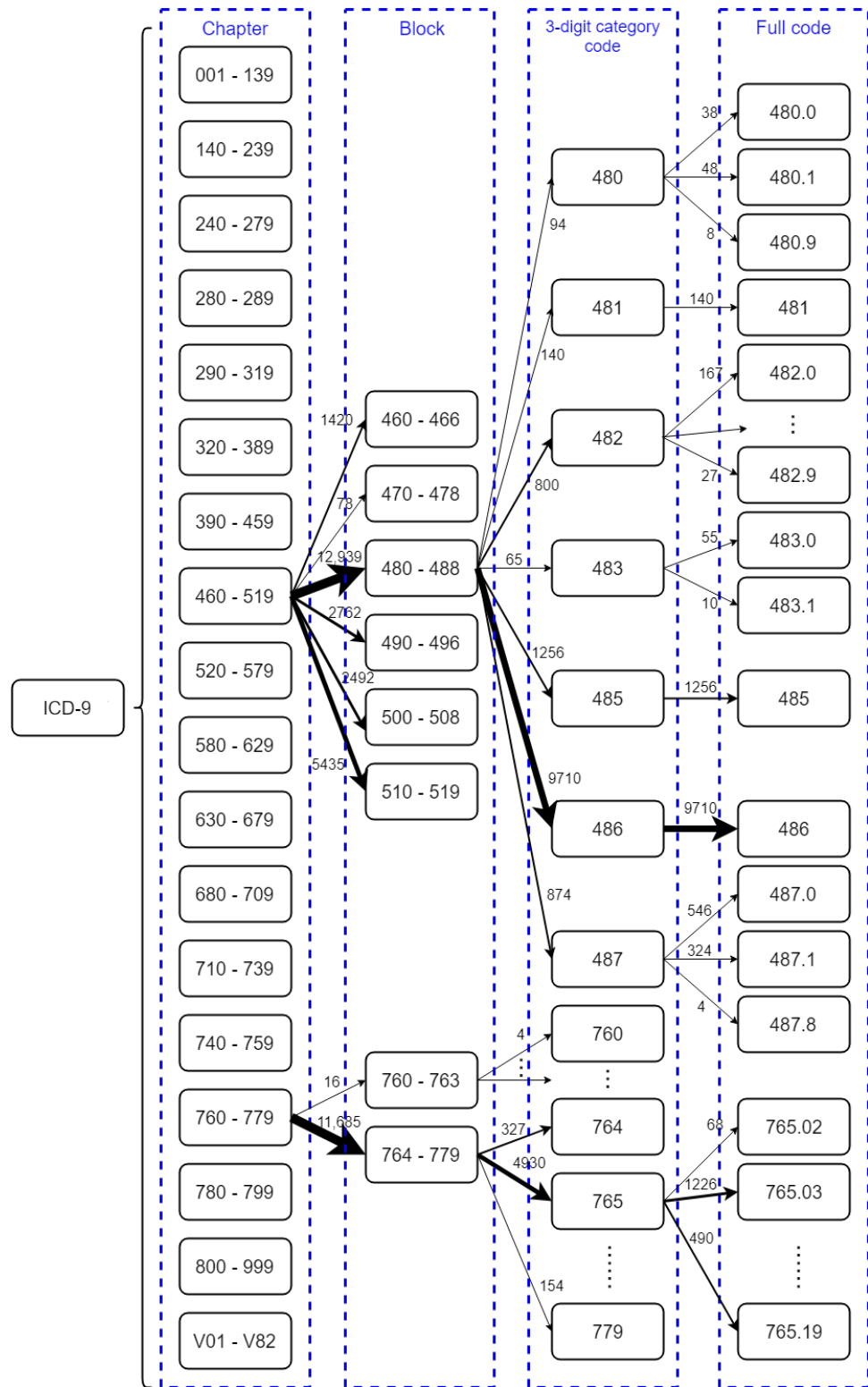


Figure 1. ICD-9 Structure.

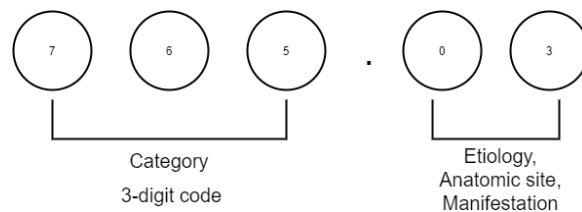


Figure 2. The format of ICD-9 code.

2. Related Work

Multi-label classification approach captures raw nursing notes to predict ICD-9 chapters [16]. The majority of studies utilize nursing notes for the ICD-9 code prediction [17,18]. In our study, a subjective component is used as training data to predict disease codes. This model aids doctors in deciding relevant ICD-9 code between the vast number of ICD codes available. Our model aids the medical personnel in choosing a reliable ICD-9 code. Additionally, the concept is used in a variety of applications, including medical chat-bots. At the same time, a deep learning based model for ICD-10 is emerging in the current scenario. The recommendation system for ICD-10 is implemented using a GRU based model [19]. Previous research creates a deep learning model to minimize human effort and automatically detect ICD-10 codes. This model is based on diagnosis records to forecast the ICD codes [20]. A recent study performs an ICD-10 automatic coding for primary diagnosis in the Chinese context. This method works with the help of discharge procedure and diagnosis texts and performs well in cardiovascular diseases [21].

A recent study developed an app that aids medical personnel in predicting incisional hernia occurrence [22]. ICD-9 codes are utilized to cross verify the eligible patients. Their study depicts a perfect demonstration of transferring an institutional dataset into a application. Our prediction model can assist in developing a remote assistance type of application. Prior research proposes an ICD-9 prediction method that uses a transfer learning approach to transform MeSH index information into ICD-9 codes [23]. A previous study proposes an approach for assigning ICD-9 codes to track a disease history of patients, which helps in the billing system [18]. In this process, the ICD-9 code prediction is made based on the clinical notes. Clinical notes are used to classify the top 10 ICD-9 codes and blocks. Enhanced Hierarchical Attention Network (EnHAN) [24] utilizes discharge summary to solve ICD-9 prediction problems. To deal with multi-class label problems, the method uses topical word embedding. Moons et al. [25] solve the ICD-9 prediction using discharge summaries. Multiple approaches are used in the study to identify ICD-9 codes. On the discharge summary, Word2Vec is used in the multi-label classification [26]. A recent study uses TF-IDF in the process of generating embedding vectors [27]. The CNN is used to identify ICD-9 codes in their process. On the subjective components, we have used Word2Vec in our study. In addition, we use LSTM and GRU to predict ICD-9 codes in our study.

In our previous study [13], we employed a CNN based network to predict ICD-9 code, which scores the recall of 58% in chapter level. In this work, we applied the LSTM and GRU network to predict ICD codes. We removed the list of stop words and applied TF-IDF in this approach. Similarly, this approach scores a recall of 57.54%. In addition, the top-10 prediction model scores a recall of 67.37%.

The prior study takes its base knowledge from ontologies to understand clinically related features for developing a robust deep neural framework that achieves disease diagnosis [28]. The SVM classifier specifically predicts ICD-9 codes of mechanical ventilation, phototherapy, jaundice, and Intracranial hemorrhage using ICU notes. N-gram feature extraction methods are utilized in their approach [29]. Previous studies evaluate supervised learning approaches to predict 1231 ICD-9 codes using the EMR dataset. In this method, the EMR is gathered from three types of datasets that have differences in the number of codes and size of the EMRs. This study implements the model with the discharge summary in the prediction process [30]. Prior research proposes a hierarchical model along with an attention mechanism to assign ICD codes with the use of diagnosis description. This

study assigns just 50 ICD-9 codes and scores 0.53 of F1 scores [31]. Our study handles 1871 codes in the process of ICD prediction. Another study deals with diagnosis description to forecast ICD-9 codes. The study creates a neural structure for automatic ICD coding. The neural structure scores 0.29 of sensitivity. The study handles 2833 ICD codes in the process of computing ICD code prediction [32]. The reason behind less sensitivity is that the study involves a large number of ICD-9 codes in the prediction. FarSight, a method with long-term integration technique to find the onset of the issue with early symptoms. This research consumes unstructured nursing notes to produce 19 ICD-9 chapter codes [33].

With the advancement of the medical field, a vast number of electronic health records (EHRs) are exchanged with healthcare providers in order to improve medical services. H. Li takes advantage of EHR data to develop a reliable bone disease prediction model [34], by analyzing and interpreting the data in the EHR, which helps to develop a medical system. A previous study was designed to predict heart disease based on that data [35]. To train our prediction model, we have used subjective components that reflect a patient's feelings about their illness. The prediction model can work as a self assistance tool that aids to identify ICD-9 codes.

Table 2 shows the comparison of our study with other approaches. In this study, we predict ICD-9 codes with the help of subjective components. This comparison shows the uniqueness of our study.

Table 2. Comparison with other approaches.

Method	Approach	Dataset	Input	Target	Performance	Taxonomy
Marafino et al. [29]	Traditional	MIMIC-II	Noteset (All ICU notes)	VENT, PHOTO, JAUND, ICH	Accuracy: 0.98 (VENT), 0.94 (PHOTO), 0.89 (JAUND), 0.93 (ICH)	Full code
Kavuluru et al. [30]	Traditional	UKLarge	Discharge Summary	1231 ICD-9 codes	Micro F1 score—0.479	Full code
Shi et al. [31]	Deep Learning	MIMIC-III	Diagnosis Description (Discharge Summary)	50 ICD-9 codes	F1 score—0.53	Full code
Xie et al. [32]	Deep Learning	MIMIC-III	Diagnosis Description	2833 ICD-9 codes	Sensitivity—0.29, Specificity—0.33	Full code
Huang et al.	Deep Learning	MIMIC-III	Discharge Summary	10 ICD-9 codes and 10 blocks	F1 score: Full code—0.69, ICD-9 block—0.723	Block, Full code
Zeng et al. [23]	Deep Learning	MIMIC-III	Discharge Summary	6984 ICD-9 codes	Micro average F1 score—0.42	Full code
Samonte et al. [24]	Deep Learning	MIMIC-III	Discharge Summary	10 ICD-9 codes	Recall score—0.620; F1 score—0.678	Full code
Hsu et al. [26]	Deep Learning	MIMIC-III	Discharge Summary	Chapters (19), 50 and 100 ICD-9 codes	Micro F1: 0.76—Chapter; Full code: 0.57-top-50; 0.51-top-10	Chapter, Full code
Gangavarapu et al. [16]	Deep Learning	MIMIC-III	Nursing notes	19 Chapters	Accuracy—0.82	Chapter
Gangavarapu et al. [33]	Deep Learning	MIMIC-III	Nursing notes	19 Chapters	Accuracy—0.83	Chapter
Our method	Deep Learning	Medical Center	Subjective component	1871 ICD-9 codes	Recall score: Chapter—0.579, Block—0.492, Three-digit code—0.430, Full code—0.405	Chapter, Block, Three-digit code, Full code

Notes: VENT—Ventilation, PHOTO—Phototherapy, JAUND—Jaundice, ICH- Intracranial hemorrhage.

3. Materials and Methods

3.1. The Data

The entire data holds a total of 146,343 medical records. The data include 11 fields of attributes, which are hospitalization number, date, time, medical record number, author, subject, ICD code, subjective component, objective component, assessment, and plan. The

subjective components mainly record the patient's emotions and opinions regarding the illness. In this study, we use the subjective component as the training data. According to the dataset, the listed data ranges between 2012 and 2017. In around 140,000 medical records, there are 1871 different disease codes, which are distributed in the 17 chapters of ICD-9 in the supplementary category. In our data, there are 24 types of disease codes with more than 1000 medical records, which are approximately 40% of the total data. There are 234 types of disease codes with more than 100 medical records, which are approximately 80% of the total data volume. Level 1 (chapter) consists of a total of 17 chapters with a supplementary category, level 2 (block) holds 128 codes, level 3 (three-digit code) holds 624 codes, and level 4 (full code) consists of 1871 codes.

In our dataset, the majority of records is from respiratory diseases (chapter 8). More than 25,000 (17%) medical records are related to the respiratory diseases. Secondly, tumors (chapter 2) data has a higher volume in our data set with a proportion of 15%. Subsequently, circulatory system diseases (chapter 7) consist of 13% of data; Digestive system diseases (Chapter 9) hold 12% of data in the total data set. Further, chapter 18 has accumulated more than 6000 medical records.

3.2. Word2Vec

Word2vec is a model that aids in the translation of vocabularies into vector representations [36,37]. The core concept of the code suite is derived from the concept of word vectorization [38]. The Word2Vec model consists of CBOW (Continuous Bag-of-Words) and skip-gram model. Based on the previous and next words, the CBOW predicts the current target words. The input layer intakes the previous and next word to produce the (current) target word. On the other hand, the skip-gram model predicts the previous and next word using the current target word as an input. The input layer intakes the target word as a input to produce the previous and next word as a output. According to the semantic sense, the Word2vec model transforms each vocabulary into a word vector. As a result, similar words are grouped together in a high-dimensional space; Figure 3 clearly illustrates the process of Word2vec.

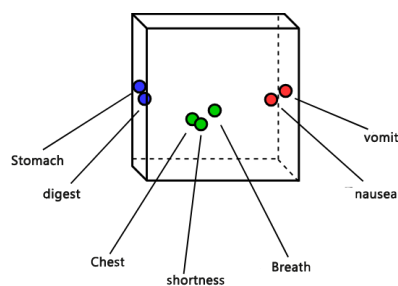


Figure 3. The visualization of Word2vec.

3.3. Data Cleaning and Word Segmentation

Numerous numerical values, such as record time, date, and so on, are used in subjective components. The first step in pre-processing is used to remove numeric characters from the text. Subsequently, we convert all English words to lowercase letters; We remove all punctuation marks and special symbols to achieve more reliable word segmentation results. In the segmentation stage, we used the jieba segmentation suite (Jieba: <https://github.com/fxsjy/jieba> (accessed on 3 September 2019)) which is widely used in the field of Chinese word segmentation. The native kit is developed based on simplified Chinese. In our research, subjective components are based on traditional Chinese. Accordingly, we have utilized jieba-tw (Jieba-tw: <https://github.com/APCLab/jieba-tw> (accessed on 3 September 2019)) to solve this problem. The total number of words in our text, after the data cleaning and word segmentation is 27,196. We also added English Stop Words (ESW) and Chinese Stop Words (CSW) to the stop word list in this study. We involved

custom stop words such as One Count Stop Words (OCSW), One Character Stop Words (OCharSW), TF-IDF (Term Frequency-Inverse Document Frequency) stop words (TSW), and IDF stop words. OCharSW are used to deal with special symbols that contain a single character. The results of word segmentation are considered as an input for OCSW, which handle statistically trivial words that are used only once.

3.4. Term Frequency-Inverse Document Frequency

Additionally, we utilize the approach of TF-IDF [39] to calculate the frequency of each word in all texts to observe the significance of each word in each medical record.

TF in TF-IDF stands for Term Frequency, which is represented by $tf_{t,d}$, which is depicted in Equation (1), where the subscript t represents a specific word (term), d represents a specific article (document), $n_{t,d}$ represents the number of occurrences of the word t in the article d ; $\sum_k n_{k,d}$ represents the sum of the number of occurrences of all words in the article d . The TF value can help us to understand the frequency of a word in an article. IDF stands for Inverse Document Frequency, which is represented by idf_t , which is depicted in Equation (2), where df_t represents the number of articles containing the vocabulary t , and N represents the total number of articles. The IDF value can help us to understand the frequency of a word in all articles.

$$tf_{t,d} = \frac{n_{t,d}}{\sum_k n_{k,d}}, \quad (1)$$

$$idf_t = \log \frac{N}{df_t}, \quad (2)$$

$$TF-IDF_{t,j} = tf_{t,d} \times idf_t. \quad (3)$$

The value of TF-IDF is the result of multiplying TF and IDF, which is depicted in Equation (3). The larger value represents the higher importance of the word in the article. The TF-IDF value of each word in each medical record is different, we take the sum of the value of each word in different medical records and take the average as the criterion, and rank them according to this value. Consequently, we can find from this ranking that the words with lower weights are not directly related to the disease code or that the word segmentation process has not been processed properly.

3.5. Word Encoding and Word Embedding

After pre-processing, our medical record text data contains a total of 14,767 words with the longest medical record, which comprises 147 words. Subsequently, we build a dictionary with all of the words in the text and their corresponding number. As a result, each word is represented by a number, and each text medical record is then converted into a vector representation. After the text is transcoded, we utilize Word2vec to convert the text of the medical record into a vector. We organize words that have the same meaning, which are then allocated to a similar place in the higher dimension space. The common length is stretched in order to use the embedding layer to translate the encoding into a vector.

Medical documents have a maximum word count of 147, which is the regular limit for all texts. As a result, short texts are padded in order to meet the limit. Finally, one hot encoding is used to convert each word into a one-dimensional vector. Each word vector becomes one hot vector with a length of 14,767 after the above conversion process. As a result, the input dimension of the embedding layer is 14,767, and the output dimension is 300. Each medical record vector contains 147 word vectors, which represents the maximum length of training data.

3.6. Deep Learning Methods

We utilize the pre-processed data to implement LSTM and GRU. Figure 4 depicts the model architecture, which includes both LSTM and GRU. The architecture of the two

models is identical. The input of network is a subjective component. ICD-9 codes are the output of this network. A total of 1871 ICD-9 codes are predicted using our model. The main difference lies in the different parameter settings of LSTM and GRU units and detailed model parameter settings.

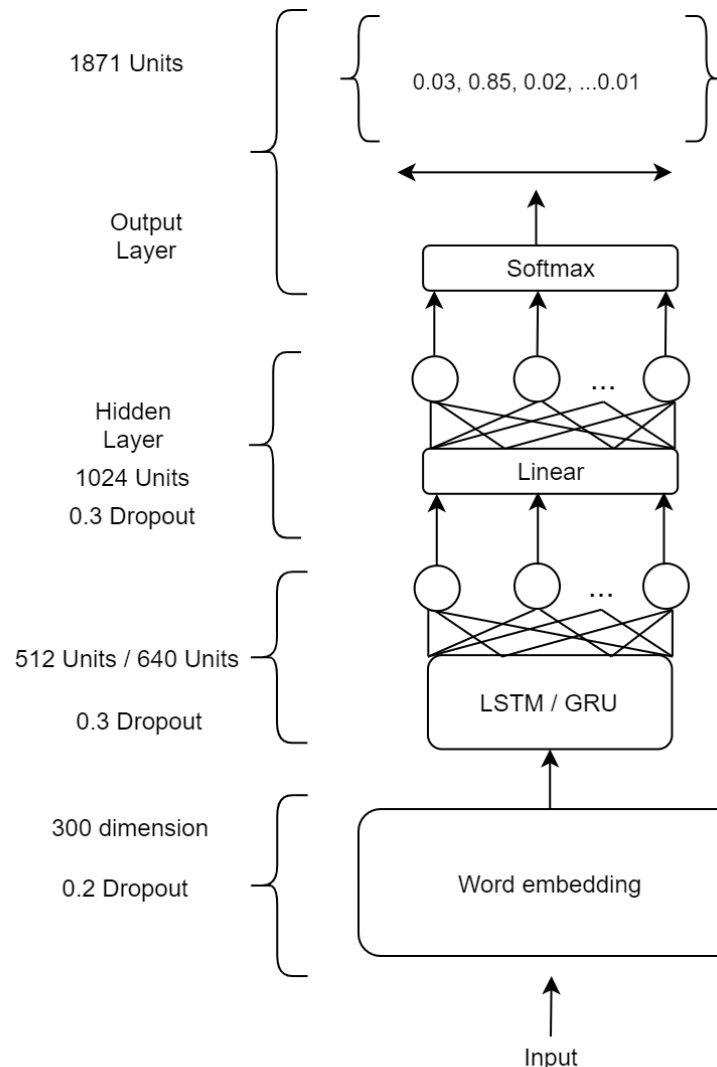


Figure 4. The architecture of our Model.

4. Experiments

Our study utilizes two distinct deep learning approaches to achieve a model. In this study, we use the Keras 2.1.1 (tensorflow backend) package and Python 3.6 to implement our model. The different hyper-parameters and experimental settings are used for each approach. The output dimension of an embedding layer in both networks is 300. LSTM units are set to 512. GRU units are set to 640. At this stage, the dropout is 0.3 in both networks. Linear activation function is applied in LSTM and GRU networks. The learning rate is 10^{-5} in this study. Adam optimizer is applied in both networks. We use the Keras model in our experiment. The categorical accuracy in the group is used as the evaluation standard. Finally, 10-fold cross-validation is applied on the best model and evaluation methods, such as precision and f1 score, are added for comparison. The baseline is considered as an evaluation benchmark. In this study, the highest numbered disease code is 486 (pneumonia) with a total of 9710 cases and the ratio is 6.64% of the overall data. The confusion matrix is shown in the Table 3.

Table 3. Confusion matrix.

	Positive Prediction	Negative Prediction
Positive actual class	True positive (TP)	False negative (FN)
Negative actual class	False positive (FP)	True negative (TN)

4.1. LSTM

We use the ESW, OCSW, OCharSW, and other stop words as the basis for excluding the stop words in the LSTM experiment. This time, the highest number of words in the medical record is 158. The number of words in the longest medical record is decreased from 158 to 147 after adding TSW, and the recall is increased a little bit. LSTM units are set to 512 for achieving the best performance.

The network settings are constantly adjusted during the experiment to improve the results. In the settings, the activation function of the hidden layer is changed from the general function. Replacing the ReLu function with a linear function helps to increase the result by more than 3%. Table 4 shows the experiment results of LSTM.

Table 4. Experiment results of LSTM.

Epochs	Recall	Stop Word List
30	34.65%	Basic
60	35.10%	Basic + TSW
60	38.38%	Basic + Linear f(x)
60	38.93%	Basic + Linear f(x) + TSW

4.2. GRU

The specifications of the input and output data are exactly the same as the previous experiment. In this experiment, the embedding output setting is the same as the LSTM experiment. Accordingly, the output setting remains at 300 and the hidden layer activation function also uses a linear function. The major differences between LSTM and GRU are identified during the process of training. In a shorter time, GRU has reached convergence with less repetitive training times. The LSTM produces a good result after adding TSW in the list of stop words. On the other hand, GRU achieves a better result by retaining data without using TSW. The GRU experiments are listed in Table 5.

Table 5. Experiment results of GRU.

Epochs	Recall	Notes
40	39.59%	Basic
60	38.94%	Basic + TSW
40	40.22%	Only OCSW

5. Results

The GRU experiment serves as a benchmark for determining model parameters in our experiments. Furthermore, 10-fold cross validation is carried out in this study, and the results of 10-fold validation are reported in Table 6. We have achieved the highest recall of 40.54%, the highest precision of 42.37% and the highest f1 score of 39.86% in the 10-fold validation. Moreover, our research finds that the averaged recall is 40.01%, precision is 41.99%, f1 score is 39.33%, and baseline is 6.63%. The results of the 6-fold experiment obtains the best prediction results on the two evaluation criteria of recall and f1 score. Therefore, subsequent experimental analyzes are performed using the model trained in this experiment.

Table 6. The performance of 10-fold cross validation.

Fold	Recall	Precision	F1 Score	Baseline
1	39.85%	41.98%	39.18%	6.79%
2	40.21%	41.91%	39.40%	6.75%
3	39.80%	42.09%	39.06%	6.55%
4	39.89%	42.04%	39.07%	6.53%
5	39.88%	41.89%	39.29%	6.89%
6	40.54%	42.25%	39.86%	6.61%
7	40.41%	42.37%	39.76%	6.70%
8	39.44%	41.77%	38.96%	6.36%
9	40.09%	41.70%	39.38%	6.83%
10	39.98%	41.91%	39.38%	6.33%
Average	40.01%	41.99%	39.33%	6.63%

5.1. ICD Coding Prediction Results of Each Level

In this study, the GRU model is utilized to predict disease codes at each level. The levels of ICD code prediction are chapter, block, three-digit code, and full code. Table 7 shows the prediction results of each level. In our study, the full code has scored a recall of 40.54%. The three-digit recall reaches 43.05% higher than the previous level. Comparatively, the block has exceeded the full code and three-digit code with a recall rate of 49.20%. Overall, the chapter has the highest recall rate of 57.91%.

Table 7 shows the prediction results of each level, where the recall is increased by 2% from level 4 (full code) to level 3 (three-digit code). Level 3 is surpassed by level 2 (block), a nearly 6% rise. Level 2 increased from level 1 (chapter) with an improvement of almost 9%.

Figure 1 depicts the ICD-9 structure, which demonstrates the data imbalance. The disparity is caused by the accumulation of a large number of medical records in a few specific disease codes. For instance, the highest numbered disease code is 486 (pneumonia), which has 9710 entries. The code 486 seems the same in both levels, the three-digit code and full code. The number of disease codes are increased from level 4 to level 1. For instance, the chapter has 18 codes and the block has 128 codes. Similarly, the number of codes are increased up to the full code, which has 1871 number of codes. The imbalance causes a deficiency in the prediction performance at higher level.

Figures 5–8 depict a variety of confusion matrices for the top 20 disease codes of various level. The actual codes are listed vertically, the predicted codes are listed horizontally, and the normalized outcome of the correct prediction is listed diagonally. Equation (4) is used to calculate the normalized result. The higher value represents a higher prediction rate. The weighted average of the correct prediction values determines the average score. The figures clearly show that our model has achieved reliable results in predicting chapters with high accuracy.

$$\text{Normalized} = \frac{\text{Correctly predicted number}}{\text{Total case number}} \quad (4)$$

Table 7. ICD code details.

Level	Level Definition	Codes	Recall	Baseline
4	Full code	1871	40.54%	6.61%
3	Three-digit code	624	43.05%	6.61%
2	Block	128	49.20%	8.69%
1	Chapter	18	57.91%	17.19%

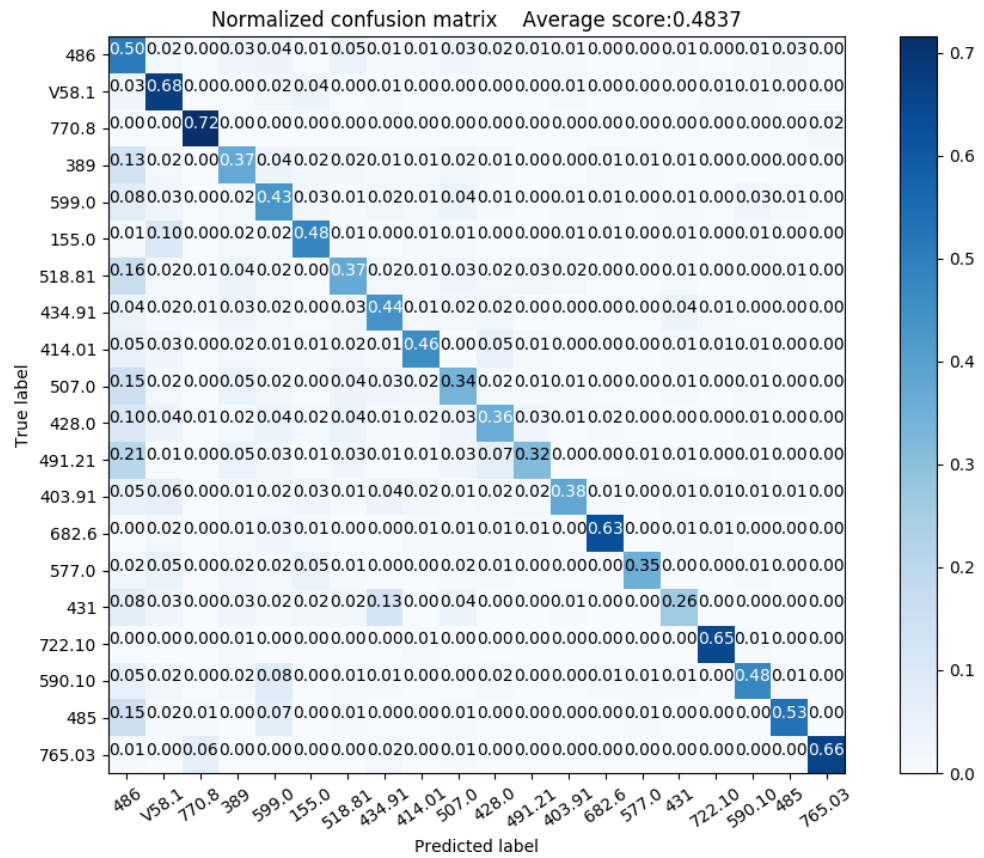


Figure 5. Confusion matrix of full code prediction.

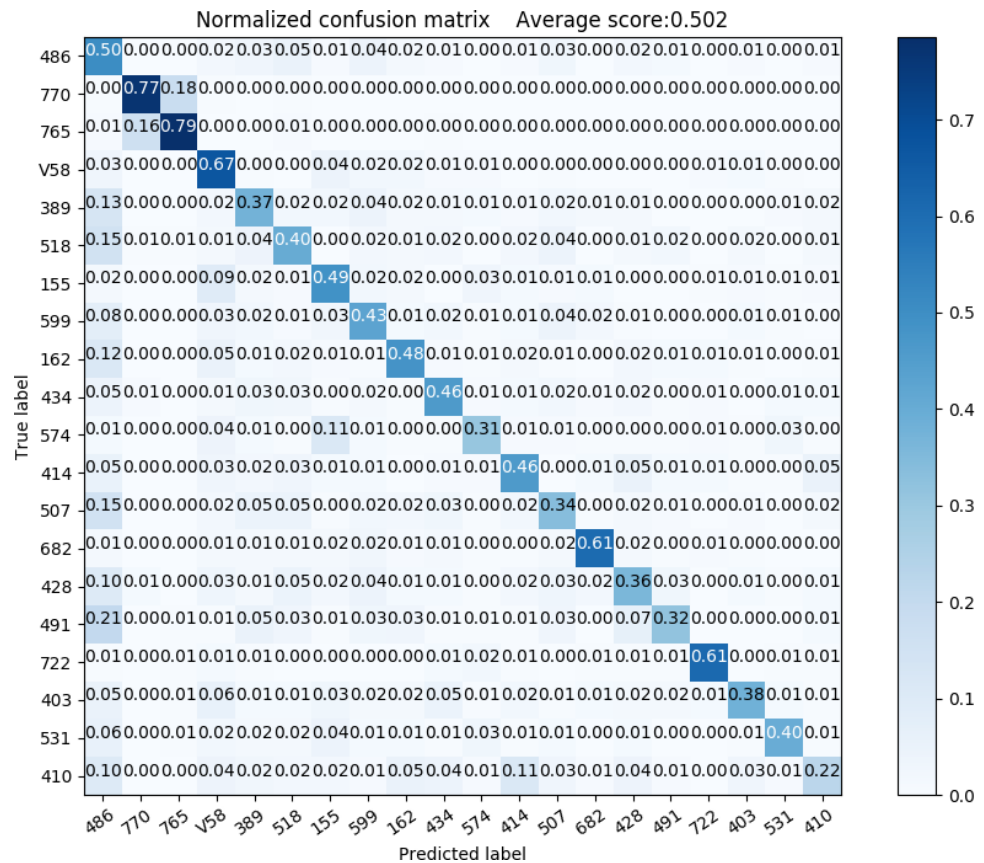


Figure 6. Confusion matrix of three-digit code prediction.

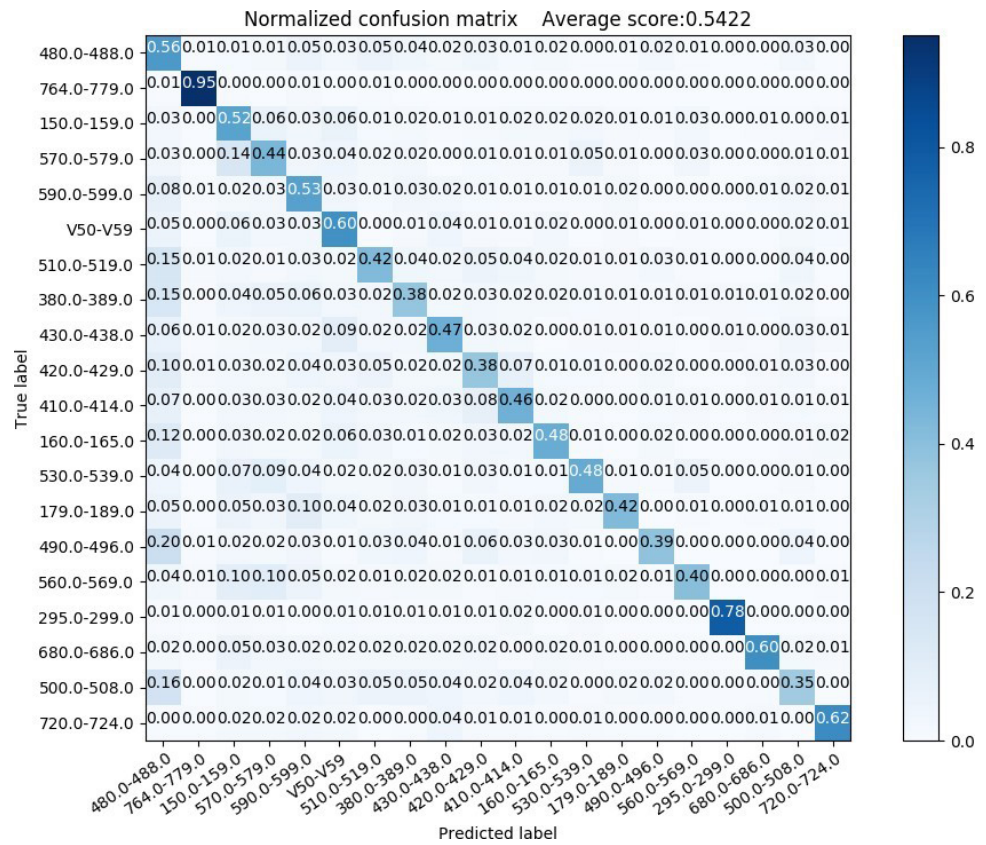


Figure 7. Confusion matrix of block prediction.

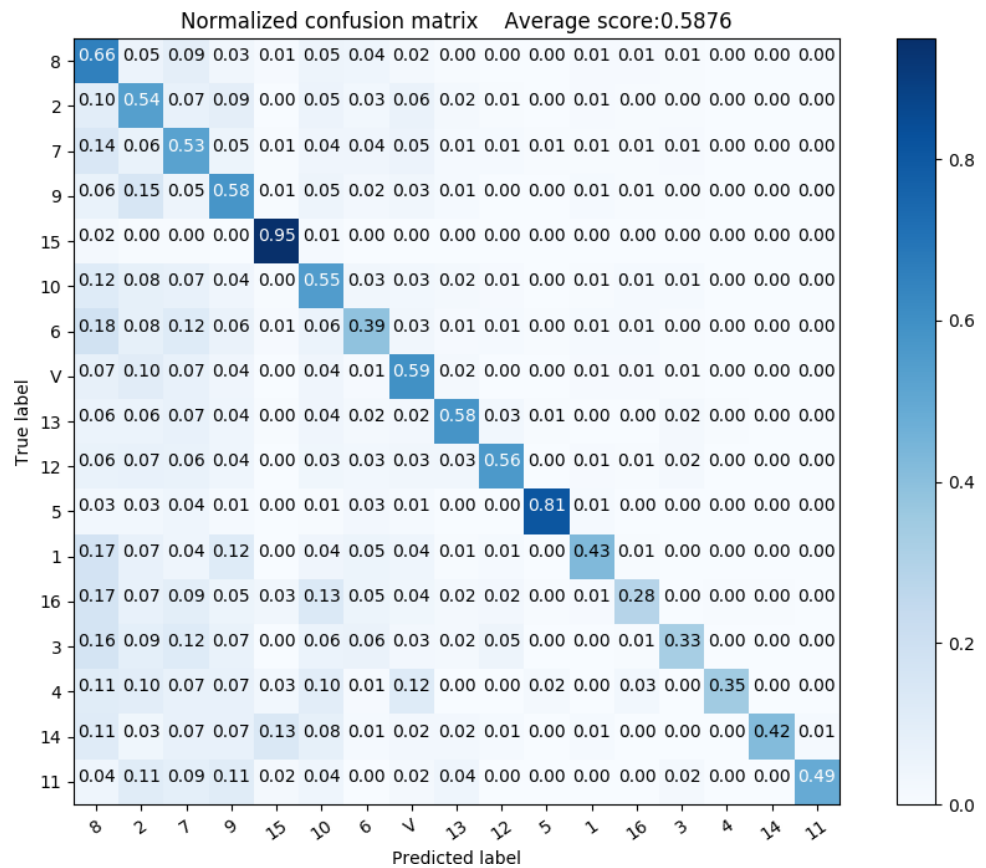


Figure 8. Confusion matrix of chapter.

5.2. Top-K

ICD-9 contains more than 13,000 disease codes. In this study, there are more than 140,000 medical records with a total of 1871 different codes, which belong to many categories. Prior research [40] classifies and predicts ICD-9 disease codes, however they classify only 45 ICD-9 codes. Another study predicts a total of 1778 ICD-9 codes in their research [41], which is nearly similar to the number of codes in this study. The Top-10, Top-20, and Top-30 evaluation models are used, respectively. As a result, professional physicians suggest to increase the number of predictive codes in the model, which is referred to as a top-K experiment, where the K value represents the number of codes to be predicted.

A one-dimensional vector with a length of 1871 is the outcome of this experimental model. The probability value of the input data belonging to 1871 label data is represented by each value in the vector, which ranges from 0 to 1. As a result, if the model only predicts one code as the output response, the mark with the highest probability value in the vector can be chosen as the top 1 prediction result. The labels with the top 3 probability values in the vector are selected in the prediction process. The model predicts three codes as potential output responses in the top 3, and so on in the top 5 and top 10. Table 8 shows the prediction results of the top-K experiment.

Table 8. The results of top-k prediction.

Top-K	Recall	Baseline
Top 1	40.54%	6.61%
Top 3	53.38%	12.63%
Top 5	59.64%	17.61%
Top 10	67.37%	27.04%

5.3. High Proportion of Disease Prediction Results

Uneven data distribution is an issue that is encountered among the disease prediction in the medical field [42,43]. The percentage of data for the most common diseases is frequently much higher than the rare diseases. Our study is not an exception. There are 1871 disease codes in total. However, the top 250 disease codes are more than 80% of total medical records. However, our study predicts 1871 disease codes in total. In this study, 24 disease codes have more than 1000 medical records. In particular, the highest amount of medical records belong to 486 (pneumonia), and the least belong to 558.9 (other non-infectious gastroenteritis and colitis). There is a total of 58,540 records, which is 40% of the overall medical records. There are 234 disease codes with more than 100 medical records, a total of 118,096, which is 80% of the overall medical records. In the Tables 9 and 10 show the disease prediction results of 24 ICD-9 codes and 234 ICD-9 codes. In particular, Table 10 shows the top 1 and top 3 prediction results of 234 ICD-9 codes. A baseline is a method that uses heuristics and/or simple summary statistics to create predictions for a dataset. The baseline is showed in Equation (5).

In our study, the baseline always predicts the most frequent label in the training set. For example, the training data consists of 5 A's, 3 B's, and 2C's. In total, there are 10 items in the training data. The baseline will be 50%, since the baseline predictor always returns the most frequent label A.

$$\text{Baseline} = \frac{\text{No. of records in most common category}}{\text{Total records}}. \quad (5)$$

Table 9. Experiment results of Top-1.

Codes	Top-K	Recall	Baseline
24	Top 1	55.50%	16.59%
24+1	Top 1	66.21%	60.00%

Table 10. Top-K results of 234 ICD codes.

Codes	Top-K	Recall	Baseline
234	Top 1	43.43%	8.22%
234	Top 3	57.85%	15.65%
234 + 1	Top 1	43.24%	19.30%
234 + 1	Top 3	60.06%	34.65%

6. Discussion

Our prediction model behaves as an assistance tool that helps in getting to know the ICD-9 codes before approaching the hospital. The subject component is a primary part of the SOAP note. Other reports hold more detailed reports including the medical results, the long process of the medical treatment, and diagnosis methods. This study aims to provide a prediction model to help patients who want to go to the hospital. To achieve that, we have chosen the subjective components (feelings, opinions, and complaints) in this study. Moreover, the research study shows the importance of subjective components with other types of reports.

In the initial stage, we applied data cleaning and segmentation. Subsequently, we removed various stop words during the experiment. Table 11 shows the effects of removing stop words in the experiment. According to the findings, the recall is increased a little bit after utilizing English and Chinese stop words. We chose to retain more original data because there is no great difference in the increase rate, so we decided to keep Chinese stop words at this level.

We used the LSTM network as the first option for training the deep learning model in this study, and we experimented with several different changes in the natural language processing steps of the training data. First and foremost, prior to the word segmentation, we removed punctuation marks, special symbols, numbers, and list of stop words from the data in the pre-processing. The OCSW was prioritized in our study. This approach helped us to remove the unimportant words, which appear only once in the training data. Table 11 illustrates the results of both English stop words (EnSW) and Chinese stop words (CSW). The recall is increased slightly after using English and Chinese stop words. Next, we found out that there are many single characters and special symbols, such as \downarrow , β and other special symbols, which come to a total of 1312 characters in our data. We removed those data with the help of OCharSW. The recall was 0.5%, which improved as a result of applying this approach.

Table 11. The performance of stop words.

Total Words	Recall	Stop Word List
15,932	33.26%	OCSW + EnSW
15,551	33.29%	OCSW + EnSW + CSW
14,934	33.76%	OCSW + EnSW + OCharSW

Figure 8 shows 95% of the prediction rate in Chapter 15 (certain conditions originating in the perinatal period). Figure 7 depicts the prediction rate of the block (other conditions originating in the perinatal period (764–779)) which reached 95% and this block belongs to Chapter 5. The chapter only has two blocks (maternal causes of perinatal morbidity and mortality (760–763)) and (764–779), with 11,685 and 16 medical records, respectively. Due to the disparity in their number distribution, the prediction rate of chapter and block remained at 95% and further improvement is not achieved. Figure 8 confirms that chapter 5 (mental disorders) achieves 81% of the prediction rate. Similarly, Figure 7 depicts the block (other psychoses (295–299)) gained 78% of the prediction rate.

Chapter 5 covers “mental disorders”, and Chapter 15 describes “certain conditions originating in the perinatal period”. In our dataset, the patient self-reports of Chapters 5 and

15 consist of relatively short sentences, and are composed of specific medical terminologies, which leads to higher accuracy.

The idea behind the preference of recall is that the cost of failing to predict the disease of a patient is much more important than the cost of admitting a healthy person to involve in more tests. This study is intended to help patients who have a need to go to the hospital. This study prefers to predict the ICD code as much as possible in this approach.

Our study has major uniqueness with previous studies in the process of ICD-9 prediction. Most of the study focuses on handling a minimal number of ICD codes. Some of the studies are a overperformed recall of this study. However, this research predicts 1871 ICD codes using a subjective component. This study scores the 0.57 of recall in the ICD chapter level prediction. A previous study utilized more medically specific data such as discharge summary, nursing notes and diagnosis summary. Numerous medical tests and treatment processes are involved in those input data. This study takes a basic complaint to compete with previous studies.

Table 12 shows that the input words of the medical record support to help the correct predictions. The long hospitalization days, similarities in the daily medical records, same consulting doctors, and similar inner feelings of patients can be the reason of the input data similarity. The wrong prediction shows that the complete different words cause the failure of correct prediction, however, the prediction results are quite close. The wrongly predicted codes 774.1 and 770.1 belong to the same block (764-779) which indicates that the chapter and block are correctly predicted.

Table 12. Discussion of 770.8 ICD-9-CM code prediction results.

Correctly Predicted		
Input Data	Predicted Code	True Code
Intermitt tachypnea no desatur no bradycardia fair satur	770.8	770.8
Intermitt tachypnea upto min no desatur mild subcost retract	770.8	770.8
Intermitt tachypnea under room air no apnea no desatur no fever	770.8	770.8
Intermitt tachypnea under nasal cpap no desatur no bradycardia no apnea note under nasal cpap	770.8	770.8
No apnea bradycardia desatur fair activ and appetit	770.8	770.8
No tachypnea no desatur toler ml meal well fair activ	770.8	770.8
No bradycardia no desatur spo under fio occasion tachypnea no vomit under ml qh yellowish skin discolor	770.8	770.8
Incorrect Predicted		
Input Data	Predicted Code	True Code
No yellowish skin no apnea no cyanosi oral feed well no vomit	774.6	770.8
Smooth respiratori pattern with fair satur oral feed well no vomit or ground note	770.1	770.8

Note: Statements of input data are translated from Chinese.

There are two disease codes that are very close in meaning among the top 20 disease codes in the research results, these are code 486 (pneumonia) and code 485 (bronchial pneumonia). Figure 5 shows that ICD-9 code 486 scored 50% and 485 scored 53% in the confusion matrix of Level 4 (full code) prediction. Repeated identical words are more likely to cause errors in the prediction of these two codes, especially when predicting code 485. The most obvious difference between the two disease codes is that the model can correctly predict that code 485 usually has the word "rhinorrhea" in the medical record, indicating that the word is a common symptom or the vocabulary is commonly used when writing

medical records, and this argument is quite reasonable from the point of view of the disease represented by code 485.

Threats to Validity

Construct validity: This research aims to predict ICD codes using our prediction model. The subjective component is a part of SOAP, which is used in this study to predict ICD codes. This model behaves as an assistance tool that helps to know the ICD codes just before reaching the hospital. Our model predicts 1871 ICD codes in this study.

Internal validity: Selection threats are handled in this study. Data cleaning, segmentation, and NLP approaches are involved in the data pre-processing to predict the prediction model. The study builds a model to solve the ICD prediction.

External validity: The study can be used to predict an ICD using the simple self-report of patients. This model can help any patient with their feelings about their medical issue. The study is generalized for any other patient.

Conclusion validity: This study uses the GRU model as a benchmark for defining model parameters in our experiment process. The 10-fold cross-validation is processed in this study and the results are discussed in this study.

7. Conclusions

In this paper, we used LSTM and GRU to procure a model. In the stage of pre-processing, NLP techniques were implemented on the subjective component. The subjective component was used as a training data to predict the ICD-9 code. The study discussed the results of chapter, block, three-digit code, and full code prediction. The different Top-K (top-1, top-3, top-5, and top-10) prediction results were discussed in our study. Top-10 prediction achieved a recall of 67.37%. On the basis of our study, we proved that the subjective component has a notable importance in predicting ICD codes. The contemporary healthcare system needs the help of various medical applications. For our future studies, we will focus on how to produce an evolved prediction system with web support such as a medical chat-bot.

Author Contributions: conceptualization, J.-L.H. and C.-H.H.; methodology, J.-L.H. and C.-H.H.; software, Y.-K.L. and A.S.; validation, J.-L.H. and C.-H.H.; formal analysis, J.-L.H. and A.S.; writing—original draft preparation, J.-L.H., Y.-K.L., C.-H.H. and A.S.; writing—review and editing, J.-L.H., C.-H.H. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: Financial support for this study was provided in part by a grant from the Ministry of Science and Technology, Taiwan, under Contract No. MOST-108-2221-E-030-013-MY2. The funding agreement ensured the author's independence in designing the study, interpreting the data, writing, and publishing the report.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The training data presented in this study, the code of our proposed methods, and the trained model are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Gong, F.; Wang, M.; Wang, H.; Wang, S.; Liu, M. SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation. *Big Data Res.* **2021**, *23*, 100174. [\[CrossRef\]](#)
2. Moor, M.; Rieck, B.; Horn, M.; Jutzeler, C.R.; Borgwardt, K. Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review. *Front. Med.* **2021**, *8*, 348. [\[CrossRef\]](#)
3. Kantardzic, M. *Data Mining: Concepts, Models, Methods, and Algorithms*, 2nd ed.; Wiley-IEEE Press: Chichester, UK, 2011.
4. Brijain, M.; Patel, R.; Kushik, M.; Rana, K. A Survey on Decision Tree Algorithm For Classification. *Int. J. Eng. Dev. Res.* **2014**, *2*, 1–5.
5. Zheng, B.; Yoon, S.W.; Lam, S.S. Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Syst. Appl.* **2014**, *41*, 1476–1482. [\[CrossRef\]](#)

6. Abreu, P.H.; Santos, M.S.; Abreu, M.H.; Andrade, B.; Silva, D.C. Predicting breast cancer recurrence using machine learning techniques: A systematic review. *ACM Comput. Surv. (CSUR)* **2016**, *49*, 1–40. [[CrossRef](#)]
7. Yang, P.T.; Wu, W.S.; Wu, C.C.; Shih, Y.N.; Hsieh, C.H.; Hsu, J.L. Breast cancer recurrence prediction with ensemble methods and cost-sensitive learning. *Open Med.* **2021**, *16*, 754–768. [[CrossRef](#)]
8. Hoover, R. Benefits of using an electronic health record. *Nursing* **2016**, *46*, 21–22. [[CrossRef](#)] [[PubMed](#)]
9. Varela, L.O.; Doktorchik, C.; Wiebe, N.; Quan, H.; Eastwood, C. Exploring the differences in ICD and hospital morbidity data collection features across countries: An international survey. *BMC Health Serv. Res.* **2021**, *21*, 1–9.
10. Cartwright, D.J. ICD-9-CM to ICD-10-CM codes: What? why? how? *Adv. Wound Care* **2013**, *2*, 588–592. [[CrossRef](#)] [[PubMed](#)]
11. Li, M.; Fei, Z.; Zeng, M.; Wu, F.X.; Li, Y.; Pan, Y.; Wang, J. Automated ICD-9 Coding via A Deep Learning Approach. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2019**, *16*, 1193–1202. [[CrossRef](#)]
12. Alcaide, D.; Aerts, J. A visual analytic approach for the identification of ICU patient subpopulations using ICD diagnostic codes. *PeerJ Comput. Sci.* **2021**, *7*, e430. [[CrossRef](#)]
13. Hsu, J.L.; Hsu, T.J.; Hsieh, C.H.; Singaravelan, A. Applying Convolutional Neural Networks to Predict the ICD-9 Codes of Medical Records. *Sensors* **2020**, *20*, 7116. [[CrossRef](#)]
14. Hsu, M.C.; Wang, C.C.; Huang, L.Y.; Lin, C.Y.; Lin, F.J.; Toh, S. Effect of ICD-9-CM to ICD-10-CM coding system transition on identification of common conditions: An interrupted time series analysis. *Pharmacoepidemiol. Drug Saf.* **2021**. [[CrossRef](#)]
15. Salmon, P.; Rappaport, A.; Bainbridge, M.; Hayes, G.; Williams, J. Taking the problem oriented medical record forward. In *Proceedings of the AMIA Annual Fall Symposium*; American Medical Informatics Association: Atlanta, GA, USA, 1996; pp. 463–467.
16. Gangavarapu, T.; Jayasimha, A.; Krishnan, G.S.; S, S.K. Predicting ICD-9 code groups with fuzzy similarity based supervised multi-label classification of unstructured clinical nursing notes. *Knowl. Based Syst.* **2020**, *190*, 105321. [[CrossRef](#)]
17. Krishnan, G.S.; S, S.K. Evaluating the Quality of Word Representation Models for Unstructured Clinical Text Based ICU Mortality Prediction. In *Proceedings of the 20th International Conference on Distributed Computing and Networking, ICDCN'19, Bangalore, India, 4–7 January 2019*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 480–485. [[CrossRef](#)]
18. Huang, J.; Osorio, C.; Sy, L.W. An empirical evaluation of deep learning for ICD-9 code assignment using MIMIC-III clinical notes. *Comput. Methods Programs Biomed.* **2019**, *177*, 141–153. [[CrossRef](#)] [[PubMed](#)]
19. Wang, S.M.; Chang, Y.H.; Kuo, L.C.; Lai, F.; Chen, Y.N.V.; Yu, F.Y.; Chen, C.W.; Li, Z.W.; Chung, Y.F. Using Deep Learning for Automatic Icd-10 Classification from Free-Text Data. 2020. Available online: shorturl.at/vBOU7 (accessed on 23 August 2021).
20. Chen, P.F.; Wang, S.M.; Liao, W.C.; Kuo, L.C.; Chen, K.C.; Lin, Y.C.; Yang, C.Y.; Chiu, C.H.; Chang, S.C.; Lai, F. Automatic ICD-10 Coding and Training System: Deep Neural Network Based on Supervised Learning. *JMIR Med. Inform.* **2021**, *9*, e23230. [[CrossRef](#)]
21. Diao, X.; Huo, Y.; Zhao, S.; Yuan, J.; Cui, M.; Wang, Y.; Lian, X.; Zhao, W. Automated ICD coding for primary diagnosis via clinically interpretable machine learning. *Int. J. Med. Inform.* **2021**, *153*, 104543. [[CrossRef](#)]
22. Mauch, J.T.; Rios-Diaz, A.J.; Kozak, G.M.; Zhitomirsky, A.; Broach, R.B.; Fischer, J.P. How to Develop a Risk Prediction Smartphone App. *Surg. Innov.* **2020**, *28*, 438–448.
23. Zeng, M.; Li, M.; Fei, Z.; Yu, Y.; Pan, Y.; Wang, J. Automatic ICD-9 coding via deep transfer learning. *Neurocomputing* **2019**, *324*, 43–50.
24. Samonte, M.J.C.; Gerardo, B.D.; Fajardo, A.C.; Medina, R.P. ICD-9 Tagging of Clinical Notes Using Topical Word Embedding. In *Proceedings of the 2018 International Conference on Internet and E-Business, ICIEB'18, Singapore, 25–27 April 2018*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 118–123. [[CrossRef](#)]
25. Moons, E.; Khanna, A.; Akkasi, A.; Moens, M.F. A Comparison of Deep Learning Methods for ICD Coding of Clinical Records. *Appl. Sci.* **2020**, *10*, 5262. [[CrossRef](#)]
26. Hsu, C.C.; Chang, P.C.; Chang, A. Multi-Label Classification of ICD Coding Using Deep Learning. In *Proceedings of the 2020 International Symposium on Community-centric Systems (CcS), Tokyo, Japan, 23–26 September 2020*; pp. 1–6. [[CrossRef](#)]
27. Gupta, S.; Dieleman, F.; Long, P.; Doyle, O.; Leavitt, N. Using SNOMED to Automate Clinical Concept Mapping. In *Proceedings of the ACM Conference on Health, Inference, and Learning, CHIL'20, Toronto, ON, Canada, 2–4 April 2020*; Association for Computing Machinery: New York, NY, USA, 2020; pp. 131–138. [[CrossRef](#)]
28. Che, Z.; Kale, D.; Li, W.; Bahadori, M.T.; Liu, Y. Deep Computational Phenotyping. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'15, Sydney, NSW, Australia, 10–13 August 2015*; Association for Computing Machinery: New York, NY, USA, 2015; pp. 507–516. [[CrossRef](#)]
29. Marafino, B.J.; Davies, J.M.; Bardach, N.S.; Dean, M.L.; Dudley, R.A. N-gram support vector machines for scalable procedure and diagnosis classification, with applications to clinical free text data from the intensive care unit. *J. Am. Med. Inform. Assoc.* **2014**, *21*, 871–875.
30. Kavuluru, R.; Rios, A.; Lu, Y. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artif. Intell. Med.* **2015**, *65*, 155–166.
31. Shi, H.; Xie, P.; Hu, Z.; Zhang, M.; Xing, E.P. Towards Automated ICD Coding Using Deep Learning. *arXiv* **2017**, arXiv:1711.04075.
32. Xie, P.; Xing, E. A Neural Architecture for Automated ICD Coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018*; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 1066–1076. [[CrossRef](#)]

33. Gangavarapu, T.; S Krishnan, G.; Kamath S, S.; Jeganathan, J. FarSight: Long-Term Disease Prediction Using Unstructured Clinical Nursing Notes. *IEEE Trans. Emerg. Top. Comput.* **2020**, *9*, 1151–1169. [[CrossRef](#)]
34. Li, H.; Li, X.; Ramanathan, M.; Zhang, A. Prediction and Informative Risk Factor Selection of Bone Diseases. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12*, 79–91. [[CrossRef](#)]
35. Jin, B.; Che, C.; Liu, Z.; Zhang, S.; Yin, X.; Wei, X. Predicting the Risk of Heart Failure With EHR Sequential Data Modeling. *IEEE Access* **2018**, *6*, 9256–9261. [[CrossRef](#)]
36. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
37. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; Dean, J. Distributed Representations of Words and Phrases and Their Compositionality. In Proceedings of the 26th International Conference on Neural Information Processing Systems—Volume 2, NIPS’13, Lake Tahoe, NV, USA, 5–10 December 2013; Curran Associates Inc.: Red Hook, NY, USA, 2013; pp. 3111–3119.
38. Bengio, Y.; Ducharme, R.; Vincent, P.; Janvin, C. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* **2003**, *3*, 1137–1155.
39. Manning, C.D.; Raghavan, P.; Schütze, H. *Introduction to Information Retrieval*; Cambridge University Press: New York, NY, USA, 2008. [[CrossRef](#)]
40. Rizzo, S.G.; Montesi, D.; Fabbri, A.; Marchesini, G. ICD Code Retrieval: Novel Approach for Assisted Disease Classification. In Proceedings of the Data Integration in the Life Sciences, Luxembourg, 14–15 November 2017; Ashish, N., Ambite, J.L., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 147–161.
41. Choi, E.; Bahadori, M.T.; Schuetz, A.; Stewart, W.F.; Sun, J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In Proceedings of the 1st Machine Learning for Healthcare Conference, Children’s Hospital LA, Los Angeles, CA, USA, 19–20 August 2016; Volume 56, pp. 301–318.
42. Gu, J.; Liang, L.; Song, H.; Kong, Y.; Ma, R.; Hou, Y.; Zhao, J.; Liu, J.; He, N.; Zhang, Y. A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in Guangxi, China. *Sci. Rep.* **2019**, *9*, 1–10. [[CrossRef](#)]
43. Xie, J.; Wu, R.; Wang, H.; Chen, H.; Xu, X.; Kong, Y.; Zhang, W. Prediction of cardiovascular diseases using weight learning based on density information. *Neurocomputing* **2021**, *452*, 566–575. [[CrossRef](#)]