healthcare

MDPI

Article

DeepDRG: Performance of Artificial Intelligence Model for Real-Time Prediction of Diagnosis-Related Groups

Md. Mohaimenul Islam [1,2,3], Guo-Hung Li [2], Tahmina Nasrin Poly [1,3,4] and Yu-Chuan (Jack) Li [1,2,3,4,5,6,*]

[1] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 110, Taiwan; d610106004@tmu.edu.tw (M.M.I.); d610108004@tmu.edu.tw (T.N.P.)
[2] AESOP Technology, Songshan District, Taipei 105, Taiwan; jacky@aesoptek.com
[3] International Center for Health Information Technology (ICHIT), Taipei Medical University, Taipei 110, Taiwan
[4] Research Center of Big Data and Meta-Analysis, Wan Fang Hospital, Taipei Medical University, Taipei 116, Taiwan
[5] Department of Dermatology, Wan Fang Hospital, Taipei 116, Taiwan
[6] TMU Research Center of Cancer Translational Medicine, Taipei Medical University, Taipei 110, Taiwan
* Correspondence: jack@tmu.edu.tw

Abstract: Nowadays, the use of diagnosis-related groups (DRGs) has been increased to claim reimbursement for inpatient care. The overall benefits of using DRGs depend upon the accuracy of clinical coding to obtain reasonable reimbursement. However, the selection of appropriate codes is always challenging and requires professional expertise. The rate of incorrect DRGs is always high due to the heavy workload, poor quality of documentation, and lack of computer assistance. We therefore developed deep learning (DL) models to predict the primary diagnosis for appropriate reimbursement and improving hospital performance. A dataset consisting of 81,486 patients with 128,105 episodes was used for model training and testing. Patients' age, sex, drugs, diseases, laboratory tests, procedures, and operation history were used as inputs to our multiclass prediction model. Gated recurrent unit (GRU) and artificial neural network (ANN) models were developed to predict 200 primary diagnoses. The performance of the DL models was measured by the area under the receiver operating curve, precision, recall, and F1 score. Of the two DL models, the GRU method, had the best performance in predicting the primary diagnosis (AUC: 0.99, precision: 83.2%, and recall: 66.0%). However, the performance of ANN model for DRGs prediction achieved AUC of 0.99 with a precision of 0.82 and recall of 0.57. The findings of our study show that DL algorithms, especially GRU, can be used to develop DRGs prediction models for identifying primary diagnosis accurately. DeepDRGs would help to claim appropriate financial incentives, enable proper utilization of medical resources, and improve hospital performance.

Keywords: deep learning; artificial intelligence; diagnosis-related groups; hospital expenditure

Citation: Islam, M.M.; Li, G.-H.; Poly, T.N.; Li, Y.-C. DeepDRG: Performance of Artificial Intelligence Model for Real-Time Prediction of Diagnosis-Related Groups Coding. *Healthcare* **2021**, *9*, 1632. https://doi.org/10.3390/healthcare9121632

Academic Editors: Mahmoud Elkhodr, Omar Darwish, Belal Alsinglawi and Daniele Giansanti

Received: 21 October 2021
Accepted: 22 November 2021
Published: 25 November 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 (http://creativecommons.org/licenses/by/4.0/).

1. Introduction

Healthcare spending has consistently been increasing globally. Inpatient care is one of the most expensive hospital services, accounting for approximately 31% of the total expenditure [1]. With the limited resources and increased complexity, policymakers are facing immense challenges of reducing health care costs while improving financial protections, high-quality care, and lowering out-of-pocket (OOP) costs for people [2]. The Fee-for-service (FFS) is a basic payment system for both national and private hospitals in which all care providers are reimbursed for each service provided [3,4]. To rein in excessive healthcare costs and maintain sustainable procedures for inpatients, prospective payment policies are often implemented to foster risk-sharing between insurers and

providers [5,6]. Nowadays, the governments of several countries have already reformed their hospital payment policy by shifting from FFS to diagnosis-related groups (DRGs).

The concept of DRGs has now been widely adopted and become the principal means of reimbursement for inpatient services globally [7]. Previous studies have shown that the implementation of DRGs helped to shorten the length of hospital stays and lower OOP by efficiently allocating hospital resources [8,9]. DRGs, a prospective payment system, are used to effectively classify and combine diseases with similar characteristics into different diagnostics and treatment groups. Indeed, DRGs are set to achieve greater equality of financing based on the homogeneity of the clinical process and the similarity of resource consumption [10]. The diagnostics codes should be accurately matched with DRGs codes to claim actual reimbursement, but this is a time-consuming process and requires expert knowledge to manually retrieve information from patients' clinical records. Selection of appropriate DRGs codes depends on several factors, such as patients' comorbidities, complications, treatments, age, discharge status, and the principal diagnoses provided by physicians. Therefore, the quality of DRGs coding is a key factor that influences a hospital's ability to receive reasonable reimbursement and its overall profits.

DRGs coding errors can influence hospitals' income, hamper proper planning, and often lead to unfair distributions of resources. Ayub et al. [11] evaluated coding accuracy and its impact on hospital costs, reporting 9.6% miscoding with a total lost billing opportunity of $587,799. Cheng et al. [12] reviewed the causes and consequences of miscoding in a Melbourne tertiary hospital, finding that 16% of the 752 cases audited reflected a DRG change and caused a loss of hospital revenue of nearly AUD 575,300. Furthermore, the incorrect selection of the principal diagnosis accounted for an additional 13% of the DRG changes, which is due to the poor quality of documentation [12]. A previous study demonstrated that the overall rate of incorrect DRGs coding was up to 52%, which may be due to a lack of professional experience, work load, and lack of automation [13].

The widespread application of electronic health records (EHRs) has generated large amounts of patient data and created immense opportunities to predict the primary diagnosis using deep learning (DL). In this study, we developed and validated DL models to predict the primary diagnosis for appropriate reimbursement and improve the quality of care.

2. Literature Review

Taiwanese residents have been benefited from the nationwide health-care coverage through the compulsory National Health Insurance (NHI) scheme since 1995 [14]. The introduction of this system has ensured high-quality care, and NHI provides reimbursement for nearly all medical fees [15]. However, given the limited resources, global health care systems are facing immense challenges in responding to burgeoning healthcare expenditures [16,17]. Therefore, several strategies have been implemented to reduce unnecessary costs and minimize financial risks from insurers to providers [5,18]. In 2010, Taiwan introduced the diagnosis-related group (DRG) payment system, aiming to improve efficiency and minimize costs. This DRG payment system has an evident impact on current health care services, including the length of hospital stay and the intensity of inpatients care [19].

Artificial intelligence (AI) has shown great promise in improving patient care and making fruitful clinical decisions [20,21]. The availability of EHRs data has created an opportunity to calculate DRGs and associated costs at the time of admission using AI algorithms. The Taiwan National Health Insurance (NHI) Research Database is one of the largest nationwide population databases in the world, which has been used to produce high-quality research [22–26]. As a result, the NHI database can be used to predict DRGs using AI algorithms to accurate reflecting the costs incurred by hospital treatments and stays. Using two cohorts, a prior study applied a deep learning model to automatically predict DRGs and the corresponding costs [27]. Moreover, another study from Germany

examined the effectiveness of statistical machine learning in the early prediction of DRG and resource allocation at a 350-bed hospital [28].

2.1. Artificial Neural Networks

Artificial neural networks (ANNs) are also simply called neural networks. ANNs are a subset of machine learning and are at the heart of DL algorithms that are inspired by the biological neural network. The main concept of an ANN was described by McCulloch and Pitts [29] and was finally developed in 1958 [30]. An ANN consists of three layers: the input layer, which receives data, the output layer, which generates valuable information from provided data, and one or multiple hidden layers that are connected to the input and output layer (Figure 1).

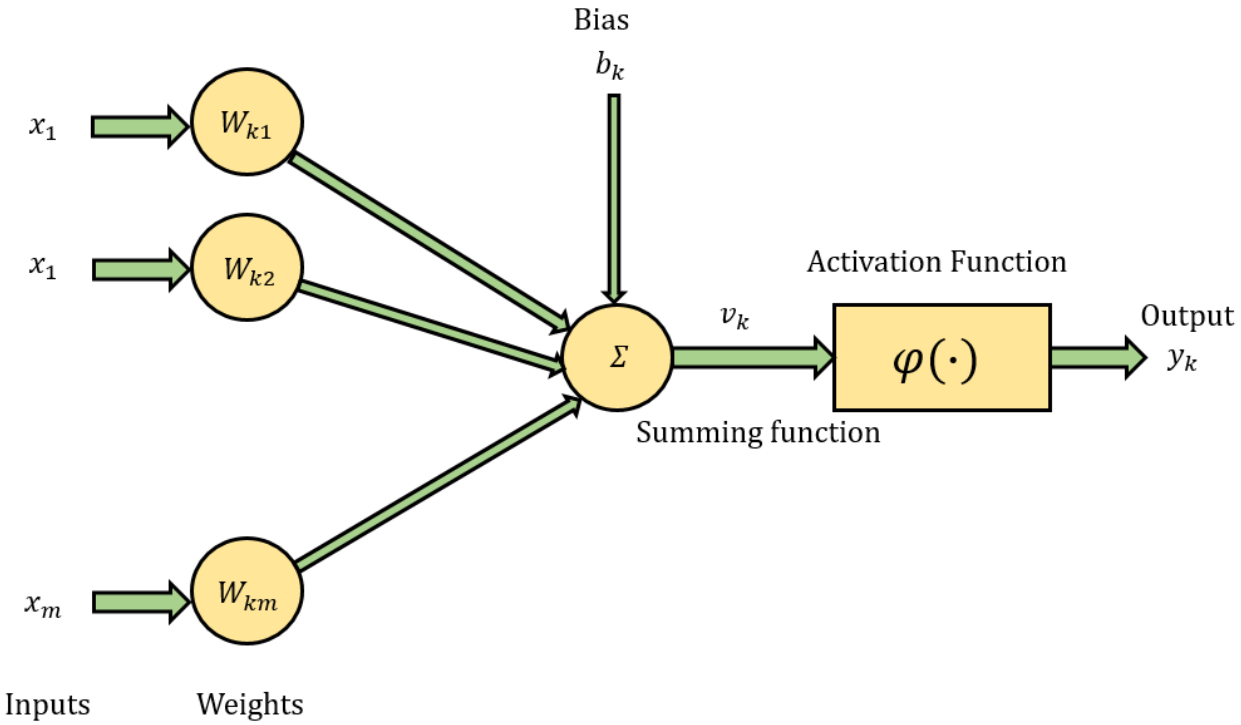

Figure 1. Simple structure of ANN.

For example, each input x_i in the input layer is multiplied by a connection weight w_{ij} between the neuron i in the input layer to the neuron j in the hidden layer. However, bias b_j is formerly summarized as net input S_j and passes it to the hidden layer by a nonlinear activation function, sigma to generate an output, y_j.

$$S_j = \sum x_i w_{ij} + b_j$$

$$f(S) = \frac{(1 - e^{-2S})}{(1 + e^{-2S})}$$

$$y_j = f(S_j)$$

However, y_i signal from the hidden layer to all k neuron in the final output layer F_k and calculate the input to the k neuron of the output layer is F'_k.

$$F'_k = \sum_{j=1}^{h} y_j w'_{jk} + b'_k$$

where w'_{jk} is the weight of the connected between the j neuron in the hidden layer to the k neuron of the output layer; b'_k is the bias.

Finally, it calculates the output layer signals by using an activation function, sigmoid.

$$F_k = f(F'_k)$$

The error between the target outcome and the observed data is measured as follows:

$$Error_k = \frac{1}{2}\sum_{k=1}^{n}(target_k - observed_k)$$

This process, utilized in all pairs in the training dataset and the training cycles, is known as epoch. The number of epochs is selected by users and repeated to reach minimum error.

The error gradient for output layer is calculated as follows:

$$Gradient\ error_k = (target_k - observed_k)f'(F'_k)$$

Rosenblatt [30] first proposed the idea of perceptron, which was used in a gradient descent-based learning algorithm. The perceptron is centered on single neuron and can be considered the primary basis of feed-forward ANNs. Perceptron is more generally a computational model, and it takes an input, aggregates it and returns 1 only if the weighted sum is more than some thresholds, otherwise it gives 0. The equation is given below:

$$f(net) = \begin{cases} 1\ if \sum_{i=0}^{n} w_i x_i > 0 \\ 0\ otherwise \end{cases}$$

However, there were several limitations of using perceptron, which were raised by Minsky and Papert [31], who mentioned that it cannot be used to implement the sample data which are not linearly separable. In the modern era, using backpropagation has immense advantages over traditional gradient descent methods. It provides a way to train networks with any number of hidden units arranged in any number of layers. In fact, the networks do not need to be organized in layers. It provides multi-layer feed-forward ANNs with a highly competitive supervised algorithm. The backpropagation helps to reduce the predefined loss function through updating the weight and bias values [32].

2.2. Gated Recurrent Unit

A gated recurrent unit (GRU) uses a gating mechanism in the recurrent neural networks and was first introduced in 2014 by Cho [33]. The GRU is quite similar to long short-term memory (LSTM) with a forget gate, but GRU has only two gates (reset and update,) and LSTM has three gates (input, output, and forget). GRU is less complex, faster, requires less memory and less time to train compared to LSTM because it has less parameters than LSTM. The main advantage of using GRU is that it solves the vanishing gradient problem, which comes with a standard recurrent neural network (RNN). Figure 2 presents the inputs for both the reset and update gates in a GRU where the input of the current time step is x_t, the hidden state of the previous time step is h_{t-1}, and output is calculated by activation function.

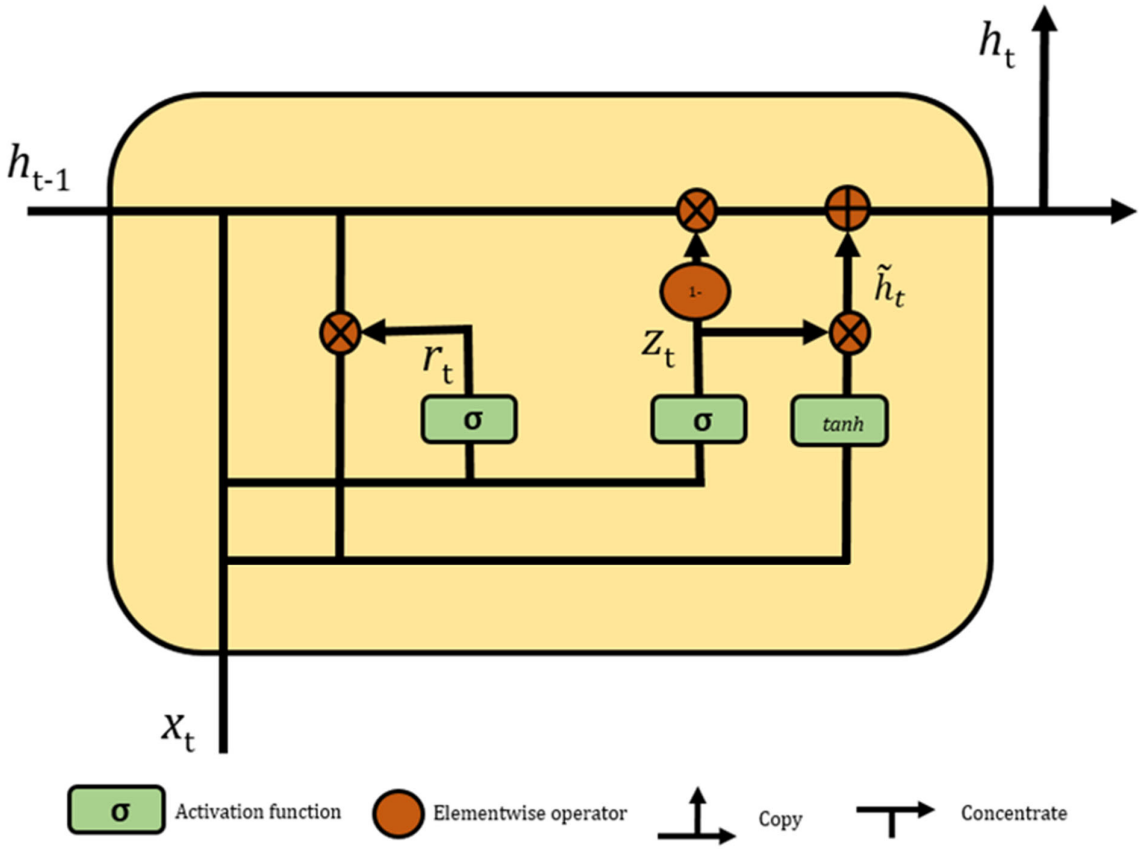

Figure 2. Architecture of gated recurrent unit.

Assume that, in a given time step t, the number of provided inputs of a small batch $x_t \in \mathbb{R}^{n\times h}$ (h is the number of hidden units). Then, the reset gate $r_t \in \mathbb{R}^{n\times h}$ and update gate $z_t \in \mathbb{R}^{n\times h}$ are calculated as follows:

$$r_t = \sigma_{GRU}(x_t w_{xr} + h_{t-1} w_{hr} + b_r)$$

$$z_t = \sigma_{GRU}(x_t w_{xz} + h_{t-1} w_{hz} + b_z)$$

Here, $w_{xr}, w_{xz} \in \mathbb{R}^{v\times h}$ and $w_{hr}, w_{hz} \in \mathbb{R}^{h\times h}$ are weight parameters, $b_r, b_z \in \mathbb{R}^{1\times h}$ are biases, and σ_{GRU} is an activation function, which converts values in both gates into 0–1.

Later, the candidate hidden state $\tilde{h}_t \in \mathbb{R}^{n\times h}$ is generated in a series of operations between the output of the reset gate and the hidden state h_{t-1} of the previous time step. It is calculated as follows:

$$\tilde{h}_t = tanh_{GRU}\,(x_t w_{xh} + (r_t \otimes h_{t-1}) w_{hh} + b_h)$$

where $w_{xh} \in \mathbb{R}^{v\times h}$ and $w_{hh} \in \mathbb{R}^{h\times h}$ are weight parameters, $b_h \in \mathbb{R}^{1\times h}$ is the bias, $\otimes$ is the elementwise product operator, and $tanh_{GRU}$ is a nonlinear activation function, which converts value between −1 and 1. In this stage, the influence of previous states can be minimized with the elementwise multiplication of r_t and h_{t-1}. The unique part of this equation is how the element value of the reset gate controls how much influence of previous hidden state h_{t-1} can have on the candidate state. If the value of reset gate r_t is equal to 1, then all the information from the previous hidden state h_{t-1} is considered. Likewise, if the value of reset gate r_t is 0, then the information from the previous hidden state is completely ignored.

After obtaining the result of the candidate state, it is used to generate the current hidden state. This is where the update gate comes on board. However, the equation here is slightly different from LSTM (input and forget gate are complementary and have certain redundancy). GRU directly uses a single update gate to control both historical information, which is the hidden state h_{t-1} of the previous time moment, and the candidate hidden state $\tilde{h}$ of the current time. The final updated equation for the GRU is as follows:

$$h_t = z_t \otimes h_{t-1} + (1 - z_t) \otimes \tilde{h}_t$$

When the update gate z_t is close to 1, it will retain the old state. In this case, the information from x_t is completely ignored. When z_t is close to 0, then the new latent state h_t approaches the candidate latent state $\tilde{h}_t$.

3. Materials and Methods

3.1. Study Approval and Propose Methodology

This study was conducted in the multiple center according to the tenets of the Declaration of Helsinki. This study was approved by the Taipei Medical University Institutional Review Board, which waived informed patient consent because all patient records and information were anonymized and deidentified before the analysis. Figure 3 shows the data analysis framework for DRGs prediction in this study.

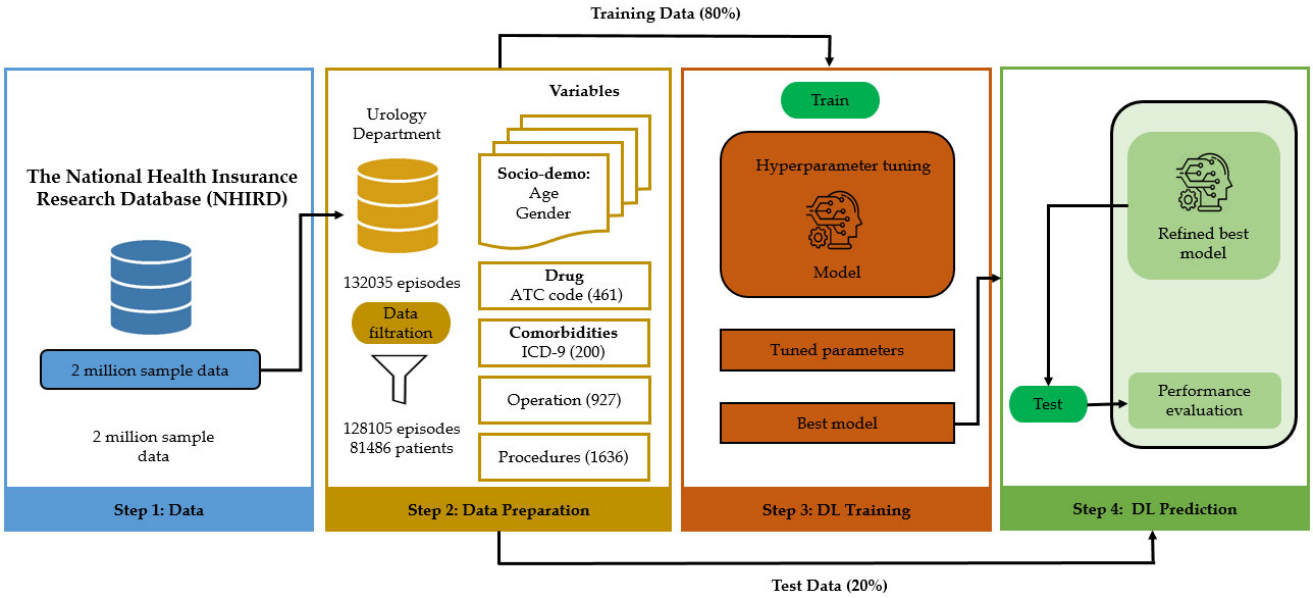

Figure 3. The methodological framework used in this study.

3.2. Data Source

This current study retrieved data from the Taiwan National Health Insurance Research Database (NHIRD) between 1999 and 2013. NHIRD offers the most comprehensive clinical information, such as demographics, diagnosis records, medication records, surgical records, and laboratory information from 23 million people in the Taiwanese population [34,35]. International guidelines were followed to record and collect data, e.g., diseases and medication prescriptions were coded and retrieved by using the International Classification of Diseases, Ninth Revision (ICD-9), and Anatomical Therapeutic Chemical (ATC). The quality and completeness of the NHIRD database is excellent, and it is used to conduct high-quality clinical research. In our study, we collected 2 million random samples from NHIRD. Afterwards, we carried out a

retrospective cohort study of individuals who visited the urinary department's inpatient care between 1999 and 2013. Patients were eligible for inclusion in this study if they were labelled as discharged and had a primary diagnosis for admission, resulting in 132,035 episodes. After filtration (e.g., missing data and infrequent comorbidities), the final study sample included 128,105 episodes from 81,486 patients.

3.3. Data Descriptions

In the era of big data, AI models have potential to make advanced clinical decision support and assist clinicians to deliver optimal care. Clinical decision support systems (CDSS) have the ability to analyze large volumes of data and recommend appropriate primary diagnosis for improving efficiency and sustainable care. However, to make a feasible implementation of a DL algorithm to predict DRGs, we included patient and clinical factors readily available in electronic health records (EHRs). More specifically, we collected patients' sociodemographic, admission status, admission history, admission diagnosis, discharge diagnosis, medications, comorbidities, operations, and procedures history to support real-time, proactive decision making related to selecting appropriate primary diagnosis for effective cost management.

Potential predictor selection was guided by previously published works and data availability. Predictor variables include the following: (i) patient demographics (age, gender), (ii) procedures (1636 types), (iii) drugs (461 types), (iv) operation (927 codes), and (v) the 200 most common comorbidities. Admission and discharge diagnoses were determined using the primary International Classification of Diseases, 9th edition (ICD-9). Moreover, drugs were determined using Anatomical Therapeutic Chemical (ATC), and operation and procedure codes were determined using standard protocol.

3.4. Data Preprocessing

Data preprocessing comprised the following steps: (1) data cleaning, (2) variable selection, and (3) one hot coding/embedding formation. NHIRD collects a vast number of variables; however, not all the variables were important to this study. In this process, we deleted irrelevant variables and kept only demographics, admission data, visit date, date of birth, department identification, medications, diseases, operations, laboratory, and procedures information. Afterwards, patients' age was calculated using their birth date and admission date information. There was no age limit included in our work. For medication information, we used the five-digit Anatomical Therapeutic Chemical Classification System., e.g., five-digit ATC code: B01AC, platelet aggregation inhibitors excluding heparin, including drugs such as B01AC01 (Ditazole), B01AC02 (Cloricromen), and B01AC04 (Clopidogrel). Additionally, seven characters (e.g., A10BA02) were considered for the chemical substances, even though five characters were used to describe the chemical substances, and all the drugs included in this group were prescribed for almost the same purpose. There were 1317 types of primary diagnosis during the study period, but all were infrequent. We therefore calculated the frequency and percentage of primary diagnoses and considered 200 primary diagnoses as our primary outcome. Those 200 diagnoses covered approximately 97% of total diagnoses. The reason for considering the top 200 primary diagnoses as targeted outcomes was because the DL algorithm needs sufficient data to train the model; otherwise, it may perform poorly.

3.5. Model Development

We split the data set into a training set (80%) and a testing set (20%). The GRU model was developed to train all the variables, and the model was assessed using the validation set to predict the primary diagnosis. GRU is a high-performing recurrent neural network. It is similar to the LSTM and RNN algorithms, but GRU consists of only two gates—a reset and an update gate [36,37]. The input of GRU moves through the layers, calculating the probability of each output. For the activation function, sigmoid was used in the hidden

layers, and Softmax was used in the output layer. The activation function is an integral part of a neural network that is often known for non-linearity, i.e., describing the input and output relations in a non-linear way. The architecture of the current model is shown in Figure 4. Six types of inputs (sex, age, drug, second diagnosis, procedure, and surgery) enter the embedding layer first. Afterwards, the embedding of sex and age enter the bilinear layer and PReLU. The output of the last step is multiplied with other embeddings of inputs. However, the other four are multiplied with embeddings, then enter the linear layers and GRU layers and calculate four statistic values for each output from step 4 and concatenate them. Finally, GRU gives the output as the probability using the activation function and residual layer.

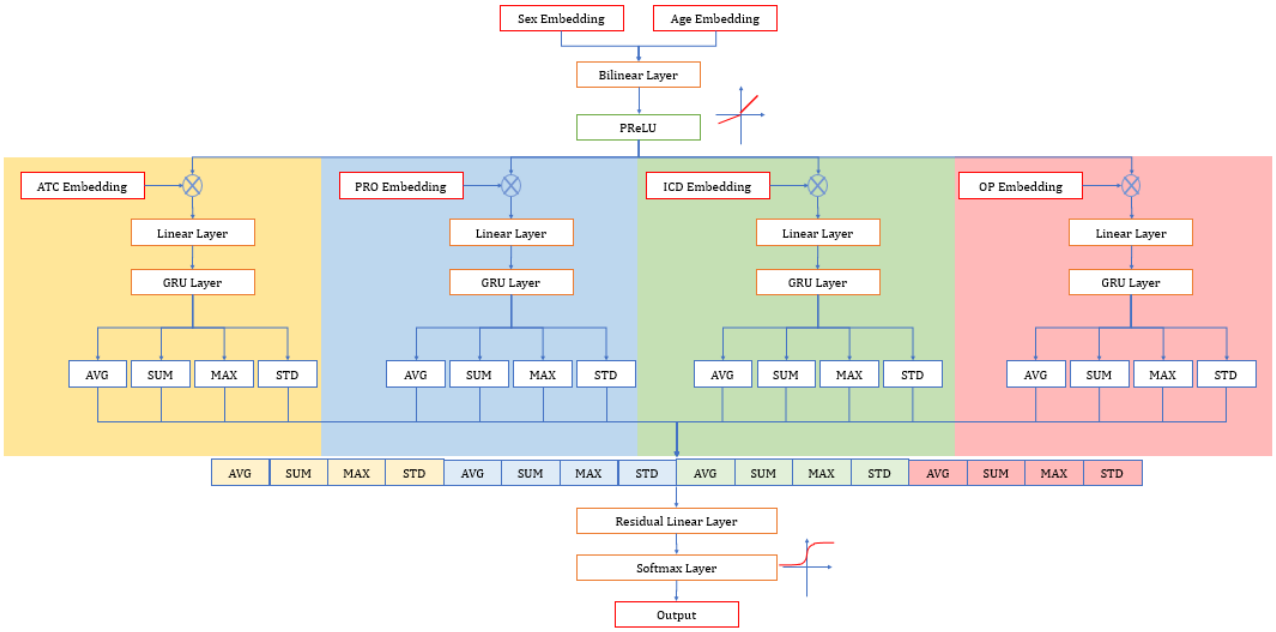

Figure 4. Architecture of GRU model.

In our study, we used 25 epochs; however, the model performed well while considering only 15 epochs (Figure 5).

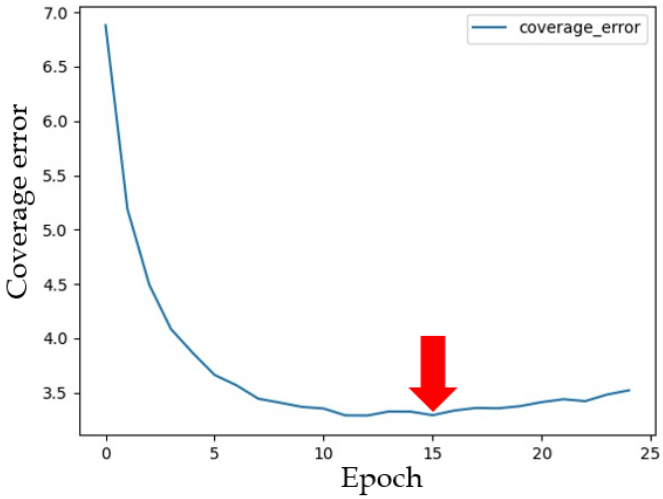

Figure 5. Testing loss of the GRU model.

3.6. Evaluation Matrices

We evaluated the performance of the DL models on the internal validation set for primary diagnosis recommendations using the following metrics.

Accuracy: It averages the entire set of data as an aggregate result, and calculates 1 metric rather than k metrics.

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

where TP = True Positive; TN = True Negative; FP = False Positive, and FN = False Negative.

Micro-F1: It measures the F1-score of the aggregated contributions of all classes. The equation is given below:

$$\text{micro} - \text{F1} = 2\frac{\text{Micro} - \text{Precision} \times \text{Micro} - \text{Recall}}{\text{Micro} - \text{Precision} + \text{Micro} - \text{Recall}}$$

Micro Precision and Recall: Average precision summarize the fraction of relevant labels ranked higher than the other relevant label. The equation is presented below:

$$\text{Precision} = \frac{\sum_c \text{TP}_c}{\sum_c \text{TP}_c + \sum_c \text{FP}_c}$$

$$\text{Recall} = \frac{\sum_c \text{TP}_c}{\sum_c \text{TP}_c + \sum_c \text{FN}_c}$$

where c is the class label.

4. Results

4.1. Patient Characteristics

A total of 128,105 patients admitted to the urinary department were included in this study. There were more male patients than female patients (74.65% vs. 25.35). The number of input and output was 3224 and 200, respectively. Table 1 shows the basic characteristics of patients.

Table 1. Basic characteristics of patients included.

Variable	Number/Percentage
Total number of episodes	128,105
Total number of patients	81,486
Age range	
Age group	
0~20	4.51%
20~40	54.76%
40~60	42.79%
>60	0.02%
Gender	
Male	74.65%
Female	25.35%
Operation	
Yes	87.78%
No	12.22%
Additional diagnosis	
Yes	70.98%
No	29.02%
Procedure	

Yes	98.82%
No	1.18%
Drug	
Yes	99.58%
No	0.42%
Number of drugs input	461
Number of diseases input	200
Number of procedures input	1636
Number of operations input	927
Number of output	200

4.2. Performance of Deep Learning Model

Table 2 shows the measurement of diagnostic performance (precision, recall, F1-score, accuracy, and AUROC) used to predict the primary diagnosis. The GRU model predicted the primary diagnosis with 83% precision, 66% recall, and 73% F1-score. The ANN model predicted the primary diagnosis with 82% precision, 57% recall, and 67% F1-score. However, the GRU performed better than ANN in predicting the primary diagnosis.

Table 2. Performance of deep learning models.

Model	Precision	Recall	F1-Score	Accuracy	AUROC	Ranking Loss
GRU	0.83	0.66	0.73	0.72	0.99	0.01
ANN	0.82	0.57	0.67	0.68	0.99	0.01

#Note: micro-AUROC.

4.3. Sensitivity Analysis

In our study, we also tested our model using different numbers of variables and achieved high performance when using all the variables (basic information, drugs, procedures, operations, and additional disease information). However, the model performed worse when only basic information was used. Table 3 shows the performance of the GRU model using different numbers of variables.

Table 3. Sensitivity analysis.

Basic Info	Drug	Procedure	Operation	Additional ICD	Precision	Recall	F1-Score	Accuracy	Micro-AUC	Label Ranking Loss
V	V	V	V	V	0.83	0.65	0.73	0.726	0.99	0.01
V	V	V	V		0.76	0.60	0.67	0.671	0.99	0.01
V	V	V		V	0.70	0.31	0.43	0.481	0.97	0.03
V	V		V	V	0.55	0.42	0.47	0.465	0.92	0.06
V		V	V	V	0.81	0.56	0.66	0.632	0.98	0.03
V	V				0.08	0.02	0.04	0.059	0.79	0.16
V		V			0.26	0.04	0.07	0.211	0.92	0.08
V			V		0.52	0.33	0.41	0.373	0.88	0.09
V				V	0.01	0.005	0.006	0.026	0.75	0.19
V					0.001	0	0.001	0.006	0.73	0.21

4.4. Evaluation

After developing and internally validating our model, we evaluated its effectiveness using several unknown cases. Table 4 shows five examples of how our model gave suggestions based on the patients' inputs. Our model gave the top five recommendations for primary diagnosis, from which the doctor could choose. If the doctor had wanted to see more recommendations, our system would have been able to provide more based on user inputs. However, the top diagnosis was supported by strong evidence. For example,

the original primary diagnosis for patient #1 was calculus of ureter, and our model predicted the same primary diagnosis. However, our model will provide additional suggestions for selecting primary diagnosis, and doctor can choose another as either a primary or secondary diagnosis. For patient#4, the original diagnosis was "Malignant bladder neoplasm, other specified sites", and our model suggested "Malignant bladder neoplasm, part unspecified". However, the original diagnosis was in our suggested lists, and the physician could have chosen any one of them. The main advantage of using a DL-based system is the availability of a list of appropriate suggestions without manual examination of patients' documentations.

Table 4. Evaluation of the performance of GRU for predicting primary diagnosis.

Example	Age	Sex	Original Primary Diagnosis	Predicted Primary Diagnosis	Top 5 Primary Diagnoses
Patient #1	20–40	Male	Calculus of ureter	Calculus of ureter	1. Calculus of ureter. 2. Calculus of kidney. 3. Urinary tract infection, site not specified. 4. Calculus in urethra. 5. Acute pyelonephritis without lesion of renal medullary necrosis.
Patient #2	20–40	Female	Calculus of kidney	Calculus of kidney	1. Calculus of kidney. 2. Acute pyelonephritis without lesion of renal medullary necrosis. 3. Urinary tract infection, site not specified. 4. Pyelonephritis, unspecified. 5. Renal colic.
Patient #3	20–40	Male	Malignant bladder neoplasm, part unspecified.	Malignant bladder neoplasm, part unspecified.	1. Malignant bladder neoplasm, part unspecified. 2. Malignant bladder neoplasm, lateral wall. 3. Malignant bladder neoplasm, other specified sites. 4. Neoplasms of unspecified nature, bladder. 5. Benign neoplasm of bladder.
Patient #4	20–40	Male	Malignant bladder neoplasm, other specified sites.	Malignant bladder neoplasm, part unspecified.	1. Malignant bladder neoplasm, part unspecified. 2. Neoplasms of unspecified nature, bladder. 3. Malignant bladder neoplasm, lateral wall. 4. Malignant bladder neoplasm, other specified sites. 5. Hematuria.
Patient #5	20–40	Male	Acute pyelonephritis without lesion of renal medullary necrosis.	Urinary tract infection, site not specified.	1. Urinary tract infection, site not specified. 2. Acute pyelonephritis without lesion of renal medullary necrosis. 3. Acute cystitis. 4. Hematuria. 5. Orchitis and epididymitis, other, without mention of abscess.

5. Discussion

To the best of our knowledge, this is the first study to examine the performance of a DL model used for the prediction of accurate primary diagnosis. This study shows the rigorous training and testing of a novel deep learning model that has been shown to achieve a high accuracy for the prediction of DRGs. The rapid rise of AI in healthcare offers great opportunities to overcome DRGs' coding limitations and to claim appropriate reimbursement for inpatient care. The satisfactory performance of DL models using a huge amount of EHRs data can now be provided as an alternative option to traditional manual DRGs coding.

The effectiveness of using DRGs in inpatient care is widely assessed because it is considered as the standard payment management system [38]. The main purpose of using DRGs in the inpatient care setting are: (a) to improve the transparency of the service provided by hospitals, (b) to utilize hospital resources efficiently, and (c) to obtain appropriate reimbursement. Hospital expenditure has increased, and the quality of care has diminished due to improper use of the current DRGs. A study from Saudi Arabia reported a coding error rate of approximately 30% for DRGs [39], while a 51% overall coding error rate was reported in a study from the UK [40]. However, incorrect selection of the primary diagnosis accounted for more than 13% of the DRG changes, and missing additional diagnosis codes accounted for 29% [12]. Moreover, Zafirah et al. reported [41] that the error rate of primary diagnoses was approximately 50%. In addition, the coding error rate of secondary diagnoses was also higher in Malaysia (81.3%), Saudi Arabia (35.6%), and Thailand (28.0%) [39,42,43].

If the DRGs' reimbursement is not correctly defined, then hospitals will lose income. A previous study reported that a potential loss of hospital income due to coding errors was equivalent to 39.1% of the total hospital income [42]. Maryati et al. [44] and Nouraei et al. [45] also mentioned that selection of inappropriate codes leads to a loss of income rather than making profits. Recoding or reassessing DRGs coding helps to generate potential income; however, it is time consuming and requires additional manpower to evaluate the coding. Our DL-based automated coding system can assist coders or physicians in selecting the best primary diagnosis for financial return. Moreover, physicians can select a secondary diagnosis from the top ten suggestions provided by our system.

Our study has several strengths. First, this is the first study to evaluate the DL performance for the prediction of DRGs. Second, the performance of our model was clinically satisfactory, which would help to reduce physicians' workload, increase accurate coding for claiming reimbursement, increase hospital income, and improve hospital performance by using proper allocation of resources. There are several limitations that also need to be mentioned. First, this study only focuses on the urinary department. However, this is because the rate of coding error is high in the urinary department and the selection of appropriate DRGs coding is challenging. Second, while our model was internally validated, the performance of the model could vary if using other countries' datasets. Third, our study did not evaluate how much money it would save or earn if the hospital chose to implement this model. However, the higher accuracy refers to the ability of our model to reduce the coding error of DRGs, which ultimately helps to save money, increase income, and improve the hospital performance for inpatient care.

6. Conclusions

This study revealed that a DL model, especially GRU, has the ability to predict DRGs' primary diagnosis with high accuracy. Using this automated DRGs coding system would help to reduce incorrect coding, which can ultimately increase hospital income, ensure the fair allocation of medical resources, and improve hospital performance. Further studies are needed to evaluate the performance of the current model using data from other countries and to assess the financial benefits of this model.

Author Contributions: Conceptualization, M.M.I. and Y.-C.L.; methodology, G.-H.L.; software, M.M.I., T.N.P., G.-H.L.; validation, Y.-C.L.; formal analysis, G.-H.L.; resources, M.M.I.; data curation, G.-H.L.; writing—original draft preparation, M.M.I.; writing—review and editing, Y.-C.L.; visualization, Y.-C.L.; supervision, Y.-C.L.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgements: We would like to thank our colleague, who is a native English speaker, for editing this manuscript.

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

References

1. Muka, T.; Imo, D.; Jaspers, L.; Colpani, V.; Chaker, L.; van der Lee, S.; Mendis, S.; Chowdhury, R.; Bramer, W.; Falla, A.; et al. The global impact of non-communicable diseases on healthcare spending and national income: A systematic review. *Eur. J. Epidemiol.* **2015**, *30*, 251–277.
2. Bayati, M.; Mehrolhassani, M.H.; Yazdi-Feyzabadi, V. A paradoxical situation in regressivity or progressivity of out of pocket payment for health care: Which one is a matter of the health policy maker's decision to intervention? *Cost Eff. Resour. Alloc.* **2019**, *17*, 1–4.
3. Jung, Y.W.; Pak, H.; Lee, I.; Kim, E.H. The Effect of Diagnosis-Related Group Payment System on Quality of Care in the Field of Obstetrics and Gynecology among Korean Tertiary Hospitals. *Yonsei Med. J.* **2018**, *59*, 539–545.
4. Chien, T.-W.; Lee, Y.-L.; Wang, H.-Y. Detecting hospital behaviors of up-coding on DRGs using Rasch model of continuous variables and online cloud computing in Taiwan. *BMC Health Serv. Res.* **2019**, *19*, 630.
5. Schwartz, A.L.; Chernew, M.E.; Landon, B.E.; McWilliams, J.M. Changes in Low-Value Services in Year 1 of the Medicare Pioneer Accountable Care Organization Program. *JAMA Intern. Med.* **2015**, *175*, 1815–1825.
6. Mihailovic, N.; Kocic, S.; Jakovljevic, M. Review of diagnosis-related group-based financing of hospital care. *Health Serv. Res. Manag. Epidemiol.* **2016**, *3*, 2333392816647892.
7. Mathauer, I.; Wittenbecher, F. Hospital payment systems based on diagnosis-related groups: Experiences in low- and middle-income countries. *Bull. World Health Organ.* **2013**, *91*, 746–756A.
8. Busse, R.; Geissler, A.; Aaviksoo, A.; Cots, F.; Häkkinen, U.; Kobel, C.; Mateus, C.; Or, Z.; O'Reilly, J.; Serdén, L.; et al. Diagnosis related groups in Europe: Moving towards transparency, efficiency, and quality in hospitals?. *BMJ* **2013**, *346*, f3197.
9. Kim, S.J.; Han, K.T.; Kim, W.; Kim, S.J.; Park, E.C. Early Impact on Outpatients of Mandatory Adoption of the Diagnosis-Related Group-Based Reimbursement System in Korea on Use of Outpatient Care: Differences in Medical Utilization and Presurgery Examination. *Health Serv. Res.* **2018**, *53*, 2064–2083.
10. Wu, S.-W.; Pan, Q.; Chen, T. Research on diagnosis-related group grouping of inpatient medical expenditure in colorectal cancer patients based on a decision tree model. *World J. Clin. Cases* **2020**, *8*, 2484–2493.
11. Ayub, S.; Scali, S.T.; Richter, J.; Huber, T.S.; Beck, A.W.; Fatima, J.; Berceli, S.A.; Upchurch, G.R.; Arnaoutakis, D.; Back, M.R.; et al. Financial implications of coding inaccuracies in patients undergoing elective endovascular abdominal aortic aneurysm repair. *J. Vasc. Surg.* **2018**, *69*, 210–218.
12. Cheng, P.; Gilchrist, A.; Robinson, K.M.; Paul, L. The Risk and Consequences of Clinical Miscoding Due to Inadequate Medical Documentation: A Case Study of the Impact on Health Services Funding. *Heal. Inf. Manag. J.* **2009**, *38*, 35–46.
13. Lee, M.-T.; Su, Z.-Y.; Hou, Y.-H.; Liao, H.-C.; Lian, J.-D. A decision support system for diagnosis related groups coding. *Expert Syst. Appl.* **2011**, *38*, 3626–3631.
14. Yang, H.-C.; Nguyen, P.A.A.; Islam, M.; Huang, C.-W.; Poly, T.N.; Iqbal, U.; Li, Y.-C.J. Gout drugs use and risk of cancer: A case-control study. *Jt. Bone Spine* **2018**, *85*, 747–753.
15. Hu, W.-Y.; Yeh, C.-F.; Shiao, A.-S.; Tu, T.-Y. Effects of diagnosis-related group payment on health-care provider behaviors: A consecutive three-period study. *J. Chin. Med. Assoc.* **2015**, *78*, 678–685.
16. Badgery-Parker, T.; Pearson, S.-A.; Chalmers, K.; Brett, J.; Scott, I.A.; Dunn, S.; Onley, N.; Elshaug, A.G. Low-value care in Australian public hospitals: Prevalence and trends over time. *BMJ Qual. Saf.* **2018**, *28*, 205–214.
17. Colla, C.H.; Morden, N.E.; Sequist, T.D.; Mainor, A.J.; Li, Z.; Rosenthal, M.B. Payer Type and Low-Value Care: Comparing Choosing Wisely Services across Commercial and Medicare Populations. *Health Serv. Res.* **2017**, *53*, 730–746.
18. Howard, D.H.; Gross, C.P. Producing Evidence to Reduce Low-Value Care. *JAMA Intern. Med.* **2015**, *175*, 1893–1894.
19. Kutz, A.; Gut, L.; Ebrahimi, F.; Wagner, U.; Schuetz, P.; Mueller, B. Association of the Swiss Diagnosis-Related Group Reimbursement System With Length of Stay, Mortality, and Readmission Rates in Hospitalized Adult Patients. *JAMA Netw. Open* **2019**, *2*, e188332.
20. Ijaz, M.F.; Attique, M.; Son, Y. Data-Driven Cervical Cancer Prediction Model with Outlier Detection and Over-Sampling Methods. *Sensors* **2020**, *20*, 2809.
21. Mandal, M.; Singh, P.K.; Ijaz, M.F.; Shafi, J.; Sarkar, R. A Tri-Stage Wrapper-Filter Feature Selection Framework for Disease Classification. *Sensors* **2021**, *21*, 5571.
22. Hsing, A.W.; Ioannidis, J.P. Nationwide population science: Lessons from the Taiwan national health insurance research database. *JAMA Intern. Med.* **2015**, *175*, 1527–1529.
23. Cheng, C.-L.; Kao, Y.-H.Y.; Lin, S.-J.; Lee, C.-H.; Lai, M.L. Validation of the national health insurance research database with ischemic stroke cases in Taiwan. *Pharmacoepidemiol. Drug Saf.* **2011**, *20*, 236–242.
24. Wang, C.-H.; Nguyen, P.A.; Li, Y.C.(.; Islam, M.; Poly, T.N.; Tran, Q.-V.; Huang, C.-W.; Yang, H.-C. Improved diagnosis-medication association mining to reduce pseudo-associations. *Comput. Methods Programs Biomed.* **2021**, *207*, 106181.

25. Yang, H.-C.; Islam, M.; Nguyen, P.A.A.; Wang, C.-H.; Poly, T.N.; Huang, C.-W.; Li, Y.-C.J. Development of a Web-Based System for Exploring Cancer Risk With Long-term Use of Drugs: Logistic Regression Approach. *JMIR Public Heal. Surveill.* **2021**, *7*, e21401.
26. Liang, C.-W.; Yang, H.-C.; Islam, M.; Nguyen, P.A.A.; Feng, Y.-T.; Hou, Z.Y.; Huang, C.-W.; Poly, T.N.; Li, Y.-C.J. Predicting Hepatocellular Carcinoma With Minimal Features From Electronic Health Records: Development of a Deep Learning Model. *JMIR Cancer* **2021**, *7*, e19812.
27. Liu, J.; Capurro, D.; Nguyen, A.; Verspoor, K. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ Digit. Med.* **2021**, *4*, 1–8.
28. Gartner, D.; Kolisch, R.; Neill, D.B.; Padman, R. Machine Learning Approaches for Early DRG Classification and Resource Allocation. *INFORMS J. Comput.* **2015**, *27*, 718–734.
29. Fitch FBMcCulloch Warren, S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **1943**, *5*, 115–133.
30. Rosenblatt, F. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **1958**, *65*, 386–408.
31. Minsky, M.; Papert, S.A. *Perceptrons: An Introduction to Computational Geometry*; MIT Press: Cambridge, MA, USA, 2017.
32. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Comput.* **1989**, *1*, 541–551.
33. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:14123555.
34. Atique, S.; Hsieh, C.-H.; Hsiao, R.-T.; Iqbal, U.; Nguyen, P.A.; Islam, M.; Li, Y.-C.; Hsu, C.-Y.; Chuang, T.-W.; Syed-Abdul, S. Viral warts (Human Papilloma Virus) as a potential risk for breast cancer among younger females. *Comput. Methods Programs Biomed.* **2017**, *144*, 203–207.
35. Chen, K.C.; Iqbal, U.; Nguyen, P.A.; Hsu, C.H.; Huang, C.L.; Hsu, Y.H.E.; Atique, S.; Islam, M.M.; Li, Yu.; Jian, We. The impact of different surgical procedures on hypoparathyroidism after thyroidectomy: A population-based study. *Medicine* **2017**, *96*, e8245.
36. Li, W.; Logenthiran, T.; Woo, W.L. Multi-GRU prediction system for electricity generation's planning and operation. *IET Gener. Transm. Distrib.* **2019**, *13*, 1630–1637.
37. Li, Z.; Wang, P.; Lu, H.; Cheng, J. Reading selectively via Binary Input Gated Recurrent Unit. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Macao, China, 15–16 August 2019; pp. 5074–5080.
38. Schlembach, D.; Hund, M.; Schroer, A.; Wolf, C. Economic assessment of the use of the sFlt-1/PlGF ratio test to predict preeclampsia in Germany. *BMC Health Serv. Res.* **2018**, *18*, 603.
39. Farhan, J.; Al-Jummaa, S.; Al-Rajhi, A.; Al-Rayes, H.; Al-Nasser, A. Documentation and coding of medical records in a tertiary care center: A pilot study. *Ann. Saudi Med.* **2005**, *25*, 46–49.
40. Nouraei SA, R.; Hudovsky, A.; Frampton, A.E.; Mufti, U.; White, N.B.; Wathen, C.G.; Sandhu, G.S.; Darzi, A. A study of clinical coding accuracy in surgery: Implications for the use of administrative big data for outcomes management. *Ann. Surg.* **2015**, *261*, 1096–1107.
41. Zafirah, S.A.; Amrizal, M.N.; Sharifah EW, P.; Aljunid, S.M. Incidence of clinical coding errors and implications on casemix reimbursement in a teaching hospital in Malaysia. *Malays. J. Public Health Med.* **2017**, *17*, 19–28.
42. Zafirah, S.A.; Nur, A.M.; Puteh, S.E.W.; Aljunid, S.M. Potential loss of revenue due to errors in clinical coding during the implementation of the Malaysia diagnosis related group (MY-DRG®) Casemix system in a teaching hospital in Malaysia. *BMC Health Serv. Res.* **2018**, *18*, 38.
43. Pongpirul, K.; Robinson, C. Hospital manipulations in the DRG system: A systematic scoping review. *Asian Biomed.* **2013**, *7*, 301–310.
44. Maryati, W.; Yuliani, N.; Susanto, A.; Wannay, A.O.; Justika, A.I. Reduced hospital revenue due to error code diagnosis in the implementation of INA-CBGs. *Int. J. Public Health Sci. (IJPHS)* **2021**, *10*, 354.
45. Nouraei, S.A.R.; Virk, J.S.; Hudovsky, A.; Wathen, C.; Darzi, A.; Parsons, D. Accuracy of clinician-clinical coder information handover following acute medical admissions: Implication for using administrative datasets in clinical outcomes management. *J. Public Health* **2015**, *38*, 352–362.