sustainability

Article

Analyzing Groups of Inpatients' Healthcare Needs to Improve Service Quality and Sustainability

Ming-Hsia Hsu [1,2], Chia-Mei Chen [2,*], Wang-Chuan Juang [3,4,*], Zheng-Xun Cai [2] and Tsuang Kuo [4]

[1] Department of Information Management, Kaohsiung Veterans General Hospital, Kaohsiung 813414, Taiwan; mhhsuchen3@gmail.com
[2] Department of Information Management, National Sun Yat-sen University, Kaohsiung 804959, Taiwan; d074020002@student.nsysu.edu.tw
[3] Quality Management Center, Kaohsiung Veterans General Hospital, Kaohsiung 813414, Taiwan
[4] Department of Business Management, National Sun Yat-sen University, Kaohsiung 804959, Taiwan; kuo@bm.nsysu.edu.tw
* Correspondence: cmchen@mis.nsysu.edu.tw (C.-M.C.); wcchuang@vghks.gov.tw (W.-C.J.)

Abstract: The trend towards personalized healthcare has led to an increase in applying deep learning techniques to improve healthcare service quality and sustainability. With the increasing number of patients with multiple comorbidities, they need comprehensive care services, where comprehensive care is a synonym for complete patient care to respond to a patient's physical, emotional, social, economic, and spiritual needs, and, as such, an efficient prediction system for comprehensive care suggestions could help physicians and healthcare providers in making clinical judgement. The experiment dataset contained a total of 2.9 million electrical medical records (EMRs) from 250 thousand hospitalized patients collected retrospectively from a first-tier medical center in Taiwan, where the EMRs were de-identified and anonymized and where 949 cases had received comprehensive care. Recurrent neural networks (RNNs) are designed for analyzing time-series data but are still lacking in studying predicting personalized healthcare. Furthermore, in most cases, the collected evaluation data are imbalanced with a small portion of positive cases. This study examined the impact of imbalanced data in model training and suggested an effective approach to handle such a situation. To address the above-mentioned research issue, this study analyzed the care need in the different patient groupings, proposed a personalized care suggestion system by applying RNN models, and developed an efficient model training scheme for building AI-assisted prediction models. This study observed several findings: (1) the data resampling schemes could mitigate the impact of imbalanced data on model training, and the under-sampling scheme achieved the best performance with an ACC of 99.80%, a PPV of 70.18%, an NPV of 99.87%, a recall of 82.91%, and an F1 score of 0.7602, while the model trained with the original data had a very low PPV of 6.42% and a low F1 score of 0.1116; (2) patient clustering with multi-classier could predict comprehensive care needs efficiently with an ACC of 99.87%, a PPV of 77.90%, an NPV of 99.90%, a recall of 92.19%, and an F1 score of 0.8404; (3) the proposed long short-term memory (LSTM) prediction model achieved the best overall performance with an ACC of 99.80%, a PPV of 70.18%, an NPV of 99.87%, a recall of 82.91%, and an F1 score of 0.7602.

Keywords: deep learning in healthcare; personalized healthcare; big data analysis; recurrent neural networks

Citation: Hsu, M.-H.; Chen, C.-M.; Juang, W.-C.; Cai, Z.-X.; Kuo, T. Analyzing Groups of Inpatients' Healthcare Needs to Improve Service Quality and Sustainability. *Sustainability* **2021**, *13*, 11909. https://doi.org/10.3390/su132111909

Academic Editors: Ganjar Alfian, Muhammad Anshari, Tony Hadibarata and Muhammad Syafrudin

Received: 24 September 2021
Accepted: 22 October 2021
Published: 28 October 2021

1. Introduction

Sustainability in healthcare is a new trend all across the globe. The term sustainability has many meanings in different contexts. Shediac-Rizkallah and Bone [1] considered it as maintaining the health benefits of the service over a long period and continuing the service within an organizational structure. In its simplest form, sustainability can be seen as "holding the gains" and "evolving as required" [2]. An example of doctors in primary and secondary care working together to develop new forms is a practical sustainability project.

The current medical and scientific knowledge could keep the majority of people healthy, preventing most heart diseases, diabetes, cancers, asthma, and other serious diseases. However, the mortality rate is still very high for patients with multiple comorbidities or elderly patients. In addition, for many years, there has been a visible increase in the death rate caused by cancer diseases as well. Therefore, medical institutes provide comprehensive healthcare services to improve sustainability, patient satisfaction, and work efficiency.

Comprehensive healthcare [3,4], aka total healthcare or holistic care, is considered as patient-centered care to achieve total patient care from different aspects, including physical, emotional, social, economic, and spiritual needs. Personalized care is the process of taking into account a patient's unique characteristics such as clinical history and risk factors to provide personalized care and treatments [5]. Both are important for treating elderly patients or patients with comorbidities or cancers, as they demand medical resources and require treatments across multidisciplinary medical professionals. Such high-quality care requires integrated healthcare resources but could improve patients living quality and health conditions.

Multiple factors play an integral part in considering whether patients need comprehensive care or not. To facilitate sustainable healthcare, medical institutes rely on an efficient comprehensive care system that keeps track of all the relevant patient records for monitoring disease progression, improving patient satisfaction, and reducing health care costs. With limited resources, an efficient method of forecasting patients' needs is critical for sustainability and service improvement.

The field of recurrent neural networks (RNNs) has evolved considerably in the last decades, and many RNN applications have been deployed in various contexts. However, it remained challenging to apply RNN technology to obtain an efficient prediction model. On the other hand, the available datasets collected from real environments contain mostly negatives and a small portion of positives, where such imbalanced data would affect the effectiveness of model training. The past studies rarely discussed the implication of imbalanced data as well as model training. Therefore, this study aimed to evaluate multiple contemporary deep learning models and the impact of imbalanced data on model training and to propose an efficient RNN-assisted prediction system that forecasts patients in need of comprehensive care.

Machine-learning-based clinical decision support systems (CDSS) play an important role in enhancing the quality of healthcare provided to patients. Neural networks (NNs) are becoming increasingly popular for a wide range of health care tasks, and among the different NN models, RNNs have been applied for prediction-related tasks by applying historical time-series data and learning the patterns from the past. Based on the literature review [6], few past studies have addressed personalized care needs or have implemented a decision-making system by applying deep learning techniques in support of the need for comprehensive care.

To address the abovementioned issues, this research analyzed the care need for different groups of inpatients and proposed a CDSS that applies deep learning techniques to predict the need for comprehensive care for different groups of inpatients. A prototype system was developed for practical evaluation with the aim to improve clinic staff work efficiency and healthcare sustainability.

This research makes the following contributions: (1) a deep-learning-based CDSS for personalized care that predicts the need for comprehensive care for different groups of inpatients; (2) an analysis of the prediction efficiency on different groupings of inpatients in need of comprehensive care; (3) an analysis of the impact of imbalanced data on deep learning model training; and (4) a prototype system to demonstrate the practicality of the proposed model.

The remainder of this article is organized as follows. Section 2 reviews the deep learning techniques applied to healthcare, and Section 3 explains the proposed model.

Experimental results applying empirical data to the proposed model are explained in Section 4, followed by concluding remarks and future work recommendations in Section 5.

2. Related Work

The Alliance for Natural Health International [7] is an internationally active non-governmental organization promoting natural and sustainable approaches to healthcare worldwide and has defined sustainable healthcare in the following way [8] as: a complex system of interacting approaches to the restoration, management, and optimization of human health that have an ecological base; that are environmentally, economically, and socially viable indefinitely; that work harmoniously both with the human body and the non-human environment; and which do not result in unfair or disproportionate impacts on any significant contributory element of the healthcare system.

Artificial intelligence is a large computer science field that encompasses logic, probability, and continuous mathematics; perception, reasoning, learning, and action; fairness, trust, social good, and safety; and performs sophisticated tasks in different industries. AI includes but is not limited to machine learning, deep learning, and natural language processing (NLP) and whichever techniques can help medical staff find important information from data. Some medical organizations have already applied AI for daily routines as a CDSS; it can assist medical staff to make decisions, reduce false rates, and increase productivity by performing risk assessments, analyzing different types of diseases, and finding the relationships among the medical data.

Cadar et al.'s study [9] on analyzing clusters of customers who use medicinal plants indicated that analyzing by groupings contributes a better understanding of user behaviors. It inspired this study to cluster patients in groups in order to have a better understanding of patients' care needs as well as an efficient comprehensive care prediction. Antoniadi and Mooney [10] advised that a fair and usable ML-assisted CDSS should provide the right decision with a clear explanation on how that decision was made and should preserve the privacy of the information used to fulfill its purpose in the healthcare environment.

Many factors affect the success of machine learning on a given task, and among them, the representation and quality of data are critical. A study [11] indicates that irrelevant, redundant, noisy, or unreliable data make knowledge discovery difficult during the training phase. Therefore, data preprocessing and feature selection are important for constructing an effective prediction model. This study selected key features to establish the RNN-based CDSS system, which provide good accuracy.

Datta et al. [12] evaluated the efficiency of deep neural network architectures and concluded that soft attention improves the image classification performance as it boosts the important features and suppresses noisy features. Dawoodbhoy et al. [13] conducted an interview and suggested that AI could improve patient flow by streamlining administrative tasks and optimizing resource allocation and that real-time data analytics systems could support clinical decision-making.

Machine learning techniques have been extensively researched for disease predictions. Data-mining algorithms like naïve Bayes, linear regression, decision tree, and random forest have been used on the dataset taken from the UCI machine learning repository. The past study [14] showed that the random forest model produced the highest accuracy for heart disease prediction among the aforementioned techniques. Nawaz et al.'s study [15] compared different ML models for predicting disorders of the heart and blood vessels, including support machine vector, K-nearest neighbor, naïve Bayes, artificial neural networks, random forest, and gradient descent optimization. It concluded that gradient descent optimization yields the best performance.

The COVID-19 pandemic has caused huge damage in the world since 2019. Several studies [16–19] applied deep neural networks to detect coronavirus diseases from the X-ray and computed tomography (CT) images, and some evaluated multiple convolutional neural network (CNN) models. Chen et al. [20] proposed an efficient deep learning model for removing irrelevant backgrounds, extracting spatial features, and automatically segmenting lung lesions from CT images. Vidal [21] et al. proposed a multi-stage transfer learning approach to obtain a robust system able to segment lung regions from portable X-ray devices despite the lack of samples. Haneczok and Delijewski performed drug discovery screenings based on supervised ML models for identifying potential coronavirus diseases and compared the prediction performance of different ML models with different molecular representations. In their study, ML approaches enable drug discovery screenings.

Khreis et al. [22] evaluated CDSS with the assistance of ML techniques to help healthcare providers in making the clinical judgement of patients' therapy and reducing medication errors. They concluded that applying ML techniques into CDSS design can remove duplicated and unimportant alerts. Baker et al. [23] applied a hybrid NN model, CNN+LSTM, for estimating blood pressure from raw electrocardiogram and photoplethysmogram waveforms and demonstrated that ML techniques are an effective approach for blood pressure estimation and are ready for implementing them into wearable devices. Minnema et al. [24] evaluated different CNN training strategies for CT image segmentation. The results indicate that analyzing the structure of the images helps performance improvement. Lakshmanaprabu et al. [25] proposed a deep neural network model for lung cancer classification on CT lung images by applying linear discriminate analysis for feature reduction and a modified gravitational search algorithm for weight optimization.

Liu et al. [26] developed a CNN model to identify tuberculosis infection from X-ray images. Dong et al. [27] applied a CNN model on chest X-ray image classification to detect multiple diseases such as pleural thickening, otosclerosis, and pulmonary interstitial hyperplasia, not just tuberculosis. This multi-classification model required more than 16,000 labeled X-ray images to train and yielded an accuracy of 82.2%.

A study [28] reviewed computerized CDSS to support automatic detection of critical conditions and observed a trend toward the use of data-driven algorithms. Given the current lack of best practice guidance to ML and AI, a study [29] highlighted that patients and healthcare professionals require clinical prediction models to accurately guide healthcare decisions and that ML and AI models potentially improve diagnostic accuracy and reliable prediction.

Most past research applied CNN-based classification models to predict diseases based on images. Only a few published studies focused on applying RNN for predicting patient care, and most discussed the impacts and needs of applying AI technologies to the healthcare sector. To our best knowledge and based on the above literature review, the present study is the first attempt to apply an RNN model to develop a CDSS that predicts personalized care needs.

3. Proposed Methodology

3.1. LSTM-Based Personalized Care Prediction System

The literature review [9,24] inspired our study of personalized health care prediction by analyzing the groups of patients for comprehensive care needs. The literature review [27] has demonstrated that multi-classifiers could yield high detection performance. Therefore, this study applied a multi-classifier for personalized health care prediction.

Figure 1 presents the proposed system architecture, where the left dashed box indicates the training process of constructing the prediction model based on the historical data, and the right box shows the process of applying it in support of personalized care decision making. The training data were the medical data of hospitalized inpatients collected from a medical center that has practiced the CC service since 2017, where the de-identified data include daily medical records extracted from EMR. The proposed method applies the undersampling technique to mitigate the impact of imbalanced data on model training, clusters patients in groups, and adopts model-selected features to improve prediction performance.

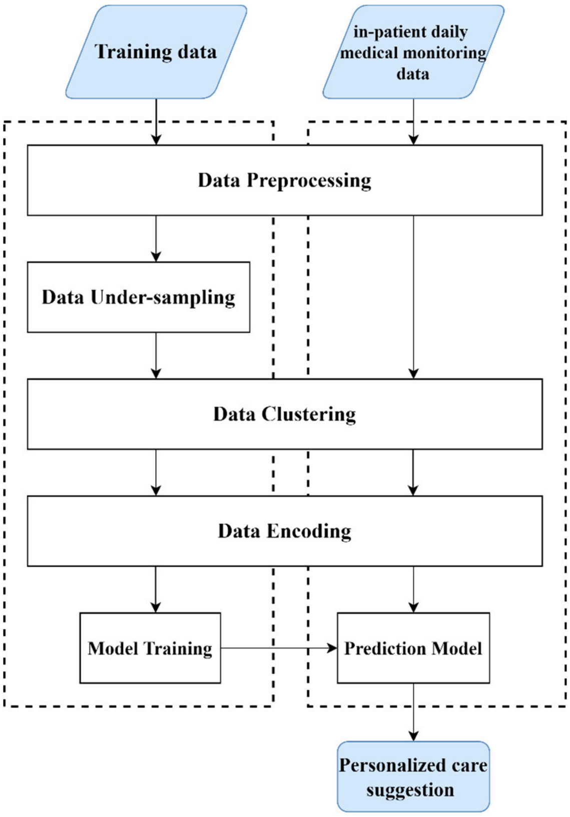

Figure 1. The proposed system architecture.

In the evaluation, the steps mentioned above in the training process were evaluated with different approaches in order to obtain the best prediction model. For example, feature selection is one of the key components of ML model training. This study considered all the vital indexes in the daily medical records as the feature set. Another method of selecting important features is relying on domain experts to select key features. This study evaluated the above two feature selection approaches, which are elaborated more in the discussion section.

This study proposed an LSTM model for predicting personalized care needs, where LSTM is an improved RNN model that learns the temporal dependency of time-series data. LSTM networks perform exceptionally well on sequential data due to their ability to "remember" what they have previously seen. The proposed LSTM network is composed of four layers, as illustrated in Figure 2. The input layer is fed with the time-series data: the current medical data (D_t) of a given patient plus his/her past N-day historical medical records ($\mathrm{D}_{\mathrm{t}-6}$~$\mathrm{D}_{\mathrm{t}-1}$), where N = 6 days. The softmax [30,31], aka normalized exponential function, is used in the dense layer, as illustrated in Figure 3. It normalizes the output into a probability distribution over the predicted output classes. The output layer forecasts if a given patient needs comprehensive care based on the given input data.

The softmax function is used as the activation function in the output layer of NN models. It transforms a vector of k values into another vector of k values that sum to 1 and has a normalized probability distribution. The softmax function $\sigma(\vec{z})_i$ is expressed by Formula 1 below, where $\vec{z}$ is the input vector, z_i is an element of the input vector, and e^{z_i} denotes the standard exponential function of element z_i.

$$\sigma\left(\vec{z}\right)_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \quad (1)$$

Neural networks learn to map a set of inputs to a set of outputs from training data and are trained using an optimization process that requires an objective function (loss function in this case) to evaluate a candidate solution (i.e., a set of weights). NN models typically are trained using the stochastic gradient descent optimization algorithm, and weights are updated using the backpropagation of error algorithm, where the gradient in gradient descent refers to an error gradient. Therefore, the objective function for NN models usually is a loss function calculating the model error.

Figure 3 illustrates the optimization process of the proposed LSTM model. It repeatedly improves the prediction by adjusting the features and their weighting in an LSTM cell in order to minimize the loss function. For a predicted result, it compares with the expected result and calculates the loss. The optimizer then revises the weighting of the features in the LSTM cell to minimize the loss.

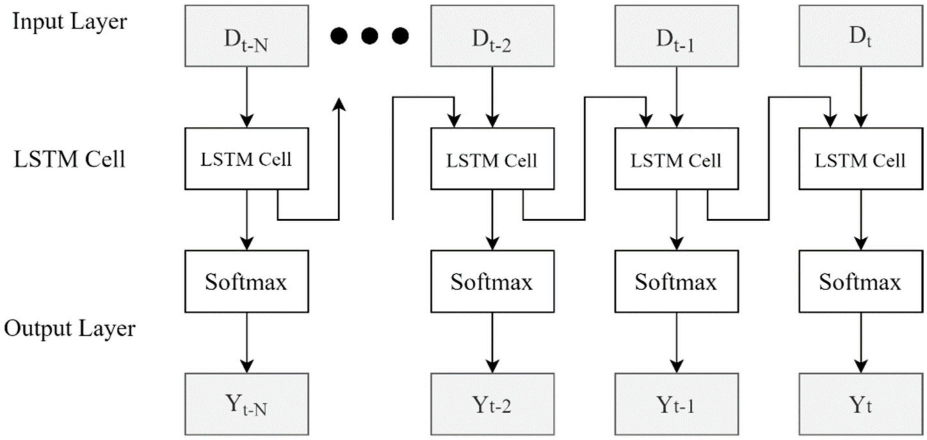

Figure 2. The proposed LSTM networks.

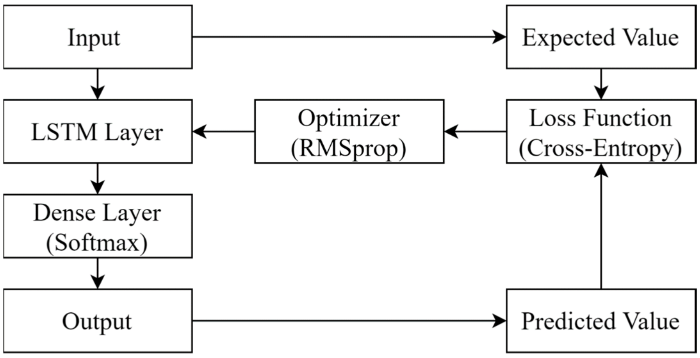

Figure 3. An illustration of the optimization process.

3.2. Hyperparameter Selection

Based on our preliminary evaluation of parameter tuning, Table 1 lists the hyperparameters chosen by the proposed model. An LSTM cell contains two activation functions: one is for the model and the other is for recurrent usage. A performance analysis study [32] comparing 23 different combinations of activation functions recommended hyperbolic tangent for activation function and sigmoid for recurrent activation function.

The literature review indicated that cross-entropy performs well at generalizing beyond the training data and initializing the weights [33] and that binary cross-entropy yields the best performance for binary classification applications [34]. Therefore, the proposed prediction model chose binary cross-entropy as the loss function.

An analysis of the optimization algorithms [35] concluded that Adam, RMSprop, and AdaDelta perform approximately well; the Keras documents [36] recommended RMSprop for RNN models. As LSTM falls into the RNN category, the proposed system implemented RMSprop as the optimizer.

Table 1. The hyperparameters of the proposed LSTM prediction model.

Name	Value
LSTM Parameters	
Learning Rate	0.01
Loss Function	Cross-entropy
Activation Function	hyperbolic tangent
Recurrent Activation Function	Sigmoid
Recurrent Dropout	0.3
Batch size	200
Number of epochs	200
Optimizer	RMSprop
Optimizer Parameters	
Learning Rate	0.01
Rho	0.9
Epsilon	None
Decay	0

The number of instances from the training data set used in the estimate of the error gradient is called the batch size and is an important hyperparameter that influences the dynamics of the learning algorithm. One training epoch means that the learning algorithm has made one pass through the training data set, where the training data are separated into randomly selected "batch size" groups [37].

A too-small batch size may fail to converge and may lead to underfitting, while a too-large batch size may exhaust computing resources. An analysis study [38] observed that controlling the ratio of the batch size to the learning rate to be not too large could achieve a good generalization ability. An investigation [39] demonstrated that a large batch tends to converge to sharp minimizers, leading to poor generalization and degrading the model quality, while another past work [40] suggested increasing the batch size instead of decaying the learning rate. Therefore, our preliminary testing ran multiple combinations of epoch and batch size and concluded with the parameter values listed in Table 1.

3.3. Data Pre-Processing

The literature review proved that data pre-processing has a significant impact on machine learning performance. The representation and quality of the instance data are some of the foremost factors for building an effective model. Noisy or invalid data may disrupt the model training.

To construct an effective ML model, the data preprocessing module examined the validity of the patient's records, including missing data points, empty values, and outlier values. It applies linear interpolation to estimate missing data, for example, a missing blood pressure measurement in a patient's daily vital record. A default value of 0 is filled in for null or empty values in a data field. A value larger than two standard deviations of a given field is considered as an outlier and is replaced by an interpolation value.

The data were collected from a medical center composed of diverse medical information such as body temperature, blood pressure, comorbidity, etc. The private information about inpatients, such as name and birthday, had been removed and de-identified. The reasonable range of each data field was verified, and unreasonable or missing data were corrected by interpolation.

3.4. Data Resampling

Effective model training requires a balanced or approximately balanced class distribution. However, the training data collected from the real environment were imbalanced, where most inpatients made up a majority class and where only a very small amount of the comprehensive care patients made up a minority class. In this study, less than 0.5% of inpatients received comprehensive care, which is extremely imbalanced data. Sampling is a process of resampling data that generates a sampling distribution based on the actual data, and the resampling is to produce a less-skewed training dataset for model training purposes. Two types of resampling strategies are commonly applied: over-sampling and under-sampling.

Over-sampling (or up-sampling) increases the number of minority class samples to reduce the degree of imbalanced distribution. For over-sampling, this study adopted the synthetic minority over-sampling technique (SMOTE) [41]. It is one of the most commonly used over-sampling approaches in which the minority class is over-sampled by creating synthetic examples. The synthetic samples are generated based on randomly selected k nearest neighbors of the minority class. They cause the classifier to create larger and less-specific decision regions so that the more general regions could be learned for the minority class.

On the other hand, since there are much more samples of the majority class than the minority one in the imbalanced class distribution problem, under-sampling approaches reduce the number of samples from the majority class. To keep the selected samples representative and to prevent the problem of information loss in most under-sampling techniques, near-neighbor methods are widely used. This study adopted the NearMiss3 under-sampling approach [42], which selects the majority class samples for which the average distance to the N nearest-neighbors is the largest. In other words, it attempts to select the samples of the decision boundary. Based on our preliminary study, under-sampling yields better performance and is suggested in the proposed method.

3.5. Data Clustering

According to our preliminary study on analyzing different patient groupings and interviewing the physicians, the proposed method divided the patients into six groups based on their age and disease, where there are three age groups (<65, 65~80, >80) and two disease groups (with or without cancer). In practice, elderly patients usually require more special care needs than young ones, so the grouping specifically ranges out two aged groups: 65~80 and >80 in order to emphasize their different needs. Likewise, cancer patients need special care as well. Table 2 summarizes the patient grouping results and the data distribution of each group, where CC denotes patients who have received comprehensive care.

Table 2. The proposed patient clustering.

Group	No. of Patients	No. of CC Patients
$C_{(<65\ \text{cancer})}$	28,936	112
$C_{(65\sim80\ \text{cancer})}$	4073	34
$C_{(>80\ \text{cancer})}$	14,068	28
$C_{(<65\ \text{no-cancer})}$	23,602	76
$C_{(65\sim80\ \text{no-cancer})}$	35,502	126
$C_{(>80\ \text{no-cancer})}$	19,826	98

3.6. Feature Selection and Encoding

There were three types of data in this study: categorical, continuous numerical, and discrete numerical data, as shown in Table 3, and all the data fields collected were considered as features in the proposed prediction model. The data need to be represented in numerical form in order to be fed into ML models as they only work with numerical values.

Table 3. The description of the proposed feature set.

Data Type	Data Name	Description
Continuous numerical	Respiration	Inpatients' respiration rate (breaths per minute).
	SpO2	Inpatients' pulse oximetry (%).
	Systolic blood pressure	Inpatients' blood pressure (mmHg).
	Pulse	Inpatients' pulse rate (beats per minute).
	Temperature	Inpatients' body temperature (°C).
Discrete numerical	Coma Status	The Coma Status sums up the three coma indicators of consciousness: eye reaction, vocal reaction, and motion reaction, and is a positive integer ranging 1~15.
	Nutrition	Nutrition scale evaluated daily by nurses.
	BSRS	Brief Symptom Rating Scale evaluated daily by nurses.
	Pain	Pain scale evaluated daily by nurses.
	Hospitalized days	The number of days that the patient is hospitalized.
Categorical	Comorbidity	Comorbidity denotes the illnesses of the patient.
	Age Clustering	Age clustering indicates the age range of the patient.
	Disease Clustering	Disease clustering denotes if the patient has a cancer or not.

Categorical data, such as disease and gender, need to perform a transformation before applying an ML model. This study applied one-hot encoding for categorical data, which converts each categorical value into one bit in a bit string.

For numerical data, this study applied data normalization to quantify the significance of the feature value. For continuous numerical data, such as temperature and blood pressure, the proposed system applied the MinMaxScaler standardization approach for data normalization in order to quantify the significance of the feature, where MinMaxScaler transforms a feature by scaling it to a given range to preserve the shape of the original distribution. For each value in a feature X, MinMaxScaler subtracts the minimum value in the feature and divides by the range, where the range is the difference between the original maximum and the original minimum value of the feature. MinMaxScaler standardization is expressed in the following formula, where max and min stand for the maximum and minimum values of feature X, respectively, and X_{std} is the standardized value of feature X ranging in 0~1.

$$X_{std} = \frac{X - min}{max - min} \tag{2}$$

Regarding discrete numerical data, such as check-in and check-out dates, this study transformed these parameters into meaningful values for the ML model, for example, in this case the number of hospitalized days. The number of hospitalized days, D_{delta}, a discrete numerical feature, can be expressed as follows, where DD and ED mean the check-out and check-in dates, respectively.

$$D_{delta} = DD - ED \tag{3}$$

A discrete numerical feature, such as coma status, combines multiple pieces of medical information into one. The Glasgow coma scale (GCS) is a neurological scale to give a reliable and objective method of recording the state of a patient's consciousness, which contains three physical aspects of coma seriousness: eye (CS_{eye}), vocal (CS_{vocal}), and motion

(CS_{motion}). This study combined the above three scales into one feature, coma status (CS_{all}), as expressed below.

$$CS_{all} = CS_{eye} + CS_{vocal} + CS_{motion} \quad \in [1, 15] \tag{4}$$

Table 4 illustrates how the patient's EMR is represented originally and encoded for applying the ML models, where the two red boxes indicate the corresponding data records before and after the feature encoding. Table 4A lists the raw data of a patient's medical records including the dates of the admission and discharge, age, daily health measurements, diseases, etc., where the first record denotes the patient's medical record on the first day of the admission (15 May 2019) and so on. Based on the aforementioned feature-encoding methods, the data were encoded as outlined in Table 4B. The LSTM acquires the past N days of the EMR data for predicting the patient's health care need, where N = 6 in this case. For example, to predict the patient's health care need on his/her 8th day (22 May 2019), the input to the prediction model is the time-series feature vectors dated from 16 May 2019 to 22 May 2019 (a total of 6 + 1 days) as marked in the red box. Figure 2 describes the neural networks of the proposed LSTM model. For this illustration, the input layer is the time-series feature vectors (D_2, D_3, … , D_8, where t = 8), and these feature vectors are the input to the corresponding LSTM cells.

Table 4. An illustration of the EMR and feature encoding.

(**A**) The EMR raw data.

ID	Admission	Discharge	Record	Age	Respiration	Cancer	Diabetes	Stroke	…
0005	20190515	20190904	20190515	74	24	N	Y	N	…
0005	20190515	20190904	20190516	74	23	N	Y	N	…
0005	20190515	20190904	20190517	74	21	N	Y	N	…
0005	20190515	20190904	20190518	74	21	N	Y	N	…
0005	20190515	20190904	20190519	74	22	N	Y	N	…
0005	20190515	20190904	20190520	74	20	N	Y	N	…
0005	20190515	20190904	20190521	74	25	N	Y	N	…
0005	20190515	20190904	20190522	74	22	N	Y	N	…
…	…	…	…	…	…	…	…	…	…

(**B**) The encoded features.

Days in hospital	Respiraion	SpO2	Pulse	Nutrition	Age cluster	Cancer cluster	Comorbidity	…
1	0.74	0.45	0.84	0.33	1	0	000010000000	…
2	0.71	0.36	0.80	0.33	1	0	000010000000	…
3	0.66	0.18	0.85	0.33	1	0	000010000000	…
4	0.66	0.18	0.79	0.33	1	0	000010000000	…
5	0.68	0.27	0.70	0.33	1	0	000010000000	…
6	0.63	0.09	0.77	0.33	1	0	000010000000	…
7	0.77	0.55	0.86	0.33	1	0	000010000000	…
8	0.68	0.55	0.79	0.33	1	0	000010000000	…
…	…	…	…	…	…	…	…	…

4. System Evaluation

The proposed system and the models used for evaluation were implemented and evaluated in the same environment. They were developed in the programming language Python with the open-source library Keras. The experiments were run on a personal computer with the computing resources described in Table 5.

Table 5. Computing environment used in the evaluation.

Hardware Specification	
CPU	Intel i7-8700
Memory	64 GB
GPU	GTX 1080 Ti
Hard Disk (Dataset store)	Toshiba 6TB 7200RPM

Table 6 outlines the evaluation that consists of the following experiments: Exp. 1 evaluated the impact of imbalanced data; Exp. 2 analyzed whether a bi-classifier or a multi-classifier is suitable for predicting personalized care; Experiment 3 analyzed which patient grouping is suitable for practicing the personalized care prediction; Experiment 4 compared two different feature selection schemes, human vs. AI; Experiment 5 analyzed the performance of different prediction models in order to construct the best model for providing personalized care suggestions.

Table 6. The design of the system evaluation.

Exp. ID	**Exp. Title**	**Exp. Purpose**
Exp. 1	Imbalanced data	Evaluate the impact of imbalanced data
Exp. 2	Bi-vs multi-classifier	Evaluate the performance of the two types of classifiers
Exp. 3	Patient clustering	Analyze the performance of different groupings
Exp. 4	Feature selection	Compare the performance of features selected by the domain experts with those selected by the ML model.
Exp. 5	Model selection	Compare the performance of different prediction models

This research adopted accuracy, precision, recall, and F1 score as primary performance measurements, where they were defined by the confusion matrix as shown in Table 7. For completeness, the negative predictive value (NPV) is provided for reference. Except presenting the mean of a result, this study also calculated its confidence interval and adopted the most commonly used confidence level, 95%. A confidence interval represents the estimation of a result and is comprised of two values: the upper and lower bound. In other words, the confidence interval displays the probability that the result will fall between the pair of the values around the mean.

$$Accuracy\ (ACC) = \frac{TP + TN}{TN + FN + FP + TP} \tag{5}$$

$$Positive\ predictve\ value (PPV) = \frac{TP}{FP + TP} \tag{6}$$

$$Negative\ predictive\ value\ (NPV) = \frac{TN}{FN + TN} \tag{7}$$

$$Recall = \frac{TP}{FN + TP} \tag{8}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{9}$$

Because the main purpose of the prediction was to identify the patients who need comprehensive care, recall was a primary performance measure considered in this study. Because of highly unbalanced data with a huge number of non-comprehensive-case cases (negatives), the negative predictive value was not considered in this study. However, all the measurements were calculated for reference as well as the overall performance analysis.

Accuracy is the degree of closeness to a true value and is defined as the proportion of the correct predictions among the total number of the data set. Precision, aka positive predictive value, is defined as the fraction of the predicted comprehensive care instances divided by the correctly predicted comprehensive care instances. Recall, aka sensitivity, is the fraction of the actual comprehensive care instances divided by the correctly predicted comprehensive care instances. Negative predictive value is defined as the fraction of the predicted non-comprehensive-care cases divided by the correctly predicted non-comprehensive-care cases.

The F1 score combines the precision and recall of a prediction model into a single metric by taking their harmonic mean. As multiple performance indices were considered in the evaluation, this study considered the F1 score as the overall performance comparison measure when there was a trade-off among the different prediction models.

Table 7. Confusion Matrix.

	Actually Positive	**Actually Negative**
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

4.1. Study Population

This study adopted de-identified inpatient medical records collected over two years from a medical center in Taiwan that has practiced comprehensive care since 2017. The physiological measurements of all hospitalized patients were available from the EMR system, including vital signs, emotional conditions, sleeping quality, food intake, medication status, etc. A total of 250,000 inpatient cases with 1.9 million medical records and 949 comprehensive care patients was referenced to model training and evaluation, where comprehensive care patients were verified manually by physicians.

The data set had the ratio of negatives to positives = 0.379%. All the neural networks were trained using 50% of the data, and the remaining 50% of the data was unseen to the networks for testing purposes. A prediction was considered valid if it fell within two weeks of the actual date of occurrence. Cross-validation tests were performed to eliminate the overfitting issue.

4.2. Patient and Public Involvement

This research was carried out without patient or public involvement in the design or result interpretation. Patients and members of the public did not contribute to the writing or editing of this manuscript.

4.3. Exp. 1: Imbalanced Data

Exp. 1 compares the prediction performance with and without data resampling to evaluate the impact of imbalanced data on model training. Two data resampling strategies were considered for mitigating the imbalanced data issue, where up-sampling applied the synthetic minority oversampling technique (SMOTE) and down-sampling adopted the Fourier method. The data resampling strategies adjusted the ratio of negatives to positives into 8:2.

Table 8 summarizes the attributes of the collected data set and those of the training data after applying the data sampling strategies. The collected data set was divided evenly into two parts: one for training and the other for testing. Therefore, both had the same amount of positive and negative cases, and the number of the associated data records in training and that in testing were approximately even with the ratio of 1.22:1 as indicated in the table. As mentioned above, the data sampling was to obtain a better data distribution for improving the model training. This study applied the resampling strategies and redistributed the positives and negatives to the ratio of 2:8, where the original data had an extremely skewed ratio of 3.79:1000, as listed in the table.

Table 8. The dataset description of the different samplings.

The Properties of the Collected Data			
No. of patients	252,014		
No. of positives	948		
No. of negatives	251,066		
Ratio of positives:negatives	3.79:1000		
No. of patients in training data	126,007		
No. of patients in testing data	126,007		
Ratio of no. of patients in training: that in testing	1:1		
No. of data records	2,518,887		
No. of data records in training data	1,387,137		
No. of data records in testing data	1,137,750		
Ratio of no. of data records in training: that in testing	1.22:1		
No. of data records in positives	948		
No. of data records in negatives	2,517,939		
Sampling strategy for training data			
Resampling strategy	None	Over-sampling	Under-sampling
No. of positives (in training data)	474	31,383	474
No. of negatives (in training data)	125,533	125,533	1896
Ratio of positives: negatives (in training data)	3.79:1000	2:8	2:8
No. of data records in positives (in training data)	474	31,383	474
No. of data records in negatives (in training data)	1,386,663	1,939,845	31,119
No. of data records (in training data)	1,387,137	1,971,228	31,593
No. of -positives in testing data	474		
No. of negatives in testing data	125,533		
No. of data records in testing data	1,137,750		

Table 9 outlines the impact of different data sampling strategies on the performance of the model training. In this study, the data set contained a very huge amount of negative samples. Therefore, even though under-sampling eliminated a large amount of negative cases from the majority class, it still retained enough representative samples of the majority class so that the ML models were able to learn the class patterns. On the other hand, with an extremely small number of positive samples, some synthetic positive cases generated by over-sampling might not be able to provide the expected patterns of the minority class and hence affect the efficiency of the model training. The results indicate that, even though under-sampling had the least number of samples as well as data records, it consumed the least training time and outperformed the other two sampling strategies. No resampling performed the worst, which demonstrates that imbalanced data may affect the effectiveness of the model training. This experiment implies that data resampling could mitigate the impact of imbalanced data, and under-sampling obtained the best prediction model.

Table 9. The evaluation of the impact of resampling.

Resampling Strategy	**None**		**Over-Sampling**		**Under-Sampling**	
Actual **Predict**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Positive	202	2944	371	1119	393	167
Negative	272	122,589	103	124,414	81	125,366
	Mean	CI (95%)	Mean	CI (95%)	Mean	CI (95%)
Accuracy (ACC)	97.45%	[97.39%, 97.51%]	99.59%	[99.54%, 99.64%]	**99.80%**	**[99.78%, 99.82%]**
Precision (PPV)	6.42%	[6.63%, 6.61%]	46.96%	[46.91%, 47.00%]	**70.18%**	**[70.14%, 70.22%]**
NPV	97.65%	[97.60%, 97.70%]	99.67%	[99.65%, 99.69%]	**99.87%**	**[99.85%, 99.89%]**
Recall	42.62%	[42.66%, 42.85%]	78.27%	[78.21%, 78.33]	**82.91%**	**[82.82%, 83.00%]**
F1	0.1116	[0.1072, 0.1159]	0.5870	[0.5814, 0.5925]	**0.7602**	**[0.7534, 0.7670]**

4.4. Exp. 2: Bi-vs Multi-Classifier

A personalized health care prediction model is a multi-classes question and has two solution options: a single multi-classifier or a binary classifier for each group. Therefore, Exp. 2 analyzed the three classification models: (1) a multi-classifier for all the patient groups; (2) one bi-classifier for each group; (3) one bi-classifier for all the patients. The results are summarized in Table 10, where $M(C)$ denotes the bi-classifier for all patients, $M(C_{group\_i})$ denotes the bi-classifier for patients in group i, and $M(C_{group\_1}, C_{group\_2}, \ldots, C_{group\_n})$ denotes the multi-classifier for the patient clustering of $C_{group\_1}, C_{group\_2}, \ldots, C_{group\_n}$. The results demonstrate that the multi-classifier yields the best prediction performance and that patient clustering improves the prediction performance as it reduces data complexity. Comparing with the best training time, the increased training time of the multi-classifier is neglectable.

Table 10. The results of Exp 2 evaluating different classifiers.

Model	**Multi-Classifier, M(C<65, C65~80, C≥80)**		**Bi-Classifier for Each Group, M(C<65) + M(C65~80) + M(C≥80)**		**Bi-Classifier for All Patients, M(C)**	
Actual **Predict**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Positive	442	143	429	161	393	167
Negative	32	125,390	45	125,372	81	125,366
	Mean	CI (95%)	Mean	CI (95%)	Mean	CI (95%)
Accuracy (ACC)	**99.86%**	**[99.80%, 99.92%]**	99.84%	[99.82%, 99.86%]	99.80%	[99.78%, 99.82%]
Precision (PPV)	**75.56%**	**[75.49% 75.63%]**	72.71%	[72.65%, 72.77%]	70.18%	[70.14%, 70.22%]
NPV	**99.89%**	**[99.87%, 99.90%]**	99.87%	[99.86%, 99.88%]	99.87%	[99.85%, 99.89%]
Recall	**93.25%**	**[93.18%, 93.32%]**	90.51%	[90.40%, 90.62%]	82.91%	[82.82%, 83.00%]
F1	**0.8347**	**[0.8230, 0.8464]**	0.8064	[0.7996, 0.8132]	0.7602	[0.7534, 0.7670]
Training Time (min)	115	[113.47, 116.53]	165	[163.06, 166.94]	**103**	**[102.08, 103.92]**

4.5. Exp. 3: Patient Clustering

Exp. 3 analyzed different patient groupings to find an optimal patient clustering for personalized care suggestions. Patient clustering was factored by age and cancer disease, and Exp. 3 analyzed the following patient groupings:

- Cluster 0: all patients in one group, no clustering.
- Cluster 1: ($C_{<65}$, $C_{\geq 65}$)
- Cluster 2: ($C_{<65}$, $C_{65\sim 80}$, $C_{\geq 80}$)
- Cluster 3: (C_{cancer}, $C_{no\text{-}cancer}$)
- Cluster 4: ($C_{<65\ cancer}$, $C_{65\sim 80+caner}$, $C_{\geq 80+caner}$, $C_{<65+no\text{-}cancer}$, $C_{65\sim 80+no\text{-}cancer}$, $C_{\geq 80\ no\text{-}cancer}$)

The results of Exp. 3 are summarized in Table 11 and Figure 4. Considering clustering in age, the clustering of ($C_{<65}$, $C_{\geq 65}$) yielded the best performance. However, based on the interviews with physicians and clinical staff who practice the comprehensive care service, they preferred clustering patients in three age groups: $C_{<65}$, $C_{65\sim80}$, $C_{\geq80}$, as it fits in their operation. Based on their practice experience, patients with cancer often require comprehensive care than those without cancer. Therefore, the age groups were further divided by cancer disease, which produced Cluster 4.

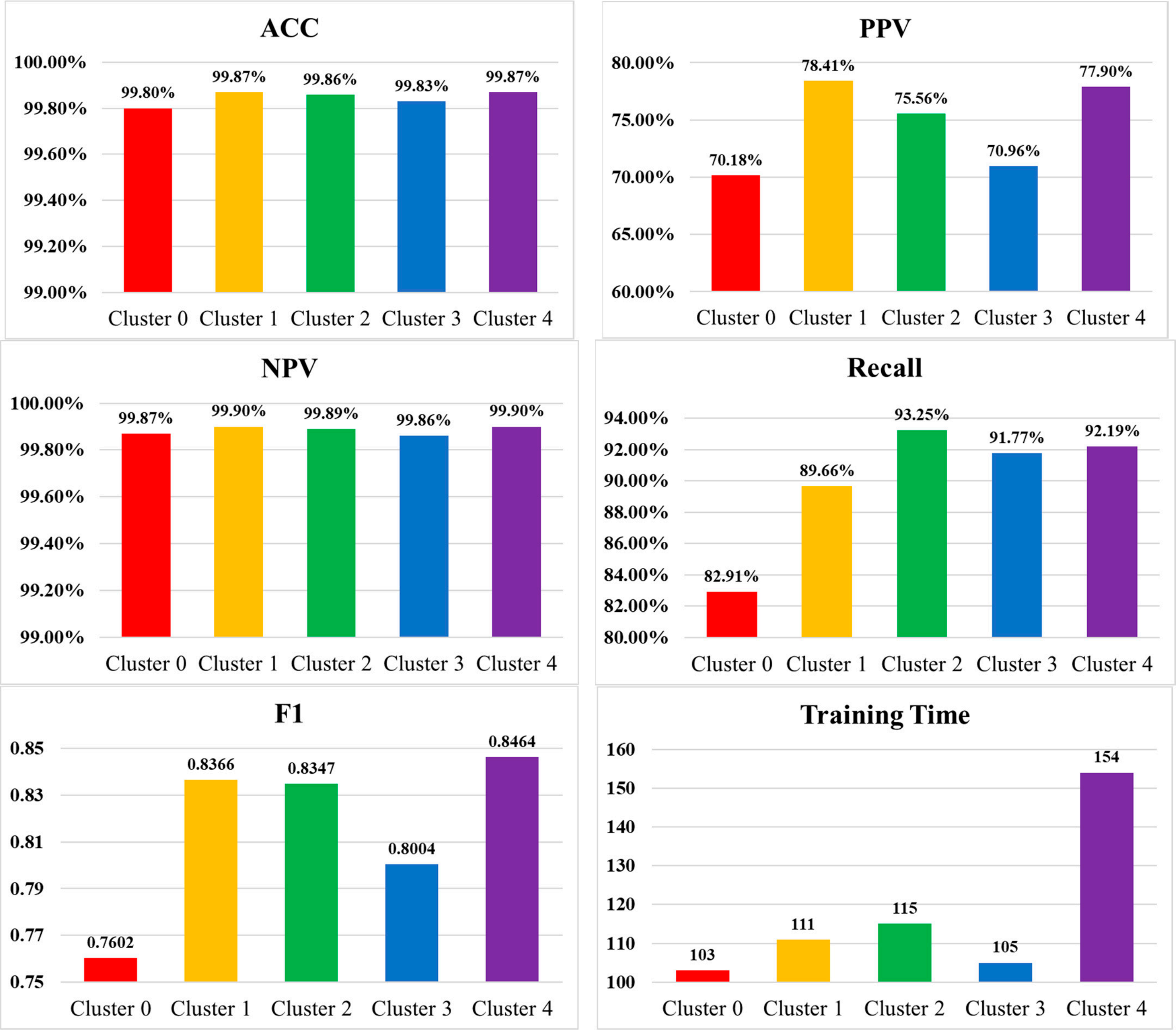

Figure 4. The results of Exp. 3 evaluating different patient clustering.

Table 11. The results of Exp. 3 evaluating different patient clustering.

Cluster	**Cluster 0**		**Cluster 1**		**Cluster 2**	
Predict \ **Actual**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Positive	393	167	425	117	442	143
Negative	81	125,366	49	125,416	32	125,390
	Mean	CI (95%)	Mean	CI (95%)	Mean	CI (95%)
Accuracy (ACC)	99.80%	[99.78%, 99.82%]	**99.87%**	**[99.85%, 99.89%]**	99.86%	[99.84%, 99.88%]
Precision (PPV)	70.18%	[70.14%, 70.22%	**78.41%**	**[78.34%, 78.48%]**	75.56%	[75.46%, 75.66%]
NPV	99.87%	[99.85%, 99.89%]	**99.90%**	**[99.89%, 99.91%]**	99.89%	[99.88%, 99.90%]
Recall	82.91%	[82.82%, 83.00%]	89.66%	[89.64%, 89.68%]	93.25%	[93.23%, 93.27%]
F1	0.7602	[0.7534, 0.7670]	0.8366	[0.8310, 0.8422]	0.8347	[0.8297, 0.8397]
Training Time (min)	**103**	**[102.08, 103.92]**	111	[109.9, 112.1]	115	[114.00, 446.00]

Model	**Cluster 3**		**Cluster 4**	
Predict \ **Actual**	**Positive**	**Negative**	**Positive**	**Negative**
Positive	435	178	437	124
Negative	39	125,355	37	125,409
	Mean	CI (95%)	Mean	CI (95%)
Accuracy (ACC)	99.83%	[99.82%, 99.84%]	**99.87%**	**[99.85%, 99.89%]**
Precision (PPV)	70.96%	[70.85%, 71.07%]	77.90%	[77.83%, 77.97%]
NPV	99.86%	[99.86%, 99.86%]	**99.90%**	**[99.89%, 99.91%]**
Recall	91.77%	[91.76%, 91.78%]	**92.19%**	**[92.14%, 92.24%]**
F1	0.8004	[0.7961, 0.8047]	**0.8464**	**[0.8433, 0.8495]**
Training Time (min)	105	[104.16, 105.84]	154	[152.79, 155.21]

Cluster 4, ($C_{<65+\text{cancer}}$, $C_{65\sim80+\text{caner}}$, $C_{\geq80+\text{caner}}$, $C_{<65\ \text{no-cancer}}$, $C_{65\sim80\ \text{no-cancer}}$, $C_{\geq80\ \text{no-cancer}}$), yielded the best performance, except that its precision was slightly less than Cluster 1. Overall, Cluster 4 outperformed all the others but required a longer training time. The results imply that a classifier with many classes requires longer training than one with not too many classes, but this would not affect its performance. The increased training time is in an acceptable range, so Cluster 4 was adopted in the proposed system.

4.6. Exp. 4: Feature Selection

Exp. 4 analyzed the performance of two methods of feature selection: human-selected vs. model-selected, where the human-selected refers to the features chosen by the domain experts and the model-selected refers to those chosen by the ML model. The model-selected applied the EMR data records without summarizing or emphasizing any indicator so that the ML model could extract the correlation among the original data.

Table 12 summarizes these two feature sets, where the human-selected feature set was adopted by the hospital that had implemented and practiced a comprehensive care suggestion system since 2017 [43]. Except the original EMR data, they added two more health indicators (EWS and health score) calculated by patient's physiological indicators. The experts intended to emphasize the importance of the health indicators by adding these two more features and assigning a human-defined weight for each of the health indicators in order to quantify the degree of his/her need for comprehensive care. Tuning the weighting has taken a large amount of trial-and-error experiments and discussion meetings in the studied hospital.

Table 12. The model-selected and the human-selected feature sets.

Model-Selected	Human-Selected	
Hospitalization date	Hospitalization date	
Coma Status	Coma Status	Early warning score (EWS)
Comorbidity	Comorbidity	
Respiration	Respiration	
SpO2	SpO2	
Systolic blood pressure	Systolic blood pressure	
Pulse	Pulse	
Temperature	Temperature	
Age Cluster		
Disease Cluster		
	Age	
Nutrition		Health Score
BSRS		
Pain		

The health score evaluates a patient's total health condition from the following three aspects: the patient's wellness evaluation from the care team (TeamEval), the comorbidity condition based on the Charlson comorbidity index (CCI) [44], and the early warning Score (EWS) [45], as summarized in Table 13, where pain and BSRS ranged from 0 to 10, nutrition ranged 0 to 6, CCI scaled from 1 to 6, and EWS scaled from 0 to 3. Each index in Table 13 contributes a ratio tuned by the experts, where B stands for the measured CCI value and C for the EWS value. Different diseases contribute different mortality risks to a patient, so weighting was given for the above three wellness aspects as listed in Table 14. The health score is defined below.

$$HealthScore = TeamEval \times W_{TeamEval} + CCI \times W_{CCI} + EWS \times W_{EWS} \tag{10}$$

where

$$TeamEval = Pain \times R_{Pain} + BSRS \times R_{BSRS} + Nutrition \times R_{Nutrition} \tag{11}$$

Table 13. Health condition attributes.

TeamEval		CCI		EWS	
Index	**Ratio**	**Index**	**Ratio**	**Index**	**Ratio**
Pain	$R_{Pain} = 5$	Min (B, 20)	$\mathrm{R}_{\mathrm{CCI}} = 5$	Min (C, 20)	$\mathrm{R}_{\mathrm{EWS}} = 5$
Brief Symptom Rating Scale (BSRS)	$R_{BSRS} = 5$				
Nutrition	$R_{Nutrition} = 30$				

Table 14. Disease weighting in health score.

Disease	$\mathbf{W_{TeamEval}}$	$\mathbf{W_{CCI}}$	$\mathbf{W_{EWS}}$
Cerebrovascular accident (CVA)	0.37	0.34	0.29
Dementia	0.7	0.19	0.10
Cancer	0.61	0.19	0.20
Others	0.10	0.80	0.10

Table 15 outlines the experimental results. Overall, the model-selected feature set yielded better performance. It required a longer training time than the human-selected one as its feature dimension was larger and required more time to find an optimal weighting for the features.

Table 15. The results of Exp. 4 evaluating two feature selection approaches.

Feature Set	**Human-Selected**		**Model-Selected**	
Actual **Predict**	**Positive**	**Negative**	**Positive**	**Negative**
Positive	397	202	393	167
Negative	77	125,331	81	125,366
	Mean	CI (95%)	Mean	CI (95%)
Accuracy (ACC)	99.78%	[99.76%, 99.80%]	**99.80%**	**[99.78%, 99.82%]**
Precision (PPV)	66.28%	[66.22%, 66.34%]	**70.18%**	**[70.14%, 70.22%**
NPV	99.84%	[99.83%, 99.85%]	**99.87%**	**[99.85%, 99.89%]**
Recall	**83.76%**	**[83.71%, 83.81%]**	82.91%	[82.82%, 83.00%]
F1	0.73998	[0.7319, 0.7480]	**0.76015**	**[0.7534, 0.7670]**
Training (min)	**66**	**[65.40, 66.60]**	104	[102.08, 103.92]

4.7. Exp. 5: Model Selection

Exp. 5 compared the performance of different prediction models including the proposed LSTM models, the baseline RNN, and a hybrid CNN+LSTM model, where the baseline RNN was considered as the previous model applied in the studied hospital, and the hybrid model applied CNN to perform feature selection and LSTM for prediction. The hyperparameters applied for the baseline RNN and hybrid CNN+LSTM models are outlined in Table 16; the model comparison results are listed in Table 17.

Table 16. The hyperparameters of the comparison models used in Exp. 5.

RNN Parameters	
Activation	Sigmoid
Kernel	Glorot_uniform
Recurrent	orthogonal
Bias vector	zeros
Return sequence	True
Return state	True
Loss	Binary Cross-Entropy
CNN + LSTM Parameters	
Convolution_2D	
Filter	64
Filter length	7
Activation	ReLU
MaxPooling_2D	
Pool length	5
LSTM Parameters	
Activation	Sigmoid
Loss	Binary Cross-Entropy

Table 17. The performance comparison of different neural networks.

	LSTM (Proposed)		**RNN (Baseline)**		**CNN+LSTM (Hybrid)**	
Actual **Predict**	**Positive**	**Negative**	**Positive**	**Negative**	**Positive**	**Negative**
Positive	393	167	385	174	404	201
Negative	81	125,366	89	125,359	70	125,332
	Mean	CI (95%)	Mean	CI (95%)	Mean	CI (95%)
Accuracy (ACC)	**99.80%**	**[99.78%, 99.82%]**	99.79%	[99.78%, 99.80%]	99.78%	[99.77%, 99.79]
Precision (PPV)	**70.18%**	**[70.14%, 70.22%**	68.87%	[68.76%, 68.98%]	66.78%	[66.65%, 66.91%]
NPV	**99.87%**	**[99.85%, 99.89%]**	99.86%	[99.85%, 99.87%]	99.84%	[99.84%, 99.84%]
Recall	82.91%	[82.82%, 83.00%]	81.22%	[81.11%, 81.33%]	**85.23%**	**[85.16%, 85.30%]**
F1	**0.7602**	**[0.7534, 0.7670]**	0.7454	[0.7392, 0.7516]	0.7488	[0.7438, 0.7538]
Training (min)	103	[102.08, 103.92]	**73**	[72.39, 73.61]	267	[260.09, 273.91]

All models performed well on accuracy and NPV as the data set contained a high number of negatives. Both models, the baseline and the hybrid, performed equally well on all the performance measurements, but the hybrid required 2.6 times more training time than the baseline. Overall, the proposed LSTM achieved the best performance and required a little bit more time than the baseline. It could forecast comprehensive care patients efficiently with a high true positive rate and achieved the purpose well.

5. Discussion

The model-selected feature set exhibited better overall performance and yielded a better F1 score, but the human-selected one had a slightly better recall rate. As the correctness of the prediction was based on the clinician's decision, they tended to activate the comprehensive care service by observing the values of the two human-defined features, the health score and the EWS. Therefore, the human-selected feature set may predict positive cases better because of the two additional features. Likewise, the patient's age and comorbidity status are two important factors for the clinician while making the decision. Therefore, the patient clustering approach according to the above two factors resulted in a better recall rate as well.

When considering the overall performance by F1 score, the LSTM model outperformed the baseline as well as the hybrid model. However, the recall rate may be important for some medical applications. Even though the hybrid model (CNN+LSTM) took two times more of the training time than the LSTM model and three times more than the RNN model, it had the best recall rate, which could offset the additional training time.

An efficient clinical decision support system assists the clinicians in predicting the patients' needs, improves the quality of healthcare service, and preserves medical resources. The proposed prediction model was designed according to the clinician's practical experiences on comprehensive care and improved the prediction performance by resampling, patient clustering, and feature selection. The system evaluation demonstrated that the proposed method is practical and efficient. It is able to improve work efficiency, and it reduces the consumed medical resources, produces less medical waste, and hence improves the environmental sustainability of the medical institute.

6. Conclusions

Good-quality care for patients with comorbidities is one of the most important and complex problems of modern clinical health care because of the high mortality rates if not managed properly. By applying AI technologies, this study developed an LSTM-based prediction system, learning from the past patients' medical records and decisions made by clinicians so that it provides comprehensive care suggestions. The proposed system can enhance the existing hospital information system with the ability to monitor the health condition of inpatients continuously and to fill the communication gap among multiple clinic staff taking care of them. By allocating healthcare resources properly based on need, the proposed system can improve staff work efficiency as well as increase medical service quality.

To our best knowledge, the present study is the first attempt to apply an RNN model to develop a CDSS that predicts personalized care needs. Only a few published studies focused on applying RNN for predicting patient care, and most discussed the impacts and needs of applying AI technologies to the healthcare sector.

This study addressed several design issues when applying AI models, including the type of classifier for a multi-class question, feature selection, imbalanced data, and model selection. The evaluation section conducted experiments to analyze the impact of the above issues. A multi-classifier performed better than multiple bi-classifiers. Human-selected features might not be able to perform as well as features selected by the machine learning model. Resampling is an effective strategy to mitigate the impact of imbalanced training data, where under-sampling is better if the data are extremely imbalanced, as is in this study. A prediction model combining two neural network models typically consumes more training time than a single neural network model but might not improve the performance significantly. This study demonstrated that patient clustering improves personalized care prediction and proposed an efficient CDSS approach.

It is difficult to account for the variability of the disease risks to a new illness, such as COVID-19. Investigations require domain knowledge and complex production rules where patterns are hard to predict. Further work could be focused on identifying the key factors of providing personalized comprehensive care for patients with different disease complications and mortality risks.

Author Contributions: Conceptualization, M.-H.H. and W.-C.J.; methodology, C.-M.C. and Z.-X.C.; software, Z.-X.C.; validation, M.-H.H., W.-C.J., C.-M.C. and T.K.; data collection, M.-H.H. and W.-C.J.; writing—original draft preparation, C.-M.C. and Z.-X.C.; writing—review and editing, M.-H.H. and W.-C.J.; visualization, Z.-X.C.; supervision, W.-C.J., C.-M.C. and T.K.; project administration, W.-C.J.; All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by Kaohsiung Veterans General Hospital with project number KSVNSU110-012.

Institutional Review Board Statement: The study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of Kaohsiung Veterans General Hospital with IRB certification number KSVGH20-CT10-08 (date of approval: 1 December 2020).

Informed Consent Statement: Not applicable.

Data Availability Statement: No data sharing agreement for the data was formulated.

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

Ethics Approval: This research was conducted with anonymized and de-identified patient records and has received ethical approval from the IRB of the designated hospital with the IRB certification number KSVGH20-CT10-08.

References

1. Shediac-Rizkallah, M.C.; Bone, L.R. Planning for the sustainability of community-based health programs: Conceptual frameworks and future directions for research, practice and policy. *Health Educ. Res.* **1998**, *13*, 87–108. [CrossRef] [PubMed]
2. The National Health Services. Improvement Leaders' Guide. Available online: https://www.england.nhs.uk/improvement-hub/wp-content/uploads/sites/44/2017/11/ILG-1.7-Sustainability-and-its-Relationship-with-Spread-and-Adoption.pdf (accessed on 3 September 2021).
3. Valizadeh, L.; Jasemi, M.; Zamanzadeh, V.; Keogh, B. A concept analysis of holistic care by hybrid model. *Indian J. Palliat. Care* **2017**, *23*, 71–80. [CrossRef] [PubMed]
4. Steiger, W.A.; Hoffman, F.H.; Hansen, A.V.; Niebuhr, H. A Definition of Comprehensive Medicine. *J. Heal. Hum. Behav.* **1960**, *1*, 83. [CrossRef]

5. Dilmegani, C. Personalized Care in 2021: In-Depth Guide. Available online: https://research.aimultiple.com/personalized-care/#:~:text=Personalized%20healthcare%20is%20the%20process,them%20personalized%20care%20and%20treatments. (accessed on 2 September 2021).
6. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef]
7. Alliance for Natural Health International. Create Health. Naturally. Available online: https://www.anhinternational.org/ (accessed on 3 September 2021).
8. Adams, K. Sustainable Healthcare. Available online: https://www.resilience.org/stories/2012-08-03/sustainable-healthcare/ (accessed on 3 September 2021).
9. Cadar, R.-L.; Amuza, A.; Dumitras, D.; Mihai, M.; Pocol, C. Analysing Clusters of Consumers Who Use Medicinal and Aromatic Plant Products. *Sustainability* **2021**, *13*, 8648. [CrossRef]
10. Antoniadi, A.M.; Mooney, C. The Multi-Dimensional Process of Developing a Clinical Decision Support System Using Machine Learning. Available online: https://womencourage.acm.org/2019/wp-content/uploads/2019/07/womENcourage_2019_paper_43.pdf (accessed on 3 September 2021).
11. Kotsiantis, S.; Kanellopoulos, D.; Pintelas, P. Data preprocessing for supervised leaning. *Int. J. Comput. Sci.* **2006**, *1*, 111–117.
12. Datta, S.K.; Shaikh, M.A.; Srihari, S.N.; Gao, M. Soft Attention Improves Skin Cancer Classification Performance. *arXiv* **2021**, arXiv:2105.03358.
13. Dawoodbhoy, F.M.; Delaney, J.; Cecula, P.; Yu, J.; Peacock, I.; Tan, J.; Cox, B. AI in patient flow: Applications of artificial intelligence to improve patient flow in NHS acute mental health inpatient units. *Heliyon* **2021**, *7*, e06993. [CrossRef] [PubMed]
14. Hashi, E.K.; Zaman, M.S.U. Developing a hyperparameter tuning based machine learning approach of heart disease prediction. *J. Appl. Sci. Process Eng.* **2020**, *7*, 631–647. [CrossRef]
15. Nawaz, M.S.; Shoaib, B.; Ashraf, M.A. Intelligent Cardiovascular Disease Prediction Empowered with Gradient Descent Optimization. *Heliyon* **2021**, *7*, e06948. [CrossRef]
16. Xu, Y.; Lam, H.-K.; Jia, G. MANet: A two-stage deep learning method for classification of COVID-19 from Chest X-ray images. *Neurocomputing* **2021**, *443*, 96–105. [CrossRef]
17. Rohila, V.S.; Gupta, N.; Kaul, A.; Sharma, D.K. Deep learning assisted COVID-19 detection using full CT-scans. *Internet Things* **2021**, *14*, 100377. [CrossRef]
18. Kassania, S.H.; Kassanib, P.H.; Wesolowskic, M.J.; Schneidera, K.A.; Detersa, R. Automatic Detection of Coronavirus Disease (COVID-19) in X-ray and CT Images: A Machine Learning Based Approach. *Biocybern. Biomed. Eng.* **2021**, *41*, 867–879. [CrossRef]
19. Ibrahim, D.M.; Elshennawy, N.M.; Sarhan, A.M. Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Comput. Biol. Med.* **2021**, *132*, 104348. [CrossRef] [PubMed]
20. Chen, C.; Zhou, K.; Zha, M.; Qu, X.; Guo, X.; Chen, H.; Wang, Z.; Xiao, R. An Effective Deep Neural Network for Lung Lesions Segmentation from COVID-19 CT Images. *IEEE Trans. Ind. Inform.* **2021**, *17*, 6528–6538. [CrossRef]
21. Vidal, P.L.; de Moura, J.; Novo, J.; Ortega, M. Multi-stage transfer learning for lung segmentation using portable X-ray devices for patients with COVID-19. *Expert Syst. Appl.* **2021**, *173*, 114677. [CrossRef] [PubMed]
22. Khreis, N.; Lau, A.S.M.; Al-Jedai, A.; Al-Khani, S.M.; Alruwaili, E.H. An Evaluation of Clinical Decision Support and Use of Machine Learning to Reduce Alert Fatigue. *Int. J. Comput. Commun. Eng.* **2019**, *8*, 32–39. [CrossRef]
23. Baker, S.; Xiang, W.; Atkinson, I. A hybrid neural network for continuous and non-invasive estimation of blood pressure from raw electrocardiogram and photoplethysmogram waveforms. *Comput. Methods Programs Biomed.* **2021**, *207*, 106191. [CrossRef] [PubMed]
24. Minnema, J.; Wolff, J.; Koivisto, J.; Lucka, F.; Batenburg, K.J.; Forouzanfar, T.; van Eijnatten, M. Comparison of convolutional neural network training strategies for cone-beam CT image segmentation. *Comput. Methods Programs Biomed.* **2021**, *207*, 106192. [CrossRef] [PubMed]
25. Lakshmanaprabu, S.; Mohanty, S.N.; Shankar, K.; Arunkumar, N.; Ramirez, G. Optimal deep learning model for classification of lung cancer on CT images. *Future Gener. Comput. Syst.* **2019**, *92*, 374–382.
26. Liu, C.; Cao, Y.; Alcantara, M.; Liu, B.; Brunette, M.; Peinado, J.; Curioso, W. TX-CNN: Detecting tuberculosis in chest X-ray images using convolutional neural network. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 2314–2318.
27. Dong, Y.; Pan, Y.; Zhang, J.; Xu, W. Learning to Read Chest X-Ray Images from 16000+ Examples Using CNN. In Proceedings of the 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Philadelphia, PA, USA, 17–19 July 2017; pp. 51–57.
28. Wulff, A.; Montag, S.; Marschollek, M.; Jack, T. Clinical Decision-Support Systems for Detection of Systemic Inflammatory Response Syndrome, Sepsis, and Septic Shock in Critically Ill Patients: A Systematic Review. *Methods Inf. Med.* **2019**, *58*, e43–e57. [CrossRef] [PubMed]
29. Vollmer, S.; Mateen, B.A.; Bohner, G.; Király, F.J.; Ghani, R.; Jonsson, P.; Cumbers, S.; Jonas, A.; McAllister, K.S.L.; Myles, P.; et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ* **2020**, *368*, l6927. [CrossRef]

30. Wood, T. What Is the Softmax Function? Available online: https://deepai.org/machine-learning-glossary-and-terms/softmax-layer (accessed on 3 September 2021).
31. Wikipedia. Softmax Function. Available online: https://en.wikipedia.org/wiki/Softmax_function (accessed on 3 September 2021).
32. Farzad, A.; Mashayekhi, H.; Hassanpour, H. A comparative performance analysis of different activation functions in LSTM networks for classification. *Neural Comput. Appl.* **2019**, *31*, 2507–2521. [CrossRef]
33. Nielsen, M. Improving the Way Neural Networks Learn. Available online: http://neuralnetworksanddeeplearning.com/chap3.html (accessed on 19 December 2020).
34. Brownlee, J. How to Choose Loss Functions When Training Deep Learning Neural Networks. Available online: https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/ (accessed on 18 December 2020).
35. Ruder, S. An overview of gradient descent optimization algorithms. *arXiv* **2016**, arXiv:1609.04747.
36. Keras. Optimizers. Available online: https://keras.io/zh/optimizers/#rmsprop (accessed on 19 December 2020).
37. Brownlee, J. How to Control the Stability of Training Neural Networks with the Batch Size. Available online: https://machinelearningmastery.com/how-to-control-the-speed-and-stability-of-training-neural-networks-with-gradient-descent-batch-size/ (accessed on 26 July 2021).
38. He, F.; Liu, T.; Tao, D. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 1143–1152.
39. Keskar, N.S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; Tang, P.T.P. On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. Available online: https://arxiv.org/abs/1609.04836 (accessed on 3 September 2021).
40. Smith, S.L.; Kindermans, P.-J.; Ying, C.; Le, Q.V. Don't Decay the Learning Rate, Increase the Batch Size. Available online: https://arxiv.org/abs/1711.00489 (accessed on 3 September 2021).
41. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
42. Yen, S.-J.; Lee, Y.-S. Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Intelligent Control and Automation. Lecture Notes in Control and Information Sciences*; Huang, D.S., Li, K., Irwin, G.W., Eds.; Springer: Berlin/Heidelberg, Germany, 2006; Volume 344, pp. 731–740. [CrossRef]
43. Hung, P.; Chen, J.; Chen, M.; Li, P.; Li, W.; Wang, Z.; Huang, T.; Liang, Y.; Lin, P. The impact of a medication reconciliation programme at geriatric hospital admission: A pre-/postintervention study. *Br. J. Clin. Pharmacol.* **2019**, *85*, 2614–2622. [CrossRef]
44. Tominaga, T.; Nonaka, T.; Takeshita, H.; Kunizaki, M.; Sumida, Y.; Hidaka, S.; Sawai, T.; Nagayasu, T. The Charlson Comorbidity Index as an Independent Prognostic Factor in Older Colorectal Cancer Patients. *Indian J. Surg.* **2016**, *80*, 54–60. [CrossRef]
45. Albur, M.; Hamilton, F.; MacGowan, A.P. Early warning score: A dynamic marker of severity and prognosis in patients with Gram-negative bacteraemia and sepsis. *Ann. Clin. Microbiol. Antimicrob.* **2016**, *15*, 23. [CrossRef] [PubMed]