future internet

MDPI

Article

Hospital Readmission and Length-of-Stay Prediction Using an Optimized Hybrid Deep Model

Alireza Tavakolian [1], Alireza Rezaee [1], Farshid Hajati [2,*] and Shahadat Uddin [3]

[1] Department of Mechatronics Engineering, Faculty of New Sciences and Technologies, University of Tehran, Tehran 14174, Iran; alireza.tavakol@ut.ac.ir (A.T.); arrezaee@ut.ac.ir (A.R.)
[2] Intelligent Technology Innovation Laboratory (ITIL) Group, Institute for Sustainable Industries and Liveable Cities, Victoria University, Footscray, VIC 3011, Australia
[3] School of Project Management, Faculty of Engineering, The University of Sydney, Sydney, NSW 2006, Australia; shahadat.uddin@sydney.edu.au
* Correspondence: farshid.hajati@vu.edu.au

Abstract: Hospital readmission and length-of-stay predictions provide information on how to manage hospital bed capacity and the number of required staff, especially during pandemics. We present a hybrid deep model called the Genetic Algorithm-Optimized Convolutional Neural Network (GAOCNN), with a unique preprocessing method to predict hospital readmission and the length of stay required for patients of various conditions. GAOCNN uses one-dimensional convolutional layers to predict hospital readmission and the length of stay. The parameters of the layers are optimized via a genetic algorithm. To show the performance of the proposed model in patients with various conditions, we evaluate the model under three healthcare datasets: the Diabetes 130-US hospitals dataset, the COVID-19 dataset, and the MIMIC-III dataset. The diabetes 130-US hospitals dataset has information on both readmission and the length of stay, while the COVID-19 and MIMIC-III datasets just include information on the length of stay. Experimental results show that the proposed model's accuracy for hospital readmission was 97.2% for diabetic patients. Furthermore, the accuracy of the length-of-stay prediction was 89%, 99.4%, and 94.1% for the diabetic, COVID-19, and ICU patients, respectively. These results confirm the superiority of the proposed model compared to existing methods. Our findings offer a platform for managing the healthcare funds and resources for patients with various diseases.

Keywords: readmission; length of stay; convolutional neural networks; genetic algorithm; diabetes; COVID-19

Citation: Tavakolian, A.; Rezaee, A.; Hajati, F.; Uddin, S. Hospital Readmission and Length-of-Stay Prediction Using an Optimized Hybrid Deep Model. *Future Internet* **2023**, *15*, 304. https://doi.org/10.3390/fi15090304

Academic Editor: Hamid Mcheick

Received: 2 July 2023
Revised: 6 August 2023
Accepted: 10 August 2023
Published: 6 September 2023

1. Introduction

Hospital readmission and length of stay (LOS) play major roles in hospital expenditures. Recently, healthcare systems' main focus is on patients being readmitted to hospitals within a short time frame (mostly on readmissions that occur within 30 days) after discharge [1]. According to the latest report, the United States healthcare system's burden was 41 billion dollars, due to the hospital readmissions of diabetic patients within 30 days [2]. A study in Spain revealed that, while the total annual cost of diabetic patients was EUR 1803.6 per person, the cost of hospitalization for these patients was EUR 801.6 [3]. Another study conducted in the United States showed that the direct annual cost of diabetes is about USD 9595 per person [4]. There are direct and indirect costs of healthcare systems that are related to inpatient hospitalization. Direct medical costs include the costs associated with the services provided at the hospital, such as inpatient stays, intensive care unit (ICU) stays, laboratory tests, and other types of hospital visits. For various diseases, the hospitalization share of the total cost is different. For diabetic patients, 35% of the total cost is considered for hospitalization [5]. The share for swine flu is 40%. The hospitalization cost for COVID-19 patients varies based on age [6].

On average, 92.6% of the total cost for COVID-19 patients is for hospitalization [7]. These facts indicate that readmission time and LOS are responsible for more than 50% of the total cost to patients. Besides the cost to the patients and healthcare systems, long LOS and repeated readmissions also lead to other problems. An increase in LOS downgrades the quality of healthcare services due to an increase in the patients-to-nurses ratio. During the COVID-19 pandemic, it was reported that for every extra patient per nurse, a 7% increase was incurred in the odds of patient failure-to-rescue rates, as well as a 7% increase in the likelihood of dying within 30 days [8]. Recently, with the emergence of COVID-19, the need for hospital beds has increased. The LOS for COVID-19 patients varies based on the level of severity and age group; the LOS of COVID-19 patients increases with age for patients older than 60. However, the LOS for COVID-19 patients in ICUs decreases for people aged 80 years or older due to a higher mortality rate [9]. The hospital readmission of diabetic patients within 30 days increases the risk of contracting COVID-19 [10]. Diabetic patients have a risk factor for hospitalization and a high mortality rate due to COVID-19. According to recent research in China, the COVID-19 mortality rate in diabetes patients is about three-fold more elevated than the general patient mortality rate [10]. Thus, a precise prediction of readmissions and LOS help the healthcare system to manage the availability of hospital beds and quality of service. Predicting LOS and readmission time frames has been investigated by other researchers in recent years. The main focus of previously published research was to use basic machine learning (ML) methods as a simple classifier for predicting the LOS and readmission time frames. Based on reviewed articles in the related work section, the main gap in previous research is the lack of evaluating new hybrid methods and in investigating the possibility of using new feature sets for LOS and readmission time frame predictions.

In this research, we propose a hybrid model, which is achieved by a combination of deep learning (DL) and evolutionary algorithms under the name of the Genetic Algorithm-Optimized Convolutional Neural Network (GAOCNN). The proposed algorithms are evaluated by three different datasets to predict the readmission time frame for diabetic patients and the LOS for diabetic, COVID-19, and ICU patients. Experimental results indicate that the GAOCNN estimates readmissions with a 97.2% accuracy. Furthermore, the accuracy of the GAOCNN for the LOS prediction is 89%, 99.4%, and 94.1% for diabetic, COVID-19, and ICU patients, respectively.

The proposed method is organized into eight main sections. Section 1 describes a summary of the research and its contribution. Subsequently, the second part presents an overview of the historical background related to LOS and readmission time frame predictions, as well as the most recent solutions. Next, Section 3 provides an outline of the dataset's specifications, along with its associated statistical characteristics. The proposed method is explained in this section as well. Section 4 describes the evaluation process employed to assess the proposed method's effectiveness and the verification of the results. In Section 5, the presented methodology is summarized, and a comparison is drawn between the proposed method and similar research. Section 6 specifies the limitations and future directions of the work. Section 7 delivers a concise conclusion that encompasses the research method, results, contributions, and future prospects of the study. The final section of the study defines the abbreviations. The contributions of this research are as follows:

- A novel cost function for the genetic optimization process to leverage the feature extraction process toward a better performance is presented;
- The proposed approach outperformed both ML and DL methods for short- and long-term LOS predictions;
- The proposed approach outperformed the surveyed ML methods in related works with respect to diabetic readmission time predictions;
- The most important features for both the LOS and readmission time frame predictions are provided.

2. Related Works

Numerous models have been developed to predict patient conditions in medical facilities [11]. Recent studies have focused on utilizing ML techniques for readmission prediction [12].

Forsman and Jonsson used k-nearest neighbour, logistic regression (LR), boosted decision trees [13], and artificial neural networks [14] for readmission prediction. Their purpose was to classify patients into two groups: patients who never returned to the hospital and patients who returned within 30 days. The best result for this research was an 80.1% accuracy with the LR model. Alloghani et al. [15] applied ML to diabetes data to recognize patterns and combinations of factors that characterize the readmission of diabetes patients. They used a range of classifiers, including linear discriminant analysis [16], random forest (RF), k-nearest neighbour, naive Bayesian, decision tree, and support vector machine (SVM) [17]. Using the naive Bayesian algorithm, their best result was the area under the receiver operating characteristic curve (AUROC) and a precision of 64% and 51%, respectively. Hammoudeh et al. [18] presented a convolutional neural network CNN model as a binary classifier to predict readmission. They aimed to distinguish between patients who returned to the hospital and those who did not return. They reported accuracy and AUROC of 80% and 85%, respectively. Mingle [19] used ML classifiers such as RF, extreme gradient boosting (XGB), balanced RF, gradient boosted trees, gradient boosted greedy trees, extreme gradient boosted trees, extreme gradient boosted classifier [20] and Nesterov kernel SVM [21] with a range of encoding procedures. The best accuracy for classifying patients as either never having returned to the hospital or having returned within 30 days was 78%. Arnaud et al. [12]. evaluated various DL models to predict the readmission time frame of emergency admitted patients. The authors evaluated a combination of MLP and CNN models to extract information from numerical and contextual features. The authors reported 0.83 AUROC for the readmission time frame prediction.

Morton et al. [22] tested supervised ML algorithms such as SVM and RF for predicting short-term stays (less than three days) at hospitals for diabetic patients. They worked on a three-class classification, and reported a 68% accuracy with 1% tolerance, which they achieved with SVM+. Yakovlev et al. [23] used a multi-layer perceptron (MLP) to predict the hospital LOS for coronary syndrome patients. They used 6000 samples, divided into 5000 training samples and 1000 testing samples. The predicted LOS's average and standard deviation were 15 and 9.5 days, respectively. Tsai et al. [24] proposed an ML algorithm for hospital management by predicting the LOS before patients' admission. They developed DL models to predict the LOS for patients with one of three primary diagnoses: coronary atherosclerosis, heart failure, and acute myocardial infarction in a cardiovascular unit. They reported a 67% accuracy with a 2-day tolerance. Wang et al. [25] evaluated various ML models such as SVM, RF, and long short-term memory (LSTM) [26] for ICU LOS prediction. The authors used features such as age, gender, and admission type of the patient for predicting LOS. The authors proposed a pipeline for preprocessing, encoding, and training the final model. The target classes for this prediction were LOS > 4 or LOS > 7 days. The authors reported 84.3% accuracy using LSTM for emergency LOS prediction. Nallabasannagar et al. [27] presented an MLP model for ICU LOS prediction. The authors used features such as age, gender, and lab events as the feature set. The authors used a customized embedding method for converting contextual information to numerical values. The authors reported 66.2% accuracy for discriminating LOS in two classes of >7 and <7.

COVID-19 can lead to pneumonia and long-term LOS for patients with underlying diseases [28]. Thus, researchers have focused on predicting long-term LOS to manage the hospital beds [29]. Manhub et al. [30] used a decision tree to predict the COVID-19 patients' LOS. They analysed 2017 patients from January to July 2020. Their work results indicate an R2-score of 49.8% and a median absolute deviation of 2.85 days. For predicting of discharge time in COVID-19 patients, Nemati et al. [31] used the health records of 1182 patients. They used only age and gender as input features for the discharge time prediction of COVID-19

patients. They tested the gradient boost algorithm, Cox regression, and fast SVM for the discharge time prediction. They reported that the gradient boost algorithm achieved the best result, with an accuracy of 71.7%. A summary of the reviewed research is shown in Table 1.

Table 1. Summary of the reviewed research.

Authors	Dataset	Model	Accuracy (%)	Strengths/Weakness
Alloghani et al. [15] (2019)	Diabetes	naive Bayesian	65	-/Weak performance, relatively old method.
Hammoudeh et al. [18] (2018)	Diabetes	CNN	80	High performance/Only two classes of readmission were predicted.
Mingle et al. [19] (2017)	Diabetes	Gradient boosted trees	78	-/Poor performance
Morton et al. [22] (2014)	Diabetes	SVM+	68	-/Poor performance for discriminating between short and long term LOS, relatively old method.
Mahboub et al. [30] (2021)	COVID-19	Decision Tree	50	-/Poor performance, relatively old method.
Mahboub et al. [30] (2021)	COVID-19	gradient boost algorithm	72	-/Poor performance, No hyperparameter tuning was performed.
Wang et al. [25] (2020)	ICU	LSTM	84	High performance/No hyperparameter tuning was performed.
Nallabasannagari et al. [27] (2020)	ICU	MLP	66	-/Weak performance, No augmentation method was used.

None of the above-mentioned methods has taken any action to predict patients' long-term hospitalisation LOS. All existing LOS classifications are restricted to three or fewer classes. Furthermore, the existing models have a low performance for classifying readmitted patients into more than two categories. Most of the reviewed works have focused on using standard ML models for LOS prediction and their performance has been reported on a single disease only. To overcome the limitations of previous works, we propose a method to predict both the readmission and the LOS in patients with various conditions using a novel hybrid deep model (GAOCNN).

In the GAOCNN, the CNN predicts the hospital readmission and the LOS, while the genetic algorithm (GA) optimizes the parameters of the layers to improve the performance. The proposed model is evaluated using three datasets of diabetic, COVID-19, and ICU patients. To compare GAOCNN performance with other artificial intelligence techniques, we used a traditional ML model such as SVM or a traditional DL model such as Visual Geometry Group (VGG16) [32]. Furthermore, we combine traditional ML with DL models such as CNN+SVM to ensure the capability of GAOCNN compared to the hybrid model. The experimental results indicate superior performance to ML, DL, and hybrid models. Compared to similar research, proposed algorithms can also help predict LOS with a lower time frame. Lower time frame length leads to a better knowledge of the number of patients each day, and this knowledge can help the hospital manage nurse scheduling programs better, especially during a pandemic.

3. Methodology

In this section, we first provided an overview of the dataset, highlighting its characteristics and conducting a statistical analysis. Next, we presented a detailed explanation of the proposed algorithm, which comprises three distinct phases: feature extraction, op-

timization, and the final classifier. Within each phase, we delved into the description of the hyperparameters for the convolutional layers, the genetic optimizer objective function, and the final discriminator.

3.1. Dataset

To show the performance of the proposed model in patients with various conditions, we evaluate the proposed model using datasets of diabetic, COVID-19, and ICU patients. The dataset we have used for diabetic patients has information on readmission and the LOS, while the other two datasets include information on the LOS. The details of each dataset are explained in the following sections.

3.1.1. Diabetes

For diabetic patients, we use a dataset of 130 hospitals in the United States from 1999 to 2008 [33]. The dataset consists of 101,766 records with 50 attributes, such as ethnicity, gender, age, weight, and hospital visits. The data also contains features such as patient identification number, admission type, hospital LOS, the speciality of the admitting physician, the number of performed lab tests, glycated haemoglobin (HbA1c) test results, diagnoses, the number of medications, diabetic medications, the number of inpatients and outpatients, and the number of emergency visits in the year before the hospitalization. Weight and age are recorded in and 25-pound and 10-year intervals, respectively. Gender reported as male, female, or unknown. The percentage of patients with male, female, and unknown gender is 53.77%, 46.22%, and 0.01%, respectively.

The hospital inpatient and outpatient visits within the year before the hospitalization have been recorded in the dataset. The speciality of the admitting physician had been recorded as 84 distinct values such as cardiology, internal medicine, family or general practice, and surgeon. In the dataset, the range of the glucose serum test result had been recorded as “normal”, “more than 200”, “more than 300”, or “not measured”. The primary, secondary, and additional secondary diagnoses have been recorded in the international-statistical classification of diseases (ICD) codes [34]. The primary, secondary, and additional secondary diagnosis attributes encoded as the first three digits of the ICD-9 [34], having 848, 923, and 954 distinct values, respectively. More than 44% of the primary diagnoses in the dataset are related to circulatory and respiratory system diseases.

3.1.2. COVID-19

We have gathered the medical records of 1085 COVID-19 patients from January to February 2020 from a publicly available COVID-19 dataset [35]. The dataset includes symptom-onset, hospital visit date, exposure date, recovered date, and death. Personal information about age, gender, location of hospitalization (country/state), and travel history from Wuhan are reported. The most significant information is the date of exposure to the public and the date before the critical condition. The LOS has not been reported in the dataset directly, but it can be extracted using the difference between the hospital visit and discharge or death. Most patients lived in China, Southeast Asia, and the United States.

3.1.3. ICU

ICU patients’ information is extracted from the MIMIC-III clinical dataset [36]. This dataset consists of 58,976 patients, 42,071 of whom were admitted to the hospital with an emergency condition. The dataset was collected from the Beth Israel Deaconess Medical Center between 2001 and 2012. The dataset consists of personal characteristics such as sex, age, ethnicity, and detailed admit information for each patient, including type and location of admission. Other information, such as the number of lab procedures, the number of transformations between hospitals, and the LOS, is reported in this dataset. Detailed information for all of the used datasets is shown in Table 2.

Table 2. Details information of the Diabetes, MIMIC-III, and COVID-19 datasets.

Dataset Name	Number of Instances	Number of Features	Gender	Collected Years	LOS Range
Diabates	101,766	50 (37 descriptive and 13 numerical features)	53.8% male, 46.2% female	1999–2008	(0–14]
COVID-19	1085	23 (17 descriptive and 6 numerical features)	64.7% male, 35.3% female	2020–2021	(0–30]
ICU	58,976	28 (9 descriptive and 19 numerical features)	59.1% male, 41.9% female	2001–2012	(0–294]

3.2. Gaocnn

We present a hybrid deep model called the Genetic Algorithm-Optimized Convolutional Neural Network (GAOCNN). In this model, the convolutional layers are used for feature extraction and the dense layers for classification, while the GA is applied to optimise the layers' parameters. The overall structure of the proposed model is shown in Figure 1. The GAOCNN has two convolutional layers. After the convolutional layers, there is a pooling layer that is specified as average pooling functions [37]. Furthermore, we use two fully connected layers with a dropout, which mitigates the risk of overfitting [38].

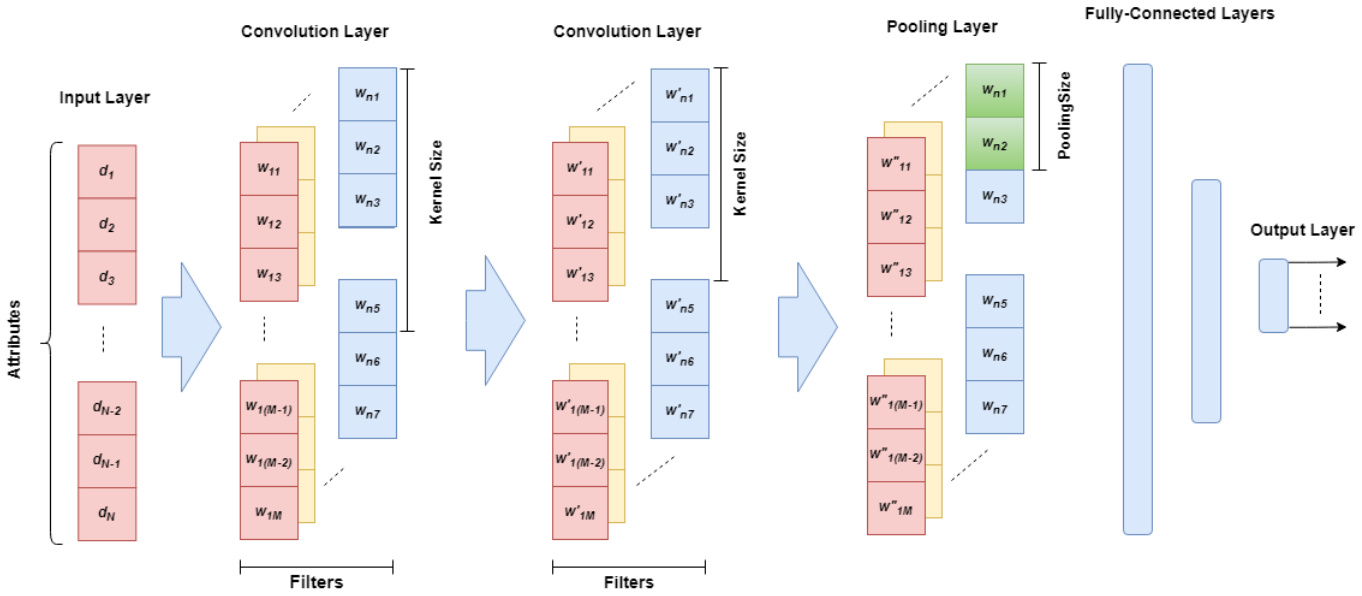


Figure 1. Structure of the proposed model.

The convolutional layers employ local connections and weights to extract features from input data and build dense feature vectors. Since the data is two-dimensional (samples, attributes), we apply one-dimensional convolutional layers. The main algorithm is a simple CNN model. The proposed method is trained for numerical, categorical, and descriptive features such as age, gender, and symptoms. One-dimensional convolutional layers can extract hidden relations between co-occurring symptoms and comorbidities. Furthermore, extracting time-dependent information from descriptive (contextual) features is conducted more efficiently using one-dimensional convolutional layers [39].

The main drawback of deep neural network models is their vast space of hyperparameters, which makes the parameter selection tedious. Most researchers use techniques such as random search [40] and grid search [41]. When we use these searching techniques, there is a trade-off between increasing the layers and the run time to reach a proper solution. To overcome this challenge, we use GA to select parameters scientifically. Furthermore, the whole process can be completed without human intervention. This automation in the learning process will help healthcare systems reach the proper performance without expert supervision.

The GA has been used widely in artificial intelligence fields, such as medical image processing, ML, and DL hybrid models [42]. The most essential elements of the GA are the environment and the fitness function. By defining a proper fitness function, we can guide the model to improve performance. The GA uses the process of selection, crossover, and mutation to choose the number of convolution kernels, the number of convolution filters, and the number of epochs and neurons of the model. The GA's standard steps are the initialization of the population, selection between the created population, logical combination (crossover), randomness (mutation), and decoding.

After making the first random generation of the population's data, according to the principle of 'survival of the fittest', only the fittest generation of the population survives. In each generation, individuals are selected according to their fitness. The surviving populations will be the parents of the next generations. In the proposed algorithm, we use a maximum filter to choose the best hyperparameters according to the highest fitness function in every step. The flowchart of the applied GA has shown in Figure 2. At the beginning of the training phase, the number of filters, convolution kernels, and neurons in each layer are randomly initialized. Then, the fitness function is calculated for the first generation. The next generations are created by conducting the crossover and possible mutation of the parents. The fitness values for new children are sorted in descending order, and the best of them are selected for the next generations. In the proposed algorithm, the model's loss decreases gradually, and the accuracy increases continuously. The proposed algorithms utilize a modified version of GA with two key differences from the original GA structure. Firstly, the objective function of the proposed method is customized to enhance both accuracy and loss. Secondly, the process of selecting the objective function is based on the calculated loss, leading to two different paths and corresponding loss functions for the optimization process.

The fitness function of the GA is defined as

$$F = \alpha \cdot a \cdot V_1 + \beta \left(\frac{1}{l + \varepsilon} \right) V_2 \tag{1}$$

where a and l are the accuracy and the loss, respectively, measured on the test set. α and β are two hyperparameters specified based on calculated loss (2). ε is a small value added to the denominator to avoid dividing by zero. V_1 and V_2 are defined below to select the best kernel size, the filter size, and the number of neurons and epochs.

$$\begin{aligned} V_1 \ = 0.3 \ * \ \left(\frac{N_F}{N_{FT}} \right) \ + \ 0.3 \ * \ \left(\frac{N_K}{N_{KT}} \right) \\ + 0.2 \ * \ \left(\frac{N_U}{N_{UT}} \right) + 0.2 * \left(\frac{N_E}{1.5} \right) \end{aligned} \tag{2}$$

and

$$\begin{aligned} V_2 \ = 0.3 \ * \ \left(1 - \frac{N_F}{N_{FT}} \right) \ + \ 0.3 \ * \ \left(1 - \frac{N_K}{N_{KT}} \right) \\ + 0.2 \ * \ \left(1 - \frac{N_U}{N_{UT}} \right) + 0.2 * \left(1 - \frac{N_E}{1.5} \right) \end{aligned} \tag{3}$$

where N_F is the number of convolution filters, N_{FT} is the total number of convolution filters, N_K is the number of convolution kernels. N_{KT} is the total number of convolution kernels, N_U is the number of neurons at the deep layer, N_{UT} is the total number of neurons at the deep layer, and N_E is the total number of epochs. Compared to traditional CNN models, the proposed methods can use dynamic kernel size to evaluate the quality of the extracted features for the final LOS and readmitted time frame prediction. Furthermore, the proposed method uses only two convolutional layers for feature extraction, and the depth of the proposed method is lower than traditional CNNs like VGG16. The search

environment for the kernel sizes are 3, 5, 7, and 9. The search environment for filter size is between 32 and 36 and 62 to 64 for the first and second convolutional layers. The search environment for the number of neurons is between 256 and 512 and 124 to 256 for the first and second dense layers. The search environment for the number of epochs is 500 to 1000. The proposed method can extract various combinations of feature sets screened by the fitness function values.

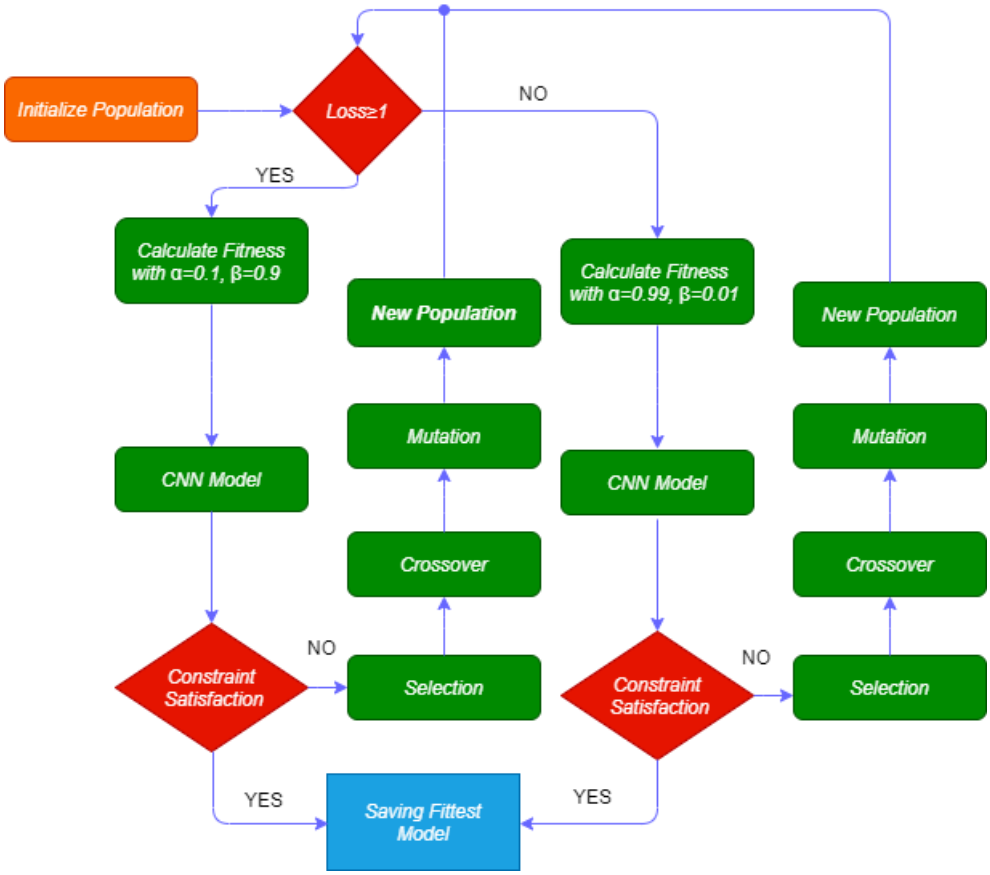


Figure 2. The flowchart of the proposed GA.

The fitness function has been defined in a way that increases the accuracy while decreasing the loss. To make sure the value of the fitness function is smooth, we define the values of α and β as follows:

- If the loss (l) is greater than or equal to 1, the categorical cross-entropy loss (CCL) varies between 1 and 10. Thus, we define alpha and beta as 0.1 and 0.9, respectively;
- If the loss (l) is less than 1, the CCL varies between 0.001 and 1. Thus, we define α and β as 0.99 and 0.01, respectively.

In summary, the proposed approach sets the number of neurons, convolutional kernel size, and convolutional filter size. Besides choosing the mentioned parameters of the model, GAOCNN indicates the number of epochs for training the model, too. Thus, GAOCNN tries to increase the number of epochs as long as the tuned structure increases the accuracy and decreases the loss. Choosing the number of epochs for training leads to the optimal time for training.

4. Experimental Results

To evaluate the proposed model for diabetic patients, we use the diabetes 130-US hospitals dataset [33]. The dataset represents ten years (1999–2008) of clinical care at 130 hospitals and integrated delivery networks in the United States. In total, there are 101,766 records (encounters) available for analysis. This data source generally has

50 attributes (13 attributes are integer types, and 37 attributes are object types). In this research, we use attributes with a missing value percentage of less than 20%. We have also removed constant and quasi-constant attributes for the dataset, as these provide no information for the classification task. Constant attributes are the features that contain a single value for all records in the dataset [43]. Quasi-constant attributes are almost stable features. Here, we consider features quasi-constant with the same value in more than 99.99% of the records. Hospital readmission was stratified into three cohorts: patients who were never readmitted after discharge, those who were readmitted within 30 days of discharge, and those who are readmitted 30 days after discharge (up to a year) [33]. Figure 3 shows the population size of each diabetic patient. As can be seen, 54% of the patients are never readmitted after discharge, resulting in imbalanced data.

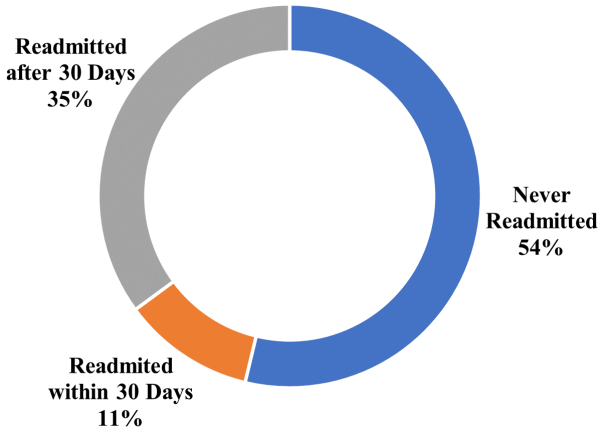
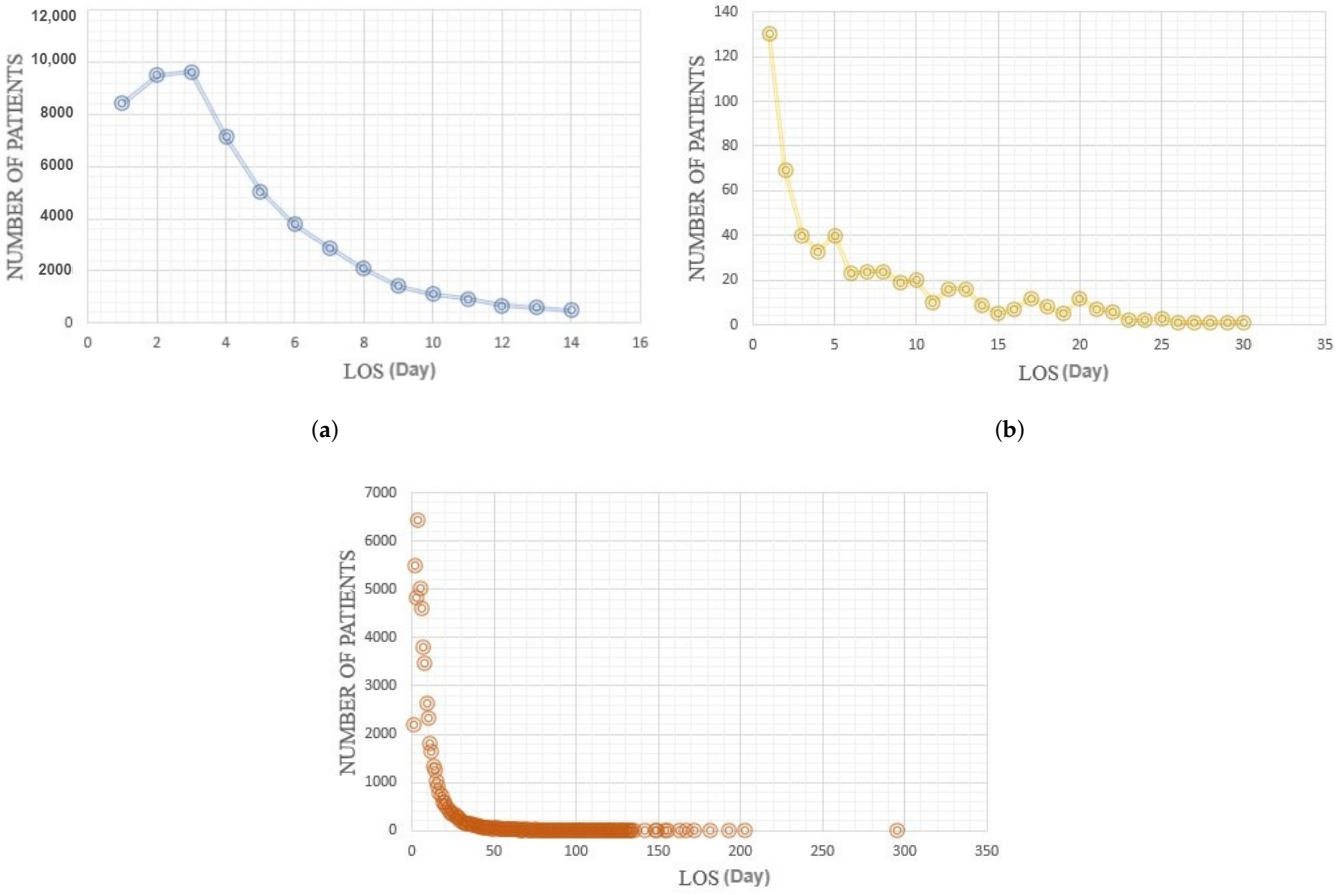


Figure 3. Distributions of the readmission in diabetic patients.

The hospital LOS range varies for different diseases. For diabetic patients, the LOS is between 1 and 14 days. For COVID-19 patients, the LOS is between 1 and 27 days in the dataset. For ICU patients, it is between 1 and 289 days in the MIMIC-III dataset. We create different classes for the LOS on each disease to consider these variations. For diabetic patients, we consider seven classes: 1–2 days, 3–4 days, 5–6 days, 7–8 days, 9–10 days, 11–12 days, and 13–14 days. For COVID-19 patients, we consider these classes: 1–2 days, 3–4 days, 5–6 days, 7–8 days, 9–10 days, 11–12 days, 13–16 days, 17–20 days, and 21–27 days. For the MIMIC-III dataset, we consider the classes the same as the COVID-19 dataset up to a 20-day LOS. However, for a LOS more extended than 20 days, we make these classes: 21–30 days, 31–50 days, 51–80 days, 81–110 days, and longer than 110 days. We have considered 3-day intervals for short-term LOSs similar to the existing research [22,44,45]. For the long-term LOS, we considered larger intervals to avoid having many classes. The narrow class division will help hospitals and the healthcare system determine hospital staff and beds for better servicing of patients.

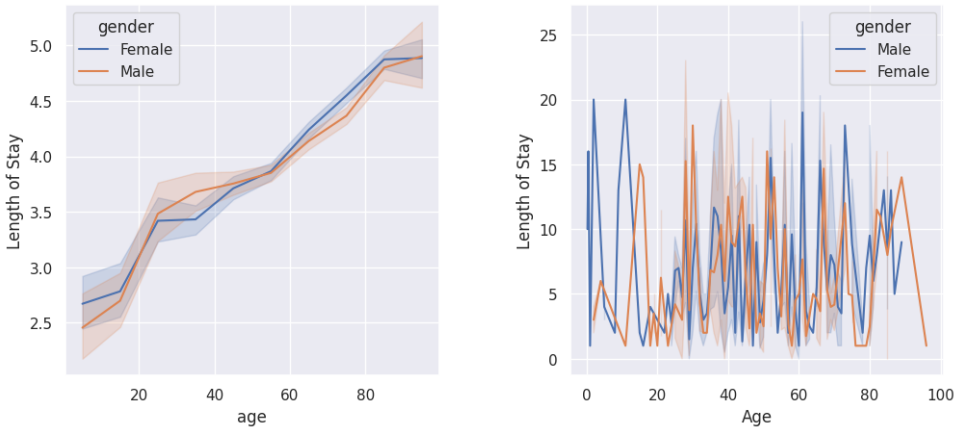
The distribution of the LOS in each dataset is shown in Figure 4. As can be seen, the density (the number of patients) decreases with the increase in the LOS. One of the most important features that affect patients' LOS is age [46]. Figure 5 demonstrates the relation between LOS and the age of each patient based on the patient's gender for instances with the highest LOS distributions. In Figure 5, each graph shows if LOS will increase or decrease by increasing each patient's age. Based on Figure 5a diabetic patient age has a direct correlation to their LOS in the hospital. By increasing the age, especially over 60 years old, the LOS has increased to more than four days. Figure 5b demonstrates various relations between age and LOS, and there are no clear trends for LOS and age for COVID-19 patients. Figure 5c indicated a high LOS for infants in the emergency room and a direct relation between LOS and age, especially for patients over 18 years old.



(**a**) (**b**)

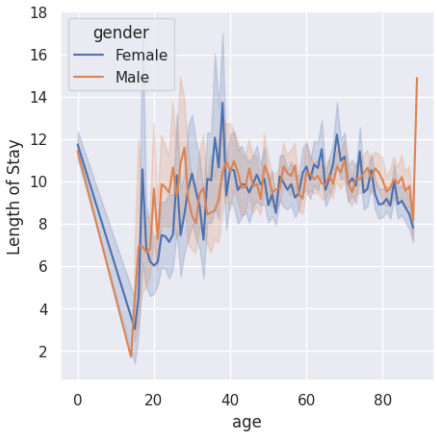
(**c**)

Figure 4. LOS distribution: (**a**) Diabetes, (**b**) COVID-19, (**c**) MIMIC-III.



(**a**) (**b**)

Figure 5. *Cont.*



(**c**)

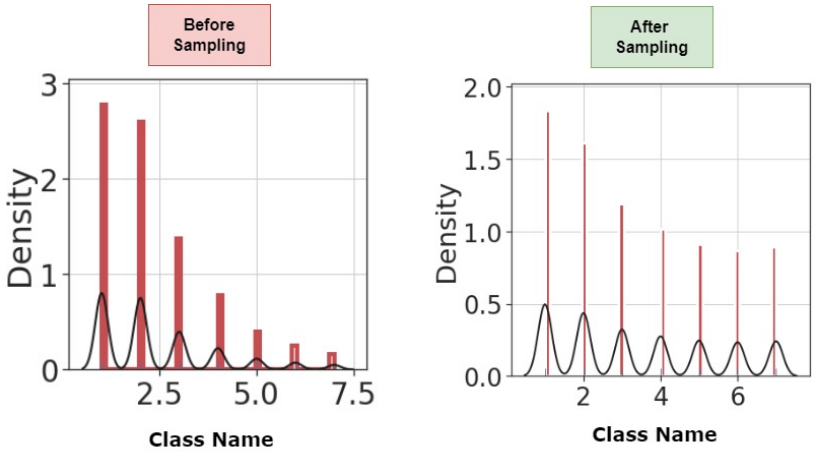
Figure 5. Relation between age and LOS based different genders for instances with the highest LOS distributions: (**a**) Diabetes, (**b**) COVID-19, (**c**) MIMIC-III.

4.1. Preprocessing

Considering the number of null values for each feature in the datasets, we ignored features with more than 20% unknown values. We imputed the features that have a null value of less than 20% with the use of k-nearest neighbour [47]. Furthermore, we computed the correlation between features and eliminated the features having more than 50% correlation. We also eliminated features with constant values. After cleaning the datasets, we applied three different encoding procedures. First, we used a label encoder that converts 'No' values to '0' and 'Yes' to '1'. Then, we applied one-hot encoding and target encoding [48] to the cleaned datasets. One-hot and target encoders have shown promising results when a CNN is used as a classifier [49].

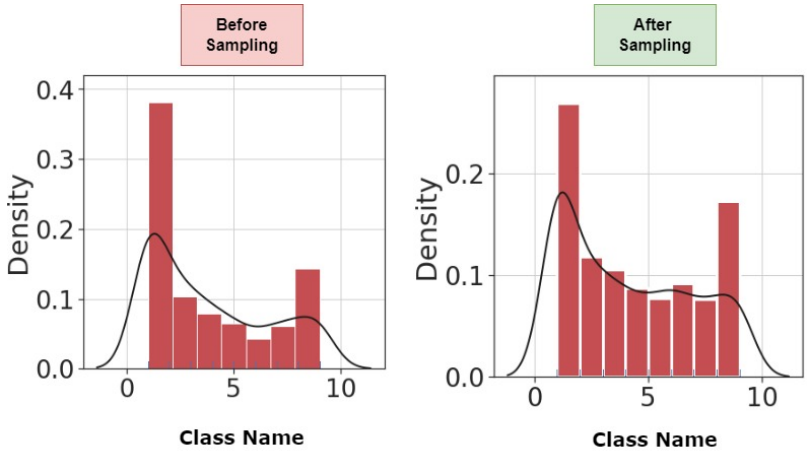
As mentioned, the used datasets are imbalanced regarding readmission and the LOS. It can affect the performance of the proposed model. Here, we use an advanced sampling technique called T-Link [50], followed by an oversampling technique to balance the datasets. This method increased the total number of instances for the readmission prediction to 33,104 samples 11,150 never readmitted, 11,150 readmitted within 30 days, and 10,804 readmitted after 30 days. The distribution of the LOS in the balanced datasets is shown in Figure 6.

As shown in Figure 6, after using sampling methods, the distance between the distribution of each class is decreased, while the original distribution is saved.

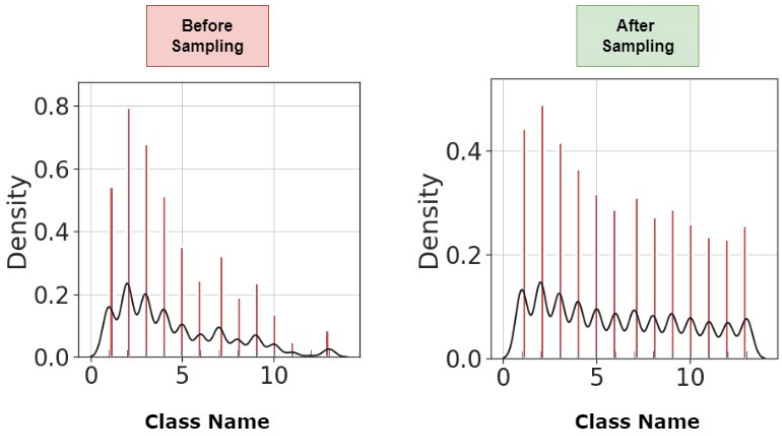


(**a**)

Figure 6. *Cont.*



(**b**)



(**c**)

Figure 6. Distributions of the LOS in the original and balanced datasets: (**a**) Diabetes, (**b**) COVID-19, (**c**) MIMIC-III.

4.2. Performance Analysis

After preprocessing, we divide each dataset into train, validation, and test. For better evaluation of the proposed model, we use k-fold cross-validation [51]. Here, we consider K as 10. A single fold acts as a test set, while the remaining nine folds are used as the training set. Finally, the results are averaged to represent a single estimation. The model was trained using the Tesla P100 graphics processing unit. The runtime for reaching the desired result differed based on the dataset and the number of classes for prediction. For MIMIC-III and diabetes datasets, the runtime to train the model was between 3 and 4 days, whereas for LOS prediction on the COVID-19 dataset, the runtime was about 6 h. To compare the proposed model, we used VGG16, ResNet, GoogLeNet [32], LR [17], RF [17], to XGB [20]) and SVM as the benchmarks. Furthermore, we implemented a combination of CNN and LR, CNN and RF, CNN and XGB, CNN and SVM, and a semi-supervised generative adversarial network (SGAN) model. We used convolutional layers to combine CNN with other ML methods as feature extractors and ML models as classifiers [52]. SGAN uses the CNN model achieved by GAOCNN as a generator and a multi-layer perception with three hidden layers and 128, 64, and 23 units, respectively, as discriminator [53]. We just converted the structure of the 2D convolutional layer of the mentioned model into 1D convolutional to match the structure of healthcare data. Table 3 indicates the performance of the readmission prediction using the proposed model (GAOCNN) and the benchmark models for diabetic patients. As can be seen, the GAOCNN outperforms all benchmarks.

Table 3. Results of readmission prediction for diabetics patients.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F-Measure (%)	Precision (%)
GAOCNN	97.2	96.7	99.3	96.9	97.1
VGG16	38.0	38.2	37.8	45.6	38.2
ResNet	38.0	38.2	38	44.2	38.2
GoogLeNet	39.6	38.4	50.3	38.4	38.4
LR	86.8	86.8	93.4	86.8	86.8
RF	90.0	94.4	96.5	90.0	90.0
XGB	94.4	94.4	97.8	94.4	94.5
SVM	94.9	94.3	98.4	94.9	94.9
CNN + LR	87.5	86.5	94.2	87.5	87.4
CNN + RF	91.7	91.4	96.8	91.7	91.7
CNN + XGB	94.8	94.6	98.9	94.8	94.8
CNN + SVM	95.1	95.1	95.1	95.1	95.1
SGANs	58.9	51.7	52.6	56.9	63.3

The classification results of the LOS for diabetic, COVID-19, and ICU patients are shown in Tables 4–6, respectively. As can be seen, the performance of the GAOCNN is higher than all benchmarks for the LOS prediction in all diseases.

Table 4. Results of the LOS prediction using different models for diabetic patients.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Measure (%)	Precision (%)
GAOCNN	89.0	89.8	97.8	90.2	90.4
VGG16	18.1	18.1	25.4	18.1	18.1
ResNet	17.7	17.7	20.8	17.7	17.7
GoogLeNet	28.6	2.3	35.6	4.5	67.9
LR	28.9	28.9	32.6	26.4	26.3
RF	79.9	79.9	92.7	79.7	79.6
XGB	78.8	78.8	92.6	78.3	77.9
SVM	36.5	33.5	42.3	32.1	31.9
CNN + LR	32.7	32.7	45.3	31.3	30.9
CNN + RF	80.0	80.0	93.4	79.7	79.6
CNN + XGB	78.8	78.8	94.4	78.3	77.9
CNN + SVM	36.2	36.2	43.3	34.8	34.5
SGANs	43.5	14.9	75.1	23.6	72.9

Table 5. Results of the LOS prediction using different models for COVID-19 patients.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Measure (%)	Precision (%)
GAOCNN	99.4	99.4	99.8	99.4	99.4
VGG16	14.1	14.6	20.5	14.6	14.6
ResNet	12.7	12.7	17.8	12.7	12.7
LR	92.1	92.1	98.8	92.1	92.3
RF	89.3	89.3	95.6	89.2	89.1
XGB	91.4	91.4	98.4	91.4	91.3
SVM	84.7	84.7	92.8	84.7	84.8
CNN + LR	70.3	70.3	89.9	70.2	70.6
CNN + RF	87.3	87.3	96.1	87.3	87.4

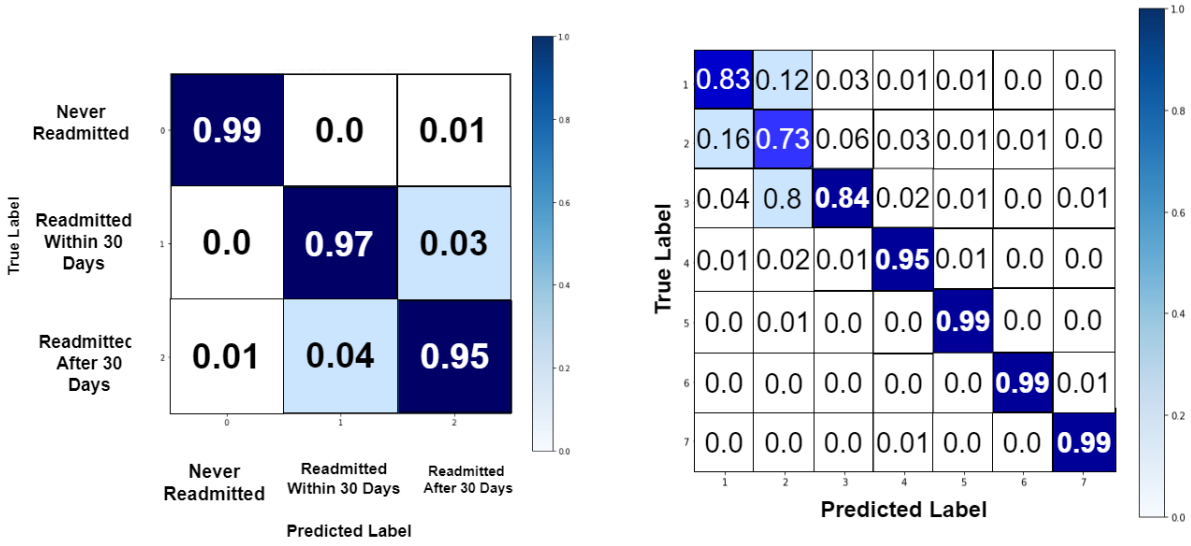
Table 5. *Cont.*

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Measure (%)	Precision (%)
CNN + XGB	87.7	87.7	96.2	87.8	88.6
CNN + SVM	81.3	81.3	92.5	81.3	81.8
SGANs	93.5	93.3	98.8	93.6	93.9

Table 6. Results of LOS prediction using different models for ICU patients.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Measure (%)	Precision (%)
GAOCNN	94.1	94.0	98.8	94.2	94.5
VGG16	10.1	10.1	20.6	10.1	10.1
ResNet	8.7	28.7	8.9	17.7	17.7
GoogLeNet	17.7	15.9	42.6	25.2	60.1
LR	43.9	43.9	65.1	38.4	36.2
RF	76.1	76.1	89.6	76.1	76.0
XGB	83.5	83.5	93.7	83.3	83.2
SVM	56.0	59.4	83.3	56.1	56.0
CNN + LR	43.6	43.6	72.7	42.4	41.8
CNN + RF	80.9	80.9	90.6	80.9	81.0
CNN + XGB	83.2	83.2	96.5	83.1	82.9
CNN + SVM	39.8	39.8	59.0	39.3	39.6
SGANs	56.1	45.7	92.6	54.5	67.7

For better observation of the performed prediction tasks using GAOCNN, we compute the model's normalized confusion matrix [54]. For readmission prediction in diabetic patients, the confusion matrix is shown in Figure 7. The result shows that there is just a 3% chance of incorrect prediction for the patients who are readmitted within 30 days. Furthermore, for the patients who are readmitted after 30 days, the error rate is 5%. For the LOS, the normalized confusion matrix of the prediction in diabetic, COVID-19, and ICU patients is shown in Figures 7–9, respectively.



(**a**) (**b**)

Figure 7. The normalized confusion matrix; (**a**) readmission prediction in diabetic patients, (**b**) LOS prediction in diabetic patients.

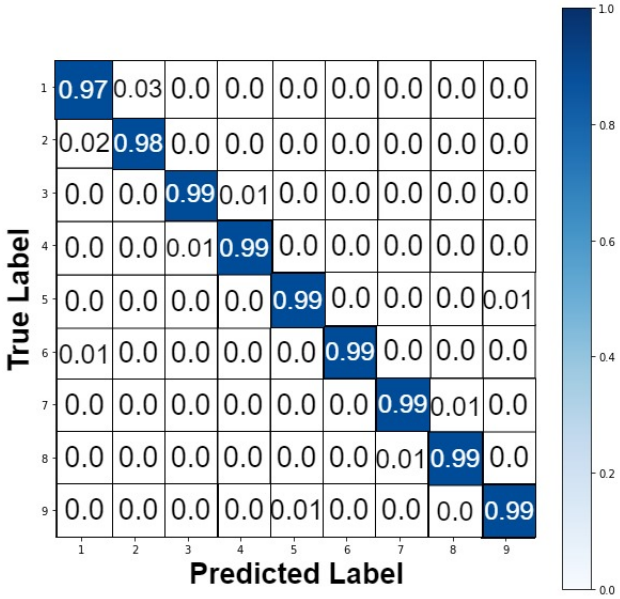


Figure 8. The normalized confusion matrix for the LOS prediction in COVID-19 patients.

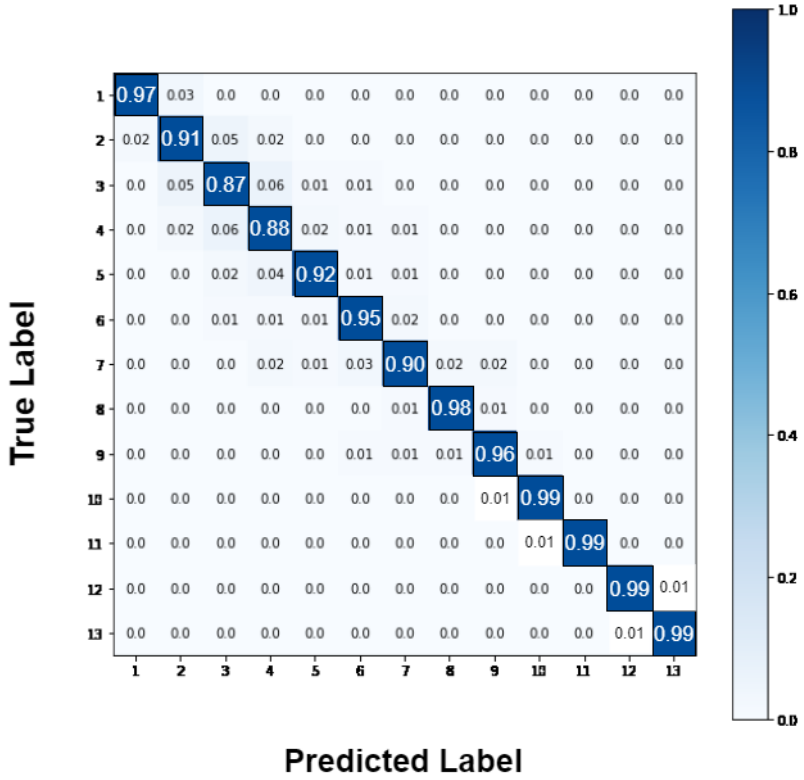


Figure 9. The normalized confusion matrix for the LOS prediction in ICU patients.

4.3. Comparison to Similar Research

To compare the performance of the GAOCNN to other research, we surveyed the recently published work on hospital readmission and LOS prediction. We used accuracy and AUROC reported in the papers in this comparison. The result of this comparison is shown in Tables 7 and 8. The missing values in the tables mean the papers have not reported them. This comparison result confirms that the GAOCNN is superior to the published works in hospital readmission and LOS prediction tasks.

Table 7. Comparison of the proposed model with the published works for readmission prediction on the diabetes dataset.

Authors	Accuracy (%)	AUROC (%)
Tamin and Iswari [55] (2017)	75.9	-
Hammoudeh et al. [18] (2018)	92	95
Popel et al. [50] (2018)	82.27	-
Alturki et al. [56] (2019)	94.8	-
Goudjerkan and Jayabalan [57] (2019)	95	95
Seraphim et al. [58] (2020)	86	66.7
Norbrun [59] (2021)	89.7	96
GAOCNN	97.2	99

Table 8. Comparison of the proposed model and the published works for the LOS prediction on diabetes, COVID-19, and MIMIC-III datasets.

Authors	Number of Classes	Accuracy (%)	AUROC (%)	Dataset
Gentimis et al. [60] (2017)	2	79.8	-	MIMIC-III
Steele and Thompson [61] (2019)	2	87.7	88	Diabetes
Alturki et al. [56] (2019)	3	85.4	-	Diabetes
Nallabasannagari et al. [27] (2020)	2	66.2	88	MIMIC-III
Wang et al. [25] (2020)	2	68.3	73.3	MIMIC-III
Wang et al. [25] (2020)	2	91.2	71	MIMIC-III
Etu et al. [62] (2022)	2	85	93	COVID-19
Alabbad et al. [63] (2022)	9	94.16	-	COVID-19
GAOCNN	7	89	96	Diabetes
GAOCNN	13	94.1	99	MIMIC-III
GAOCNN	9	99.4	99	COVID-19

Based on the contrasted results of Table 7, the proposed model outperformed similar research due to detecting the minority class (readmitting within 30 days). Table 8 indicates superior results compared to similar research, especially for discriminating between various conditions for long-term LOS. In both Figures 8 and 9, the performance of the model for separating between long and short-term LOS is outstanding. This performance indicates the ability of proposed objective functions to leverage the feature extractor and final classifier to improve the accuracy in minority and majority classes simultaneously. However, in similar research, by decreasing the number of LOS time frames only to short-term and long-term classes, excellent accuracy in short-term LOS prediction covers the poor performance of long-term LOS prediction. By increasing the number of classes and leveraging GAOCNN to increase the accuracy in all classes, GAOCNN outperformed previous research.

5. Discussion

The GAOCNN uses a hybrid structure of deep 1D convolutional networks with GA, and it is adequate for situations where the existing data is imbalanced and gathering more data is difficult. Notably, applying the proposed model helps develop an expert system to predict hospital readmission and LOS accurately. The GAOCNN is well-tuned for the readmission and the LOS prediction tasks. To evaluate the GAOCNN, we used datasets of diabetic, COVID-19, and ICU patients. The results show that the GAOCNN has a significant accuracy in predicting hospital readmission and LOS compared to existing techniques. The main contribution of this research is to help manage hospitals' resources more accurately. Furthermore, the proposed model applies to various conditions such as chronic diseases, pandemics, and intensive care. It is another contribution of this research proposing one model for different conditions.

GAOCNN presents a CNN model for accurate LOS prediction. Thus, we used forward and backward feature selection techniques [64] to specify the most critical features for LOS and readmission time frame classification. The result of feature selection based on accuracy is shown in Figure 10.

Most important features			
Diabetes (Readmission)	**Diabetes (LOS)**	**COVID-19**	**ICU**
Number of lab procedures	Number of lab procedures	Number of case in the country per day	Admission diagnosis
Number of Prescribed medications	Number of Prescribed medications	Age	Ethnicity
First diagnosis result	Number of outpatient visits by patient in proceeding year of admission	Symptoms	Age
Third diagnosis result	Number of diagnosis while admitting	Gender	Gender
Second diagnosis result	Number of inpatient visits by patient in proceeding year of admission	Number of recovered cases per day	Intake for patients monitored using the Philips CareVue system while in the ICU
Number of outpatient visits by patient in proceeding year of admission	Number of conducted procedures on admitted patients	Number of died cases per day	Output information for patients while in the ICU
Number of conducted procedures on admitted patients	First diagnosis result	Time of exposure to the public	Number of diagnosis
Measurement of Insulin in the blood test	Third diagnosis result	Time before critical condition	Type of Admission procedure for each patient
Gender	Second diagnosis result	Visiting Wuhan	Insurance status
Age	-	Location	Medications ordered for a given patient
-	-	-	Number of conducted lab test on patients
Calculated accuracy with most important features			
92.81%	87.20%	89.45%	88.47%

Figure 10. Best selected features based on accuracy with wrapper feature selection.

As shown in Figure 10, specific features such as first diagnosis, symptoms, age, and gender are more important than other features for LOS and readmission time frame classification. Using the proposed approach, there is no need to deal with hyperparameters. To achieve a balanced dataset, we considered different numbers of classes for LOS. Then, we combined over and under-sampling methods to decrease the difference between class densities. Considering the high performance of the GAOCNN model, we can develop a system that aids healthcare systems to improve their medical services allocation and apply proper management to staff and patients. To predict the readmission time frame, we have prioritized accuracy over loss, while for predicting the LOS, we have prioritized loss over accuracy.

6. Limitations and Future of the Work

The main limitation of the proposed method is the time for optimizing the feature extraction and classifier. By increasing the number of features and the length of the dataset, the search space of GAOCNN will increase, and the required time for training the model will increases exponentially. To decrease the time for training, the same objective function can be used based on the strategy of other metaheuristic optimizers, such as particle swarm and grey wolf optimizers [65]. Another area for improvement for the proposed method is the quality of extracted features from the original feature sets. This research uses a shallow CNN to extract relevant information from the available feature set. Based on the number of features, and the relation between recorded features from the patient and the final target, the structure of the feature extractor and its depth should be changed. In the future, a dynamic feature extractor based on the type of disease and the recorded features from the patients will be designed to improve the quality of the extracted features for final classification. Furthermore, to decrease the simulation time of the proposed method,

a variety of metaheuristic methods will be evaluated as the optimizer to choose the structure of the feature extractor.

7. Conclusions

Predicting hospital readmissions and the LOS for diabetic, COVID-19, and ICU patients is a challenging task essential in disease trend monitoring and cost management. With the growth in the number of patients and the emergence of COVID-19, we should equip healthcare systems with expert systems to extract useful information for resource planning. We presented the GAOCNN as a high-performing ML model to predict hospital readmissions and the LOS. The GAOCNN is robust to missing and null values, and can make precise predictions due to imbalanced data and errors in the recorded attributes. The GAOCNN model is state-of-the-art for both hospital readmission and LOS predictions. For the readmission prediction, we reached a total accuracy of 97.1%, including 97% accuracy for the patients who were readmitted within 30 days. For the LOS prediction, the proposed model reached 89.0%, 99.4%, and 94.1% accuracy for diabetic, COVID-19, and ICU patients, including 99% accuracy for long-term stays of all diseases. Using the GAOCNN, healthcare systems can develop a framework for predicting both the readmission and the LOS of diabetic patients. Furthermore, the GAOCNN can help healthcare providers in pandemic situations by providing a lower mortality risk factor for diabetic patients and preventing the prevalence of pandemic diseases.

Author Contributions: Conceptualization, A.T., A.R. and F.H.; methodology, A.T., A.R. and F.H.; validation, A.T., A.R. and F.H.; formal analysis, A.T., A.R. and F.H.; investigation, A.T.; writing—original draft preparation, A.T. and F.H.; writing—review and editing, A.T. and F.H.; visualization, A.T.; supervision, A.R., F.H. and S.U.; project administration, A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Not applicable; this study did not report any data.

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

Abbreviations

AUROC	Area Under the Receiver Operating Curve
CNN	Convolutional Neural Network
CCL	Categorical Cross-Entropy Loss
DL	Deep Learning
RF	Random Forest
GA	Genetic Algorithm
GAOCNN	Genetic Algorithm-Optimized Convolutional Neural Network
ICD	International-statistical Classification of Diseases
ICU	Intensive Care Unit
LR	Logistic Regression
LSTM	Long Short Term Memory
ML	Machine Learning
MLP	Multi-Layer Perceptron
SGAN	Semi-Supervised Generative Adversarial Network
SVM	Support Vector Machine
XGB	Extreme Gradient Boosting

References

1. Desai, D.; Mehta, D.; Mathias, P.; Menon, G.; Schubart, U.K. Health care utilization and burden of diabetic ketoacidosis in the US over the past decade: A nationwide analysis. *Diabetes Care* **2018**, *41*, 1631–1638. [CrossRef]
2. Friedberg, M.W.; Rosenthal, M.B.; Werner, R.M.; Volpp, K.G.; Schneider, E.C. Effects of a medical home and shared savings intervention on quality and utilization of care. *JAMA Intern. Med.* **2015**, *175*, 1362–1368. [CrossRef]

3. Mata-Cases, M.; Casajuana, M.; Franch-Nadal, J.; Casellas, A.; Castell, C.; Vinagre, I.; Mauricio, D.; Bolíbar, B. Direct medical costs attributable to type 2 diabetes mellitus: A population-based study in Catalonia, Spain. *Eur. J. Health Econ.* **2016**, *17*, 1001–1010. [CrossRef]
4. Huang, E.S.; Laiteerapong, N.; Liu, J.Y.; John, P.M.; Moffet, H.H.; Karter, A.J. Rates of complications and mortality in older patients with diabetes mellitus: The diabetes and aging study. *JAMA Intern. Med.* **2014**, *174*, 251–258. [CrossRef]
5. Riddle, M.C.; Herman, W.H. The cost of diabetes care—An elephant in the room. *Diabetes Care* **2018**, *41*, 929–932. [CrossRef]
6. Pasquini-Descomps, H.; Brender, N.; Maradan, D. Value for money in H1N1 influenza: A systematic review of the cost-effectiveness of pandemic interventions. *Value Health* **2017**, *20*, 819–827. [CrossRef]
7. Tsai, Y.; Vogt, T.M.; Zhou, F. Patient characteristics and costs associated with COVID-19–related medical care among Medicare fee-for-service beneficiaries. *Ann. Intern. Med.* **2021**, *174*, 1101–1109. [CrossRef]
8. Gural, A. *Algorithmic Techniques for Neural Network Training on Memory-Constrained Hardware*; Stanford University: Stanford, CA, USA, 2021.
9. Faes, C.; Abrams, S.; Van Beckhoven, D.; Meyfroidt, G.; Vlieghe, E.; Hens, N.; Belgian Collaborative Group on COVID-19 Hospital Surveillance. Time between symptom onset, hospitalisation and recovery or death: Statistical analysis of Belgian COVID-19 patients. *Int. J. Environ. Res. Public Health* **2020**, *17*, 7560. [CrossRef]
10. Muniyappa, R.; Gubbi, S. COVID-19 pandemic, coronaviruses, and diabetes mellitus. *Am. J. Physiol.-Endocrinol. Metab.* **2020**, *318*, E736–E741. [CrossRef]
11. Tavakolian, A.; Hajati, F.; Rezaee, A.; Fasakhodi, A.O.; Uddin, S. Fast COVID-19 versus H1N1 screening using Optimized Parallel Inception. *Expert Syst. Appl.* **2022**, *204*, 117551. [CrossRef]
12. Arnaud, É.; Elbattah, M.; Gignon, M.; Dequen, G. Deep learning to predict hospitalization at triage: Integration of structured data and unstructured text. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 4836–4841.
13. Shinde, P.P.; Shah, S. A review of machine learning and deep learning applications. In Proceedings of the 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pune, India, 16–18 August 2018; pp. 1–6.
14. Desai, K.M.; Survase, S.A.; Saudagar, P.S.; Lele, S.; Singhal, R.S. Comparison of artificial neural network (ANN) and response surface methodology (RSM) in fermentation media optimization: Case study of fermentative production of scleroglucan. *Biochem. Eng. J.* **2008**, *41*, 266–273. [CrossRef]
15. Alloghani, M.; Aljaaf, A.; Hussain, A.; Baker, T.; Mustafina, J.; Al-Jumeily, D.; Khalaf, M. Implementation of machine learning algorithms to create diabetic patient re-admission profiles. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 253. [CrossRef]
16. Mai, Q. A review of discriminant analysis in high dimensions. *Wiley Interdiscip. Rev. Comput. Stat.* **2013**, *5*, 190–197. [CrossRef]
17. Pranckevičius, T.; Marcinkevičius, V. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Balt. J. Mod. Comput.* **2017**, *5*, 221. [CrossRef]
18. Hammoudeh, A.; Al-Naymat, G.; Ghannam, I.; Obied, N. Predicting hospital readmission among diabetics using deep learning. *Procedia Comput. Sci.* **2018**, *141*, 484–489. [CrossRef]
19. Mingle, D. Predicting diabetic readmission rates: Moving beyond Hba1c. *Curr. Trends Biomed. Eng. Biosci.* **2017**, *7*, 555707. [CrossRef]
20. Voyant, C.; Notton, G.; Kalogirou, S.; Nivet, M.L.; Paoli, C.; Motte, F.; Fouilloy, A. Machine learning methods for solar radiation forecasting: A review. *Renew. Energy* **2017**, *105*, 569–582. [CrossRef]
21. Chauhan, V.K.; Dahiya, K.; Sharma, A. Problem formulations and solvers in linear SVM: A review. *Artif. Intell. Rev.* **2019**, *52*, 803–855. [CrossRef]
22. Morton, A.; Marzban, E.; Giannoulis, G.; Patel, A.; Aparasu, R.; Kakadiaris, I.A. A comparison of supervised machine learning techniques for predicting short-term in-hospital length of stay among diabetic patients. In Proceedings of the 2014 13th International Conference on Machine Learning and Applications, Detroit, MI, USA, 3–6 December 2014; pp. 428–431.
23. Yakovlev, A.; Metsker, O.; Kovalchuk, S.; Bologova, E. Prediction of in-hospital mortality and length of stay in acute coronary syndrome patients using machine-learning methods. *J. Am. Coll. Cardiol.* **2018**, *71*, A242. [CrossRef]
24. Tsai, P.F.J.; Chen, P.C.; Chen, Y.Y.; Song, H.Y.; Lin, H.M.; Lin, F.M.; Huang, Q.P. Length of hospital stay prediction at the admission stage for cardiology patients using artificial neural network. *J. Healthc. Eng.* **2016**, *2016*, 7035463. [CrossRef] [PubMed]
25. Wang, S.; McDermott, M.B.; Chauhan, G.; Ghassemi, M.; Hughes, M.C.; Naumann, T. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In Proceedings of the ACM Conference on Health, Inference, and Learning, Toronto, ON, Canada, 2–4 April 2020; pp. 222–235.
26. Rafiei, A.; Rezaee, A.; Hajati, F.; Gheisari, S.; Golzan, M. SSP: Early prediction of sepsis using fully connected LSTM-CNN model. *Comput. Biol. Med.* **2021**, *128*, 104110. [CrossRef]
27. Nallabasannagari, A.R.; Reddiboina, M.; Seltzer, R.; Zeffiro, T.; Sharma, A.; Bhandari, M. All Data Inclusive, Deep Learning Models to Predict Critical Events in the Medical Information Mart for Intensive Care III Database (MIMIC III). *arXiv* **2020**, arXiv:2009.01366.
28. Albahli, S.; Meraj, T.; Chakraborty, C.; Rauf, H.T. AI-driven deep and handcrafted features selection approach for COVID-19 and chest related diseases identification. *Multimed. Tools Appl.* **2022**, *81*, 37569–37589. [CrossRef] [PubMed]
29. Rehman, N.U.; Zia, M.S.; Meraj, T.; Rauf, H.T.; Damaševičius, R.; El-Sherbeeny, A.M.; El-Meligy, M.A. A self-activated CNN approach for multi-class chest-related COVID-19 detection. *Appl. Sci.* **2021**, *11*, 9023. [CrossRef]

30. Mahboub, B.; Al Bataineh, M.T.; Alshraideh, H.; Hamoudi, R.; Salameh, L.; Shamayleh, A. Prediction of COVID-19 hospital length of stay and risk of death using artificial intelligence-based modeling. *Front. Med.* **2021**, *8*, 592336. [CrossRef] [PubMed]
31. Nemati, M.; Ansary, J.; Nemati, N. Machine-learning approaches in COVID-19 survival analysis and discharge-time likelihood prediction using clinical data. *Patterns* **2020**, *1*, 100074. [CrossRef] [PubMed]
32. Ajit, A.; Acharya, K.; Samanta, A. A review of convolutional neural networks. In Proceedings of the 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), Vellore, India, 24–25 February 2020; pp. 1–5.
33. Strack, B.; DeShazo, J.P.; Gennings, C.; Olmo, J.L.; Ventura, S.; Cios, K.J.; Clore, J.N. Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records. *BioMed Res. Int.* **2014**, *2014*, 781670. [CrossRef]
34. Quan, H.; Sundararajan, V.; Halfon, P.; Fong, A.; Burnand, B.; Luthi, J.C.; Saunders, L.D.; Beck, C.A.; Feasby, T.E.; Ghali, W.A. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* **2005**, *43*, 1130–1139. [CrossRef]
35. Xu, B.; Kraemer, M.U. Open access epidemiological data from the COVID-19 outbreak. *Lancet* **2020**, *20*, 534. [CrossRef]
36. Johnson, A.E.; Pollard, T.J.; Shen, L.; Li-Wei, H.L.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.A.; Mark, R.G. MIMIC-III, a freely accessible critical care database. *Sci. Data* **2016**, *3*, 160035. [CrossRef]
37. Sun, Y.; Xue, B.; Zhang, M.; Yen, G.G.; Lv, J. Automatically designing CNN architectures using the genetic algorithm for image classification. *IEEE Trans. Cybern.* **2020**, *50*, 3840–3854. [CrossRef] [PubMed]
38. Peng, Y.; Nagata, M.H. An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. *Chaos Solitons Fractals* **2020**, *139*, 110055. [CrossRef] [PubMed]
39. Tavakolian, A.; Hajati, F.; Rezaee, A.; Fasakhodi, A.O.; Uddin, S. Source code Optimized Parallel Inception: A fast COVID-19 screening software. *Softw. Impacts* **2022**, *13*, 100337. [CrossRef] [PubMed]
40. Luo, G. A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Netw. Model. Anal. Health Inform. Bioinform.* **2016**, *5*, 18. [CrossRef]
41. Isa, S.M.; Suwandi, R.; Andrean, Y.P. Optimizing the Hyperparameter of Feature Extraction and Machine Learning Classification Algorithms. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*, 69–76. [CrossRef]
42. Kumar, A.; Kumar, D.; Jarial, S. A review on artificial bee colony algorithms and their applications to data clustering. *Cybern. Inf. Technol.* **2017**, *17*, 3–28. [CrossRef]
43. García, S.; Luengo, J.; Herrera, F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowl.-Based Syst.* **2016**, *98*, 1–29. [CrossRef]
44. Daghistani, T.A.; Elshawi, R.; Sakr, S.; Ahmed, A.M.; Al-Thwayee, A.; Al-Mallah, M.H. Predictors of in-hospital length of stay among cardiac patients: A machine learning approach. *Int. J. Cardiol.* **2019**, *288*, 140–147. [CrossRef]
45. Gowd, A.K.; Agarwalla, A.; Amin, N.H.; Romeo, A.A.; Nicholson, G.P.; Verma, N.N.; Liu, J.N. Construct validation of machine learning in the prediction of short-term postoperative complications following total shoulder arthroplasty. *J. Shoulder Elb. Surg.* **2019**, *28*, e410–e421. [CrossRef]
46. Guo, A.; Lu, J.; Tan, H.; Kuang, Z.; Luo, Y.; Yang, T.; Xu, J.; Yu, J.; Wen, C.; Shen, A. Risk factors on admission associated with hospital length of stay in patients with COVID-19: A retrospective cohort study. *Sci. Rep.* **2021**, *11*, 7310. [CrossRef]
47. Kumar, R.N.; Kumar, M.A. Enhanced fuzzy K-NN approach for handling missing values in medical data mining. *Indian J. Sci. Technol.* **2016**, *9*, 1–6.
48. Rodríguez, P.; Bautista, M.A.; Gonzalez, J.; Escalera, S. Beyond one-hot encoding: Lower dimensional target embedding. *Image Vis. Comput.* **2018**, *75*, 21–31. [CrossRef]
49. Gikunda, P.K.; Jouandeau, N. State-of-the-art convolutional neural networks for smart farms: A review. In Proceedings of the Intelligent Computing-Proceedings of the Computing Conference, London, UK, 16–17 July 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 763–775.
50. Popel, M.H.; Hasib, K.M.; Habib, S.A.; Shah, F.M. A hybrid under-sampling method (HUSBoost) to classify imbalanced data. In Proceedings of the 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 21–23 December 2018; pp. 1–7.
51. Li, Y.; Xia, J.; Zhang, S.; Yan, J.; Ai, X.; Dai, K. An efficient intrusion detection system based on support vector machines and gradually feature removal method. *Expert Syst. Appl.* **2012**, *39*, 424–430. [CrossRef]
52. Yang, S.; Gu, L.; Li, X.; Jiang, T.; Ren, R. Crop classification method based on optimal feature selection and hybrid CNN-RF networks for multi-temporal remote sensing imagery. *Remote Sens.* **2020**, *12*, 3119. [CrossRef]
53. Miao, X.; Wu, Y.; Wang, J.; Gao, Y.; Mao, X.; Yin, J. Generative semi-supervised learning for multivariate time series imputation. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 8983–8991.
54. Balasch, A.; Beinhofer, M.; Zauner, G. The Relative Confusion Matrix, a Tool to Assess Classifiablility in Large Scale Picking Applications. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 8390–8396.
55. Tamin, F.; Iswari, N.M.S. Implementation of C4. 5 algorithm to determine hospital readmission rate of diabetes patient. In Proceedings of the 2017 4th International Conference on New Media Studies (CONMEDIA), Yogyakarta, Indonesia, 8–10 November 2017; pp. 15–18.

56. Alturki, L.; Aloraini, K.; Aldughayshim, A.; Albahli, S. Predictors of Readmissions and Length of Stay for Diabetes Related Patients. In Proceedings of the 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 3–7 November 2019; pp. 1–8.
57. Goudjerkan, T.; Jayabalan, M. Predicting 30-day hospital readmission for diabetes patients using multilayer perceptron. *Int. J. Adv. Comput. Sci. Appl.* **2019**, *10*. [CrossRef]
58. Seraphim, I.; Ravi, V.; Rajagopal, A. Prediction of Diabetes Readmission using Machine Learning. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 42–49.
59. Norbrun, G. Reduction of Hospital Readmissions in Patients with a Diagnosis of COPD: An Integrative Review. Doctoral Dissertation, Liberty University, Lynchburg, VA, USA, 2021.
60. Gentimis, T.; Ala'J, A.; Durante, A.; Cook, K.; Steele, R. Predicting hospital length of stay using neural networks on mimic iii data. In Proceedings of the 2017 IEEE 15th International Conference on Dependable, Autonomic and Secure Computing, 15th International Conference on Pervasive Intelligence and Computing, 3rd International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech), Orlando, FL, USA, 6–10 November 2017; pp. 1194–1201.
61. Steele, R.J.; Thompson, B. Data mining for generalizable pre-admission prediction of elective length of stay. In Proceedings of the 2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 7–9 January 2019; pp. 0127–0133.
62. Etu, E.E.; Monplaisir, L.; Arslanturk, S.; Masoud, S.; Aguwa, C.; Markevych, I.; Miller, J. Prediction of Length of Stay in the Emergency Department for COVID-19 Patients: A Machine Learning Approach. *IEEE Access* **2022**, *10*, 42243–42251. [CrossRef]
63. Alabbad, D.A.; Almuhaideb, A.M.; Alsunaidi, S.J.; Alqudaihi, K.S.; Alamoudi, F.A.; Alhobaishi, M.K.; Alaqeel, N.A.; Alshahrani, M.S. Machine learning model for predicting the length of stay in the intensive care unit for COVID-19 patients in the eastern province of Saudi Arabia. *Inform. Med. Unlocked* **2022**, *30*, 100937. [CrossRef] [PubMed]
64. Déjean, S.; Ionescu, R.T.; Mothe, J.; Ullah, M.Z. Forward and backward feature selection for query performance prediction. In Proceedings of the 35th Annual ACM Symposium on Applied Computing, Brno, Czech Republic, 30 March–3 April 2020; pp. 690–697.
65. Pellerin, R.; Perrier, N.; Berthaut, F. A survey of hybrid metaheuristics for the resource-constrained project scheduling problem. *Eur. J. Oper. Res.* **2020**, *280*, 395–416. [CrossRef]