The First International Conference on
Machine Learning and Knowledge Discovery (MLKD 2024)
Amirkabir University of Technology, December 18-19, 2024

Predicting Patient's Recovery Process Using Electronic Health Records With Supervised Variational Autoencoder

Amirhossein Batouei* Ehsan Nazerfard† Amirhossein Askari‡

Abstract

Historically, human beings have employed every technology and tool available in order to improve their health and treatment. Various methods and tools were utilized in the past to predict an individual's health status. As a result of recent advancements in pervasive computing, data mining and deep learning, individual health prediction and assistance can be offered in a more effective way. Although research into electronic health records opened up new avenues, but also created new challenges. Researchers face an uphill battle when it comes to predicting a patient's physical state after discharge from a hospital or while in a hospital. In this paper, we introduce a novel approach, a supervised variational autoencoder, as a solution to predicting patient health status, particularly focusing on post-discharge and in-hospital scenarios. This approach is positioned as offering comparable or superior performance to existing state-of-the-art models while requiring fewer input variables and simplifying preprocessing steps. The research promises real-world data analysis to validate the proposed method, indicating a practical application of the model in healthcare settings.

Keywords: Health Prediction, Supervised Variational Autoencoder, Pervasive Computing, Deep Learning

1 Introduction

As human beings, we have always striven to maintain good health and have spared no effort to achieve this objective. In order to help individuals recover or regain their health, several tools and methods have been developed and introduced over the years. These methods have both positive and negative points when compared to other methods. Nonetheless, individuals have never stopped seeking new and innovative ways to achieve his goals. Novel methods in this field usually involve the combining of several different scientific disciplines in order to provide a new solution.

Examination of brain and heart activity with the help of an electroencephalogram (EEG) and electrocardiogram (ECG), and medical images is one of the common methods in assessing the patient's health and following the disease process. In the past, this was only possible with the help of specialist physicians and relying on their knowledge. With scientific advances, artificial intelligence has come to the aid of physicians in some of these analyzes and predictions. Before the advent of deep learning methods, research on physiological signals such as EEGs and ECGs used traditional machine learning methods. These methods required preprocessing and feature engineering. Nevertheless, deep learning methods require less of these two steps and make the job easier for researchers.

Only certain areas of medicine could be reviewed and examined using these methods. Therefore, the need for new solutions was felt to use similar methods in other fields of medicine. Electronic health records (EHRs), created for hospital management and financial applications, attracted the attention of some researchers and opened new doors for researchers [27, 30, 1]. By examining this data set and the valuable and comprehensive information found in it, researchers have been able to find useful medical information in various fields and offer new applications for it.

EHR systems store data on each patient admission, including information on diagnosis, treatment, demographics, tests and laboratory results, prescriptions, radiographs, clinical notes, and more. Although these systems were designed and created for other purposes, they gradually found new applications. For example, purposes such as extracting medical concepts, modeling the patient's recovery process, diagnosing the disease, and the like were added to the applications of these systems. Before the advent of deep learning, traditional machine learning methods were used to analyze this data. The increasing popularity of deep learning models, on the one hand, and the need for traditional methods of preprocessing and feature engineering have led researchers to apply more and more deep learning approaches in this area [16, 37, 9].

Predicting the patient's return to the hospital or readmission after discharge is one of the significant challenges in this area. Many researchers have tried to predict this issue in different ways. However, one of the

*Department of Computer Engineering, Amirkabir University of Technology, amirhbn@aut.ac.ir
†**Corresponding author**: Department of Computer Engineering, Amirkabir University of Technology, nazerfard@aut.ac.ir
‡Department of Computer Engineering, Amirkabir University of Technology, aha079@aut.ac.ir

most common methods is to predict the patient's return in the next few days. This period is usually equal to 30 days, and the issue becomes a binary classification issue: whether the patient will return to the hospital within 30 days of discharge. Although the output of the problem is defined and the same, the input data of different studies are different. Here we will make this prediction in a new and innovative way. Therefore, the model input will only include a sequence of diagnosis and treatment codes and demographic information to predict the patient's return to the hospital.

Although some studies in this field have used the variational autoencoders (VAEs) model [25, 36], to the best of our knowledge, this research is the first work that utilizes the SVAE for predicting patient's recovery process.

The simplicity of the model inputs provides the ability to implement it practically in the real world. These features are such that they can be collected even in hospitals with minor facilities. As a result, the proposed model can be easily implemented. The main advantage of this research over other methods is to achieve close and even better results than others, despite more straightforward inputs.

In this research, we develop a prediction model that employs VAEs in a novel binary classification that simplifies predictions compared to previous methods. The proposed model predicts the recovery phase of patients, but without significant feature engineering or pre-processing, and it works with basic datasets that may be gathered in hospitals with less technological sophistication. A strategy based on SVAE is proposed to predict whether a patient would return to the hospital within 30 days of discharge. To sum up, the contributions of the paper can be summarized as follows:

(i) It proposes a novel approach, a supervised variational autoencoder, as a solution to predicting patient health status, particularly focusing on post-discharge and in-hospital scenarios.

(ii) The propose approach is positioned as offering comparable or superior performance to existing state-of-the-art models while requiring fewer input variables and simplifying preprocessing steps.

(iii) The research promises real-world data analysis to validate the proposed method, indicating a practical application of the model in healthcare settings. Overall, the paper highlights the importance of the problem addressed, the innovation of the proposed solution, and the potential impact on healthcare practice.

The rest of the paper is organized as follows. First, we present some related studies on deep learning and EHRs. Next, a background summary of variational autoencoder models is provided. The details of the proposed model method are then presented. Then, we present the evaluation findings for the proposed prediction strategy. We conclude the paper with some concluding remarks and recommendations.

2 Related Work

While initially, patients' EHRs were used only to store patient information and perform medical and regulatory care, such as paying bills and fees, some researchers have suggested other uses of these records for various medical applications. In the new application, with the help of data in the patient's medical records, items such as predicting the patient's return time, predicting the patient's future condition, extracting information and relationships, and the like are done. With the remarkable advancement of deep learning methods, the distribution of articles each year in various areas related to EHR has increased. Unsupervised methods and recurrent neural networks (RNNs) have been the most used among alternatives [12].

Predicting patient outcomes is often the ultimate objective of Deep EHR systems. Two categories of outcome prediction are recognized in the research of [10]:

1. Static or one-off prediction (e.g. heart failure prediction using data from a single encounter).

2. Predicting future results in time (e.g. heart failure prediction within the next 6 months, or disease onset prediction using historical data from sequential encounters).

Researchers have used a recurrent neural network called RETAIN to predict heart failure patients after using a recurrent neural network [28]. Based on this model, a binary prediction can be made regarding whether the patient will develop heart failure in the future. Based on a comparison of this model with traditional machine learning models, the introduced model has a significant advantage over the logistic regression method in this case. The transfer of knowledge is also an issue that is addressed in this work. The data set of this study includes information from more than 150 thousand patients from about 400 hospitals. Consequently, knowledge transfer from one hospital to another has been addressed as part of the research [32]. According to the results, this model has an excellent ability to transfer knowledge between different hospitals. Using the trained model of one hospital for another hospital will only result in a small amount of error.

A type of restricted Boltzmann machine developed by [34] that can be used to represent medical topics.

This new representation facilitates algebraic and statistical tasks such as 2D spatial mapping, thematic grouping (to identify new phenotypes), and risk categorization. RBM represents raw, high-dimensional EMR data consisting of different data types homogeneously [29]. This simplifies the process of performing the operations described. The modified RBM is called eNRBM. The purpose of this study is to examine the risk of suicide among individuals with mental illness. Although estimating the risk of suicide is incorrect, it can lead to an uncomfortable feeling regarding suicide. The data for this study was collected from a hospital in Australia between January 2009 and March 2012. Participants were evaluated if they had attempted suicide at least once in the past. Each assessment can be used to predict future behavior. A suicide risk prediction is made for the next three months and categorized into three categories: low risk, moderate risk (non-lethal), and finally high risk, which will lead to death. In total, 15,272 (86.7%) of the participants had no risk outcomes, 1,436 (8.2%) had moderate risks, and 858 (4.9%) had high risks. Two eNRBM models are implemented and utilized in this study. The first model uses diagnosis (DIAG model) and the second model uses diagnosis and treatment procedures (DIAG + PROC model). To provide a comparison, an RBM model is trained with diagnosis codes. At first, data are mapped to a k-dimensional space and finally mapped to a two-dimensional space for display. Then, the feature vector produced by these models is fed into a logistic regression model, which determines whether suicide is likely. In order to compare the newly introduced model with the previous models, the support vector machine model is introduced as a representative of the traditional models. The results indicate that the model is superior to the previous models in describing the information for classification. This model with logistic regression has a significant advantage over support vector machines in the classification of suicide risk.

Based on deep belief networks, [21] identified risk factors and predicted progression of the bone disease. Osteoporosis is the most common type of bone disease. However, despite its absence of clinical symptoms, the disease is capable of causing significant mortality and complications after onset. Thus, it is extremely important to be aware of the risk factors for this disease. Because of the complexity and diversity of data, however, it can be challenging to predict disease progression and identify risk factors based on the characteristics of the disease and clinical differences. It is possible to determine both the leading causes of the disease as well as the distinction between the sick and healthy individuals with the aid of deep belief networks. Based on a data set concerning bone disease, the proposed method performs well in predicting the progression of osteoporosis.

The article [26] describes a deep, dynamic neural network that reads medical data, stores illness history, infers current disease states, and predicts future medical outcomes. DeepCare, which is based on Long Short-Term Memory (LSTM), enables the management of irregularly timed events by managing forgetting and consolidation. In addition, DeepCare replicates medical actions that modify disease progression and influence future medical risk. Upon ascending to the health state level, past and present health states are combined via multiscale temporal pooling, and then sent to a neural network that predicts future outcomes. Experiments demonstrate the efficacy of DeepCare in modelling the progression of sickness, providing remedies, and predicting future danger. DeepCare provides four unique features:

- Capturing long-term dependencies is made possible through memory maintenance.
- Continuous distributed vector space contains admissions of varying sizes.
- The forget gate is made a function of the irregular time lag between successive time steps for irregular timing.
- DeepCare is an end-to-end prediction model that doesn't depend on manual feature engineering, can read generic medical data, infer current disease states, and estimate future risk. It also doesn't need manual feature engineering..

The results show an increase in prediction accuracy for diabetes and mental health, two significant cohorts with high social and economic costs. The proposed model predicts medical diagnoses at the next visit as code for disease progression. The treatment proposal section proposes the necessary treatment model for each admission. Finally, the risk prediction is whether the patient will return within a certain discharge period. This period is three months for mental illness and 12 months for diabetes.

A research group in [15] have studied the effect of attention mechanisms on forecasting issues in this area. Since medical records are a reliable source of information, it is important to use them when analyzing patient records and predicting clinical outcomes. On the other hand, unstructured data, and their analysis requires models that facilitate human comprehension. Even though the predictive power of a model is important, its interpretability is also significant. As a result, by developing a mechanism to pay attention to the coding part, they examined its effect in five sections: predicting, re-admission, death, phenotype, and predicting hip and knee surgery complications. The authors assert that although the attention mechanism improves the performance of predictions, there is no explanation or

reasoning for the relationship between attention weights and performance.

By utilizing several deep and sophisticated architectures, the authors in [35] present several different representations of medical data and then use these features to predict two issues. The first is to predict the risk of heart failure, and the second is to predict how long the patient will stay in the intensive care unit (ICU). According to the proposed method, the data is preprocessed to one of the deep models before assigning an output representation vector to each category. After evaluating the results, the best is introduced. The deep architectures in this study are stacked autoencoders, deep belief networks (DBNs), and VAEs. The benefit of VAE over conventional autoencoders is that it learns the real distribution of the training data, as opposed to simply memorizing the specific training dataset [13]. In the reported findings for the first problem, which consists of a tiny data set, the disparities between the models are rather minimal. In contrast, for the second challenge, which has a larger data set, the framework is extremely beneficial for leveraging a vast quantity of unlabeled health records in order to extract a high-level representation of labeled data for supervised learning tasks.

Other works in the field of predicting the patient's condition and assessing the severity of the disease can be referred to the research of [31]. One of the most common criteria for assessing the status of patients admitted to the ICU is the sequential organ failure assessment (abbreviated SOFA). The SOFA uses 13 variables to evaluate the health of the body's six vital organs. This criterion can be used to assess the severity of the disease and predict death. This paper present a new model called DeepSOFA.

A new model named Resset, introduced in the paper [24], can be used to interpret clinical data and predict future outcomes. In this model, the patient's diseases are modeled as a set of treatments in one admission. Finally, the health recovery process is modeled as a sequence of admissions using an RNN. The experiments were performed on both diabetes and mental illness. The model's output predicts re-admission, treatment recommendation, disease prognosis, and disease progression. The results obtained in the re-admission prediction section show the superiority of this method over the basic model of machine learning (which is a combination of the bag of words and logistic regression). Of course, the model's performance is almost similar to that of a deep recurrent model. However, in the other two areas (treatment recommendation and disease prediction), the performance of the proposed model was quite superior to the others.

The authors in [6] have developed a model that predicts re-hospitalization over the next 30 days while using deep learning techniques. They claim that the proposed model can be interpreted and identified by important features by physicians. This model includes a convolutional network that takes clinical notes during discharge as input and predicts whether the patient's return within 30 days after discharge. The model is designed with a few layers to facilitate interpretation, so each expression can be examined to see how it affects the patient's re-admission.

There has also been proposed a method for predicting the patients in ICU developed by [25]. The method uses vital and laboratory information in the medical record to predict re-admission within 24 hours, 72 hours, seven days, 30 days, or between 24 and 72 hours. The prediction was made by XGBoost models, random forest, and logistic regression.

Researchers in [2] have developed a model to predict the re-admission of patients with heart failure. The proposed model is a deep RNNs made of LSTM units. The presented framework tries to eliminate three shortcomings in other research:

1. Exclusively uses human-derived characteristics or only machine-derived features. The former discards a substantial amount of information from each patient's record, while the later disregards human intelligence's knowledge and rules.

2. Ignores the chronological or sequential progression of events included in EHRs. EHRs comprise a series of measures (clinical visits) throughout time that contain crucial information on the course of illness and patient condition.

3. Fails to assess the skewness in terms of class imbalance and varied costs of misclassification mistakes (class imbalance problems are common with EHRs data).

A deep prediction model by [3] provides a prediction of the patient within 30 days following discharge. This model uses two types of data:

- Sequential data (eg: historical medical events)
- Static data (eg: age)

In this research, two deep learning models are introduced. The first is a modified model, a well-known model, and is a convolutional network, and the other is a bidirectional RNN. The data set used in this study is MIMIC. The time sequence data is first converted to new vectors by a skip-gram model in both proposed models. By combining these vectors with static properties, the models can determine whether a patient needs to be readmitted after being discharged. The effect of static features on improving the model's performance is one of the outstanding features of this research.

It is a challenging and time-consuming process to extract and insert the correct codes into a patient's file. Hence authors in [14] proposed a deep learning model that analyzes individual clinical notes and extracts associated ICD codes. This model and similar cases can be beneficial to researchers in the field of information extraction and preprocessing function.

According to the study of [23], VAE was used to represent patient information. The recurrent model regarding to the time sequence of data are discussed in several studies [26, 28, 24, 18]. Moreover, the researchers in [4] employed SAVE in order to enhance the image data. The SVAE model has not been applied to data obtained from EHRs until now, and this is the first time that this type of model has been applied to data obtained from EHRs.

The research presented in this paper differs from the earlier efforts in two significant ways. The model proposes a novel supervised inference process based on VAEs. The second distinction is that we are able to accurately predict the conditions of patients using more straightforward inputs that are practical for the real world.

3 Background

Health records are intricate, diverse, and multidimensional. Therefore, representation learning is required to tackle these challenges and abstract medical data to a higher level in order to deliver stronger features. In contrast, labeling clinical data is difficult and expensive in many instances (such as certain disorders) where the data may be unlabeled. Representation learning with an unsupervised method is a potent and valuable technique for extracting features from labeled or unlabeled data that can enhance the performance of models [23].

The following are the primary issues associated with the processing of electronic health records [5]:

1. Hyper dimensionality
2. Temporality referring to the succession of clinical occurrences
3. Sparsity
4. The EHRs are characterized by irregularity, which entails substantial variation
5. Including systemic flaws, bias is present in the medical data

Representation learning can assist in overcoming these obstacles and enhancing the performance of machine learning algorithms. Due of this, several researchers have concentrated their efforts on representation learning techniques. Deep-learning approaches translate data to a higher level of abstraction using simple yet nonlinear transformations, and have demonstrated promising performance in computer vision, speech recognition, and natural language processing [23].

We briefly overview the variational autoencoder models (VAEs), and introduce supervised variational autoencoder models (SVAEs).

3.1 Variational Autoencoder

In recent years, VAEs have been developed as a useful method for learning to represent complex data. These networks have already shown promising performance in complex data including handwritten figures, faces, and speech models. VAEs include an encoder, decoder and hidden layers. Moreover, these models are probabilistic generators [23].

Assume that our input data X and the variable z is hidden, we have a total probability by Eq. (1):

$$P(X) = \int P(X, z)dz = \int P(X|z)P(z)dz \qquad (1)$$

During the production phase, the VAE attempts to maximize the probability of each X in the training set according to Eq. (1). Also, $P(X|z)$ is the probability function of the observed data in terms of the hidden variable, which specifies how to determine the distribution of the input data based on the sample distribution of the hidden variable. The fundamental concept underlying VAEs is to sample the values of the hidden variables z and derive $P(X)$ from them. Therefore, a new function $Q(z|X)$ is required to explain the distribution of z dependent on the value of X. In other words, z is sampled from an arbitrary distribution, and Q can be any distribution, such as the normal distribution, in order to compute $E_z \sim QP(X|z)$. To do this, we begin by matching $P(z|X)$ and $Q(z)$ based on the Kullbackâ€“Leibler divergence between $P(z|X)$ and $Q(z)$ for a desired Q:

$$D[Q(z)\|P(z|X)] = E_z \sim Q[\log Q(z) - \log P(z|X)] \quad (2)$$

Now, by applying the Bayesian rule to $P(z|X)$, Eq. (2) can be transformed into Eq. (3):

$$D[Q(z)\|P(z|X)] = E_z \sim Q\big[\log Q(z) - \log P(X|z) - \log P(z)\big] + \log P(X) \qquad (3)$$

In Eq. (3), the expression $\log P(x)$ is excluded from mathematics because it does not depend on z. Now, by rewriting Eq. (3), we come to Eq. (4):

$$\log P(X) - D[Q(z)\|P(z|X)] = E_z \sim Q[\log P(X|z)] - D[Q(z)\|P(z)] \quad (4)$$

Eq. (4) is known as the core of VAEs. In fact, the right side of the expression acts as a autoencoder, because Q encodes X in z, and P decodes it to reconstruct X [7].

The distribution type $Q(z|X)$ is specified in the next step. The proposed distribution is a normal distribution with parameters $\mu(X)$ and $\Sigma(X)$. In such a case, $Q(z|X)$ becomes $N(\mu(X), \Sigma(X))$, and $P(z)$ becomes $N(0, 1)$. Using the encoder, the means and covariances are predicted (close to the previous distribution) and then the decoder reconstructs the sample using the sample data. Since practicable sampling is not derivative and derivability is required to propagate the gradient backwards, a procedure called the reparameterization trick is used. For example, in a normal distribution with mean μ and standard deviation σ, sampling can be done as Eq. (5):

$$z = \mu + \sigma \odot \varepsilon \quad (5)$$

Here z is a random variable with normal distribution with mean 0 and standard deviation of 1 ($N(0, 1)$). With this change, the variable can be derived from z and $f(z)$ to the mean and standard deviation parameters. As a result, the problem of derivation and gradient propagation is solved. Figure 1 on the left shows the sampling operation before reparameterization and on the right, after reparameterization. In Figure 1, the circle symbolizes randomness and the rhombus symbolizes certainty. As it turns out, after reparameterization, z changes from random to definitive, making gradient propagation possible [19, 9].

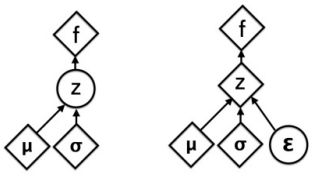

Figure 1: Orginal vs Reparameterized [19, 9].

3.2 Supervised Variational Autoencoder

Recently, in several studies, the effect of adding a classification layer attached to the hidden layer of autoencoders has been investigated [20, 4]. Le and his colleague theoretically and practically examine the effect of adding a classification layer to an autoencoder and call it a supervised autoencoder. This neural network pursues two goals simultaneously: The first is the reconstruction of samples, and the second is the classification of samples. Experiments also show that the supervised model not only has no defects in reconstruction but can also improve generalizability [20].

Moreover, researchers in [4] believe that unlabeled data is not only useful in reconstructing samples, but also improving classification performance. In addition to classification, labels also contribute to the reconstruction function.

4 Proposed Model

In this section, we present our proposed model for predicting the patient's return to the hospital or the patient's readmission after discharge.

Many researchers have tried to predict this issue in different ways. However, one of the most common methods is to predict the patient's return in the next few days. This period is usually equal to 30 days, and it becomes a matter of two categories: whether the patient will return to the hospital within 30 days of discharge. Although the output of the problem is defined, the inputs of the proposed models are different. Here we are going to make this prediction in the simplest possible way. Therefore, the model's input will be only the ICD codes (International Classification of Diseases) and the patient's demographic, such as age. Several advantages will result from this:

- This model can be transferred to hospitals in each country and region. Because it is not dependent on language and is trained with only a few diagnostic codes and the patient's demographic. Therefore, it can be practically implemented even in hospitals with other languages.
- Since it uses only ICD codes and does not interfere with the details and vital signs of the patient, it can be used in hospitals and areas with the low-order facilities.
- Avoid heavy preprocessing such as language processing operations on clinical notes.
- Time irregularity of data will not affect its performance. However, the timing of care will be considered.

As input to the model, two sequences of ICD codes are provided simultaneously. A third input provides the model with the patient's demographic. The output also has three parts. The first and second outputs are a sequence of codes reconstructed by the autoencoder, and the third output is a prediction of patient readmission.

4.1 The proposed model structure

In this section, we present the structure of our proposed model to solve the prediction problems. The proposed model consists of two autoencoders that are trained in parallel. Their hidden layer is sampled during the training and given as input to a classification. Thus, the error obtained by this model during training consists of 3 parts: two autoencoder reconstruction errors plus the classification error. The ultimate goal of the model is to minimize the sum of these three errors. Here the first and second errors are each equal to the sum of the reconstruction errors and the Kullbackâ€“Leibler difference between the hidden distribution learned and the prior distribution (Fig. 2), and the third error is the binary classifier error.

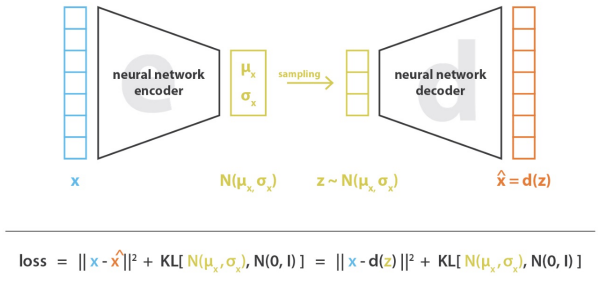

Figure 2: Computing loss in VAE

Figure 3 represents the proposed model. The proposed model consists of 3 primary parts.

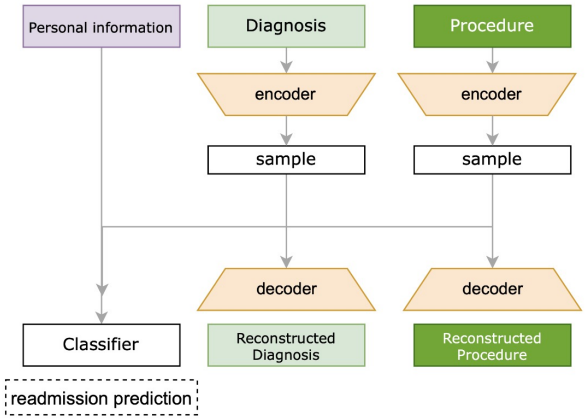

Figure 3: The proposed model

Part i: Procedure Autoencoder

This autoencoder receives the procedure codes in sequence. This structure is similar to an encoder-decoder for sentences. Each code is similar to a word, and each sequence is similar to a sentence. This autoencoder is responsible for representing and reconstructing the treating process. Since these sequences have a temporal property, the encoder uses recurrent units. This part is made up of three smaller components: the encoder, the decoder, and the hidden layer. The encoder and decoder are made of recurrent units, and the hidden layer consists of two vectors of mean and standard deviation. A representation of the procedure autoencoder can be found in figure 3.

Part ii: Diagnosis Autoencoder

This part is also similar to the procedure autoencoder. A sequence of codes is considered input, and the reconstructed sequence is considered output. This part is also made up of three smaller components: an encoder, decoder, and hidden layer. The encoder and decoder are made of recurrent units, and the hidden layer will consist of two mean vectors and a standard deviation. A representation of the diagnosis autoencoder can be found in figure 3.

Part iii: Classification

The last part of this model will be the binary classification. The classification input will be a sample of the hidden procedure coding layer, a sample of the diagnosis coding, and the patient's demographic. In fact, after merging these three vectors, one vector is given to the classification as the individual vector, and the prediction will be made. Figure 3 also shows the classification and its inputs.

5 Experimental Results

In this section, we give the experimental outcomes of the proposed model, as well as the outcomes of related algorithms. Before evaluating the suggested model, the Experimental Setup is described.

5.1 Experimental Setup

The dataset used in this study is called MIMIC-III. MIMIC-III is a large and free dataset that includes health and care data of more than 40,000 patients. This data set was collected from patients who were hospitalized in the intensive care unit of the Beth Israel Deaconess Medical Center from 2001 to 2012 [17].

The database contains demographic information, hourly vital sign measurements, laboratory results, treatment methods, drugs, clinical comments, imaging data, and death records (inside or outside the center).

MIMIC is applicable to a variety of fields, including epidemiology, clinical decision improvement, and the creation of electronic devices. MIMIC-III possesses three outstanding characteristics:

- It is freely accessible to researchers around the globe.
- It includes a very large and diversified population of ICU patients.
- It includes information such as vital signs, test findings, and prescriptions.

The relational database MIMIC-III has multiple tables. Identifiers, typically suffixed with "ID," connect these tables; for example, HADM_ID refers to unique hospital admission, and SUBJECT_ID refers to a unique patient. Tables with the prefix "D_" are dictionaries and contain definitions for identifiers. For example, each row of OUTPUT- EVENTS contains a single ITEMID that represents the measured concept but does not include the actual name of the drug. OUTPUTEVENTS and D_ITEMS can be connected to ITEMID to determine what an ITEMID represents.

MIMIC-III tables include several categories:

1. The first category is used to define and follow the patient's stay in the hospital (such as PATIENTS and ADMISSIONS)
2. The second category includes data recorded in the ICU (such as CHARTEVENTS and NOTEEVENTS)
3. The third category is tables that contain information about the hospital records system (such as CPTEVENTS and DIAGNOSES_ICD)
4. The fourth category is dictionary tables (such as D_CPT and D_ICD_DIAGNOSES)

A pre-processing step has been performed to prepare the data for entering the introduced model; it will be explained in the following.

1. **Shortening diagnosis and procedure codes:**

   As mentioned earlier, diagnosis and procedure codes are very long. On the other hand, the main code also includes unnecessary details, which ultimately reduces the model's generalizability. Therefore, only the first three digits of the ICD code are used.

2. **Build code sequences:**

   The codes of each admission are sequenced by inserting the space between the two codes.

3. **Unify pad sequence length:**

   All code sequences are padded to ten sequences. If the sequence length is more than ten, the first ten are selected for diagnosis and the last ten for treatment. Otherwise, it fills to zero until the number reaches ten.

4. **Eliminate elective admissions and infants:**

   Because the goal is to predict future emergency admissions, optional admissions are eliminated to anticipate only non-scheduled admissions. Also, admissions for infants are excluded from the set of tests.

5. **Changing genetic classification:**

   Genetic classifications were changed to more general categories for better generalizability. For instance, several groups of Guatemalans and Dominicans of the Latin race were classified as Latin.

6. **Change in marital status:**

   There were several cases for the status of isolated individuals, all of which generally referred to one type of category. So they became a single group.

7. **Convert attributes to One-Hot encoding:**

   Batch characteristics such as place of admission, type of reception, place of clearance were transformed into OneHot vectors.

8. **Normalization of numerical properties:**

   Numerical characteristics such as age and number of days of stay were normalized.

9. **Labeling:**

   All admissions that the patient returns to the intensive care unit within 30 days of discharge are labeled as one. Other admissions will be labeled zero. On the other hand, the patient may have died during discharge or within 30 days after discharge, in which case the patient will be removed from the data set. (Patients who die within 30 days of discharge will be labeled 1 for hospital testing).

5.2 Evaluation Metrics

In this section, we present the outcomes of running the model on our preprocessed data and compare them to the relevant prediction methods. The majority of prediction methods were tested using these standards. Predicting a binary classifier for positive and negative classes or one and zero can be in two ways:

- Identify exactly the category type for each sample. The output should be zero or one, and the class of that sample should predict exactly.

- The output is likely to be positive or class one. In this case, the output is numerically between zero and one, which indicates the probability that the sample belongs to a category.

In this latter scenario, the class matching to the sample is typically predicted by establishing a threshold, which is typically 0.5. Prediction with the second method gives better control over the model's performance. As a result, the predicted classes of the model can be changed by changing the threshold. By changing this threshold, the model evaluation criteria also change. Therefore, it is possible to create multiple confusion matrices and calculate different criteria several times, which would not be rational.

Two criteria widely employed in this field's study will be introduced. The first criterion is the area beneath the chart, while the second is the F-score:

- ROC Curve: The ROC curve is a graph depicting the false positive rate versus the true positive rate of the model for various thresholds. The horizontal axis indicates the proportion of false positives, while the vertical axis reflects the proportion of real positives. A true positive rate is calculated by dividing the total number of accurate forecasts by the sum of true positives and false negatives. The rate of true positives is also known as sensitivity or recall [22]. By calculating false positives and true positives for different threshold values, a curve is drawn from the bottom left to the top right, which is called the ROC curve. The area below this graph, a number between zero and one, evaluates the model's performance. The larger the level below this graph, the better the model performance in classifying the data. The classifier that does not have the power to distinguish between positive and negative classes is a diagonal line from point $(0, 0)$ to point $(1, 1)$, and the area below the graph is 0.5. Although ROC graphs are commonly used to evaluate classifiers in the presence of class imbalance, they have a drawback: under class rarity, when the problem of class imbalance is accompanied with a small sample size of minority occurrences, the estimates may be incorrect.

- F-score: This criterion is the harmonic mean of the other two criteria, precision and recall. Precision is the true positive divided by true positive plus false positive [11].

5.3 Evaluation of the proposed model

In this section, various experiments will be performed, and the results will be presented. Initially, several experiments are performed to find some suitable parameters for the model. The effect of input characteristics on the model's performance is then investigated. The performance of the model handler is then explicitly compared with older or so-called traditional machine learning models. Finally, the results of the model are compared with previous studies. To perform these experiments, 85% of the data is used for training and 15% for testing. 10% of training data is also used for validation. RMSprop with a training rate of 5e−5 is used as a model optimizer. The criterion used for comparison will also be the area below the graph.

In the first part, the size of the hidden layer vector, and in the second part, the effect of the standard deviation of the sampling layer on the model's performance is investigated.

In the first experiment, different values are used for the size of the hidden layer of two VAEs to determine the effect of size. The results are presented in Table 1. The first column shows the size of the vector, that is, the parameter under consideration. In order to ensure the test conditions, each test is performed in 4 different randomly distributed data modes, and the results are averaged. Each random mode divides the data into training and testing sections differently. It also (testing in several different data distribution modes) increases the model's assurance of generalizability and reduces the chance of random results. And, in the second column, the average and the final result are displayed with a 95% confidence interval.

The results shown in Table 1 indicates that the best value for the hidden layer size is 256. So by the end of the experiments, the hidden layer's size will be the same.

Table 1: Determining the appropriate value for the size of the hidden layer.

Number of Neurons	**Final Result**
32	$67.8 \pm 0.6\%$
64	$69.45 \pm 0.5\%$
128	$70.88 \pm 0.2\%$
256	$72.05 \pm 0.4\%$

In the second experiment, the standard deviation used for sampling is tested. In this case, as in the previous experiment, the test results will have the same conditions. The results of these experiments are shown in Table 2. The first column shows the standard deviation of the sampling, and the second column shows the final result, which is presented in the 95% confidence interval. The table shows that the values 0.1 and 0.01 have similar and close functions. Here, a standard deviation of 0.1 will assist in the generalizability and sampling of relatively more extensive space.

As mentioned earlier, the input features of the model fall into three categories: procedure sequence, diagnosis

Table 2: Determining the appropriate value for the standard deviation of the sampling.

σ	**Final Result**
0.01	$76.3 \pm 0.1\%$
0.1	$76.45 \pm 0.4\%$
1	$71.95 \pm 0.6\%$

sequence, and individual characteristics. In this section, we will examine the effect of each of these features on the model's performance by performing experiments. Table 3 shows the results of these experiments. The zero or one column of each attribute indicates that it is in the model or that attribute is removed from the model.

Table 3: Impact of input features on model performance.

Procedure Sequence	**Diagnosis Sequence**	**Individual Characteristics**	**Final Result**
1	**0**	**0**	$69.25 \pm 5.2\%$
0	**1**	**0**	$72.5 \pm 2.2\%$
0	**0**	**1**	$73.3 \pm 0.8\%$
1	**1**	**0**	$75.65 \pm 0.2\%$
1	**0**	**1**	$73.2 \pm 0.4\%$
0	**1**	**1**	$74.2 \pm 0.8\%$
1	**1**	**1**	$76.1 \pm 0.2\%$

In the performance of single-feature models, the best results are related to individual or static features. On the other hand, the model's performance has been outstanding and acceptable, even with removing individual characteristics and only using procedure and diagnosis sequences. In fact, among the models with two features, the best performance combines two features of diagnosis and treatment sequences. With all these interpretations, the model's performance in the case that uses all three mentioned features has a significant advantage over other test modes.

There are two sorts of output views for models. In the first type, the patient returns to the hospital after 30 days, whereas in the second type, the patient returns to the intensive care unit. In this study, all tests were focused on returning to the ICU. In the two parts that follow, the results will be contrasted with those of others.

In the first view, that is, returning to the ICU and on the same dataset, the research of [25] is available. The comparison of the results of the mentioned research with the present research is given in Table 4. The SVAE model is our proposed model. Pakbin-1 is logistic regression and Pakbin-2 is XGBoost. As it turns out, our proposed model performs far better than the model proposed by [25]. In our model test, the threshold for determining whether the output label is zero or one to calculate the F-score is 0.35. If the predicted probability for the label is less than 0.35, that label is 0, and if it is above 0.35, the label is considered 1.

Table 4: Comparing the performance of the proposed model with other approaches (ICU).

Model	**Accuracy**	**F1-Score**
SVAE	$76.28 \pm 0.3\%$	0.45
Pakbin-1 [25]	73%	0.34
Pakbin-2 [25]	75%	0.37

In the second view, and for other comparisons, the patient's return to the hospital is chosen as the label. In this case, our method is compared with the research of Balan et al. In the research of Balan et al.; several methods have been tested, the names of which can be seen in Table 5. DeepR is one of the convolutional neural network models previously introduced by academics. Balan et al. were able to increase the performance of this model, however, by adding static features and making small adjustments. The Balan model is one of the better models in the comparison table and is presented by the same researcher. Our proposed model performs better than logistic and recursive regression models but worse than the modified Balan and DeepR models. Of course, it should be noted that the Balan model uses more features as input, while our proposed model achieves such a function only with a series of diagnoses and treatments and a limited number of individual features.

Table 5: Comparing the performance of the proposed model with other approaches (Hospital).

Model	**Accuracy**
SVAE	$71.3 \pm 0.4\%$
Logistic Regression	70%
RNN	68.9%
DeepR (Non-Static)	68%
DeepR (Static)	74%
Balan (Non-Static) [3]	67%
Balan (Static) [3]	74%

6 Conclusions and Future Work

The proposed method in this paper is an SVAE that predicts the outcome of the problem and can be used to reconstruct the diagnosis or procedure sequences. Our proposed method has acceptable performance but can also have better results by adding more features.

Since the proposed model has simple inputs, the results can be improved by adding more features. On

the other hand, using other deep models can be another solution. One of the limitations of this study was the amount of RAM consumed on the Google Colab. Therefore, with the help of more advanced systems, more available data can be used. For example, the use of clinical notes and laboratory results can effectively aid in increasing the quality of prediction. On the other hand, some researchers have taught weights for the hidden layer that can be used with more RAM. The presence of a medical specialist can also be helpful in future research. In terms of output, the proposed model also includes sequences reconstructed by the autoencoder, which can be used to predict the diagnosis and treatment of patients or as assistant systems for medical diagnosis. The simplicity of the input variables and the independence of this model from the natural language provide the ability to implement it with minor facilities.

References

[1] K. Alaboud, I. E. Toubal, B. M. Dahu, A. A. Daken, A. A. Salman, N. Alaji, W. Hamadeh, and A. Aburayya. The quality application of deep learning in clinical outcome predictions using electronic health record data: A systematic review. *South Eastern European Journal of Public Health*, 12:9–23, 2023.

[2] A. Ashfaq, A. Sant'Anna, M. Lingman, and S. Nowaczyk. Readmission prediction using deep learning on electronic health records. *Journal of biomedical informatics*, 97:103256, 2019.

[3] M. Balan U, M. Gandhi, and S. Rammohan. Predicting unplanned hospital readmissions using patient level data. 2021.

[4] F. Berkhahn, R. Keys, W. Ouertani, N. Shetty, and D. Geißler. Augmenting variational autoencoders with sparse labels: A unified framework for unsupervised, semi-(un) supervised, and supervised learning. *arXiv preprint arXiv:1908.03015*, 2019.

[5] Y. Cheng, F. Wang, P. Zhang, and J. Hu. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM international conference on data mining*, pages 432–440. SIAM, 2016.

[6] E. Craig, C. Arias, and D. Gillman. Predicting readmission risk from doctors' notes. *arXiv preprint arXiv:1711.10663*, 2017.

[7] C. Doersch. Tutorial on variational autoencoders. *arXiv preprint arXiv:1606.05908*, 2016.

[8] R. Dwivedi, D. Mehrotra, and S. Chandra. Potential of internet of medical things (iomt) applications in building a smart healthcare system: A systematic review. *Journal of Oral Biology and Craniofacial Research*, 12(2):302–318, 2022.

[9] J. Egger, C. Gsaxner, A. Pepe, K. L. Pomykala, F. Jonske, M. Kurz, J. Li, and J. Kleesiek. Medical deep learning—a systematic meta-review. *Computer Methods and Programs in Biomedicine*, 221:106874, 2022.

[10] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[11] C. Goutte and E. Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer, 2005.

[12] S. Graham, C. Depp, E. E. Lee, C. Nebeker, X. Tu, H.-C. Kim, and D. V. Jeste. Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21(11):1–18, 2019.

[13] G. Harerimana, J. W. Kim, H. Yoo, and B. Jang. Deep learning for electronic health records analytics. *IEEE Access*, 7:101245–101259, 2019.

[14] J. Huang, C. Osorio, and L. W. Sy. An empirical evaluation of deep learning for icd-9 code assignment using mimic-iii clinical notes. *Computer methods and programs in biomedicine*, 177:141–153, 2019.

[15] S. Jain, R. Mohammadi, and B. C. Wallace. An analysis of attention over clinical notes for predictive tasks. *arXiv preprint arXiv:1904.03244*, 2019.

[16] H. Javidi, A. Mariam, L. Alkhaled, K. M. Pantalone, and D. M. Rotroff. An interpretable predictive deep learning platform for pediatric metabolic diseases. *Journal of the American Medical Informatics Association : JAMIA*, 31:1227 – 1238, 2024.

[17] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[18] G. KH, W. L, Y. AYK, D. YY, A. LSY, P. HMN, L. K, Y. JJL, and T. GYH. Prediction of readmission in geriatric patients from clinical notes: Retrospective text mining study. *J Med Internet Res*, 23(10), 2021.

[19] D. P. Kingma and M. Welling. An introduction to variational autoencoders. *arXiv preprint arXiv:1906.02691*, 2019.

[20] L. Le, A. Patterson, and M. White. Supervised autoencoders: Improving generalization performance with unsupervised regularizers. *Advances in neural information processing systems*, 31, 2018.

[21] H. Li, X. Li, M. Ramanathan, and A. Zhang. Identifying informative risk factors and predicting bone disease progression via deep belief networks. *Methods*, 69(3):257–265, 2014.

[22] D. K. McClish. Analyzing a portion of the roc curve. *Medical decision making*, 9(3):190–195, 1989.

[23] M. Z. Nezhad, D. Zhu, N. Sadati, and K. Yang. A predictive approach using deep feature learning for electronic medical records: A comparative study. *arXiv preprint arXiv:1801.02961*, 2018.

[24] P. Nguyen, T. Tran, and S. Venkatesh. Resset: A recurrent model for sequence of sets with applications to electronic medical records. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2018.

[25] A. Pakbin, P. Rafi, N. Hurley, W. Schulz, M. H. Krumholz, and J. B. Mortazavi. Prediction of icu readmissions using data at patient discharge. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4932–4935. IEEE, 2018.

[26] T. Pham, T. Tran, D. Phung, and S. Venkatesh. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*, 69:218–229, 2017.

[27] Y. Ramakrishnaiah, N. Macesic, G. I. Webb, A. Peleg, and S. Tyagi. Ehr-ml: A generalisable pipeline for reproducible clinical outcomes using electronic health records. In *medRxiv*, 2024.

[28] L. Rasmy, Y. Wu, N. Wang, X. Geng, W. J. Zheng, F. Wang, H. Wu, H. Xu, and D. Zhi. A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous ehr data set. *Journal of biomedical informatics*, 84:11–16, 2018.

[29] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, 113:103627, 2021.

[30] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny. The application of deep learning in analysing electronic health records for improved patient outcomes. *International Journal of Intelligent Systems and Applications in Engineering*, 12:223—-228, 2024.

[31] B. Shickel, T. J. Loftus, L. Adhikari, T. Ozrazgat-Baslanti, A. Bihorac, and P. Rashidi. Deepsofa: a continuous acuity score for critically ill patients using clinically interpretable deep learning. *Scientific reports*, 9(1):1–12, 2019.

[32] F. Wang and A. Preininger. Ai in health: state of the art, challenges, and future directions. *Yearbook of medical informatics*, 28(01):016–026, 2019.

[33] W. Wang, Y. Feng, H. Zhao, X. Wang, R. Cai, W. Cai, and X. Zhang. Mdpg: a novel multi-disease diagnosis prediction method based on patient knowledge graphs. *Health Information Science and Systems*, 12:1–18, 2024.

[34] F. Xie, H. Yuan, Y. Ning, M. E. H. Ong, M. Feng, W. Hsu, B. Chakraborty, and N. Liu. Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *Journal of Biomedical Informatics*, 126:103980, 2022.

[35] J. Xu, X. Xi, J. Chen, V. S. Sheng, J. Ma, and Z. Cui. A survey of deep learning for electronic health records. *Applied Sciences*, 12(22), 2022.

[36] P. Yadav, M. Steinbach, V. Kumar, and G. Simon. Mining electronic health records (ehrs) a survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40, 2018.

[37] D. Zhao, Y. Shi, L. Cheng, H. Li, L. Zhang, and H. Guo. Time interval uncertainty-aware and text-enhanced based disease prediction. *Journal of Biomedical Informatics*, 139:104239, 2023.