View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Central Archive at the University of Reading

# *Predicting risk of hospital readmission for comorbidity patients through a novel deep learning framework*

Conference or Workshop Item

Published Version

Creative Commons: Attribution-Noncommercial-No Derivative Works 4.0

Open Access

Dashtban, M. and Li, W. (V.) (2020) Predicting risk of hospital readmission for comorbidity patients through a novel deep learning framework. In: 53rd Hawaii International Conference on System Sciences, 7-10 Jan 2020, Maui, Hawaii. Available at http://centaur.reading.ac.uk/86410/

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

Published version at: http://hdl.handle.net/10125/64137

All outputs in CentAUR are protected by Intellectual Property Rights law, including copyright law. Copyright and IPR is retained by the creators or other copyright holders. Terms and conditions for use of this material are defined in the End User Agreement.

University of Reading

www.reading.ac.uk/centaur

**CentAUR**

Central Archive at the University of Reading

Reading's research outputs online

# Predicting risk of hospital readmission for comorbidity patients through a novel deep learning framework

M Dashtban[1,2]
[1]Informatics Research Centre, Henley Business School, University of Reading,
[2]Informatics, Royal Berkshire NHS Foundation Trust, UK
Dashtban.edu@gmail.com

Weizi Li
Informatics Research Centre, Henley Business School, University of Reading
weizi.li@henley.ac.uk

## Abstract

*Hospital readmission is widely recognized as indicator of inpatient quality of care which has significant impact on healthcare cost. Thus, early recognition of readmission risk has been of growing interest in various hospitals. Additionally, there has been growing attention to provide better care to patients with more complications, whose care would impact the quality of care in multiple directions. To this regard, this research specifically targets comorbidity patients i.e., the patients with chronic disease.*

*This research proposes a novel deep learning-framework termed SDAE-GAN. The presented approach consists of three phases. Firstly, various groups of variables from heterogeneous sources are collated. These variables mainly include demographic, socioeconomic, some statistics about patient's frequent admissions and their diagnosis codes. Then, more processing applies dealing missing values, digitization and data balancing. Afterwards, stacked denoising auto-encoders function to learn underlying representation; and technically to forms a latent space. The latent variables then are used by a Generative Adversarial Neural Networks to evaluate the risk of 30-day readmission. The model is fine-tuned and being compared with state-of-the-arts. Experimental results exhibit competitive performance with higher sensitivity.*

## 1. Introduction

Hospitalized readmission has been receiving growing attentions [1] because of its implications on cost and quality of care in the most recent decade. Professional experts in the field mostly believe a remarkable number of readmissions are truly preventable [2]–[5]. Moreover, the recent focus on readmissions in some countries, including the United States, Germany, Switzerland, and England, underlies a much more global concern about patients' safety [6]. For policymakers, reducing the rates of readmission is considered a key issue to improve patient's outcomes and contain hospital costs [5], [7].

Although numerous reasons may lead to the occurrence of 30-day hospital readmission, recent studies have shown that the increased risk of readmission is linked to comorbidities [6], [8]. The evidence shows that higher comorbidity has been shown to be associated with an increased risk of readmission [9]–[11]. The comorbidity related issues will take on greater importance as a growing percentage of the world's population becomes older and the incidence of comorbidities rises. However, hospital readmission represents a multifaceted problem and the complexity of care and patient's comorbidity condition hinders deeper understanding of readmission patterns and relatively few studies have looked at this. Therefore, a better understanding of the causes and patterns of readmissions in patients with common comorbidities may lead to more accurate prediction, targeted and successful interventions.

One of the strategies to reduce the unplanned hospital readmission rate is the application of predictive models to identify patients at high risk for readmission. Preventive approaches can then be developed and applied to target the identified high-risk patients. However, the performance of traditional risk predictive model for readmission are poor and inconsistent according to the systematic review by [12]. The limited applicability of hospital readmission risk predictive models is partly due to the lack of data quality and the robustness of the statistical model. In the meantime, with new technologies and automations huge amount of data have been created in different domain especially in healthcare and genomics [13]–[15]. However, utilization of large amount of data has been particularly a great challenge for different machine learning and AI application. There are many common and difficult challenges such as high-dimensionality, heterogeneity,

URI: https://hdl.handle.net/10125/64137
978-0-9981331-3-3
(CC BY-NC-ND 4.0)

sparseness, incompleteness, random errors, and systematic biases [16]–[18]. To deal with such challenges depends on the problem, various kinds of kinds of statistical and machine learning methods have been employed.

Meanwhile, deep learning methods revealed some promising results in handling noise through representation learning [19]. Hence, such learning methods have attracted many researchers and institutions in clinical research tasks which were quite difficult to solve [20] [21]. Although a series of excellent work have been conducted in seek of novel deep learning solutions in different healthcare applications, there is very few deep learning-based researches available to predict readmission risk for comorbidity patients. Furthermore, the challenging nature of EPR such as inherent noises, and missing value make it extremely difficult to apply most deep learning models for accurate prediction [22].

Recently, the development of generative adversarial network (GAN) [23] provides a new capability to provide more robust model to noisy data with considerable missing values. In this study, we employ the full power of deep learning by introducing a new methodology comprising of representation learning and GANs. Additionally, this research work specifically targeted chronic comorbidity patients although provide the applicability for using across whole board of specialties. We believe, studding different cohort of patients separately and take proper course of actions accordingly will help to provide more robust and applicable solution in real-life environment. Many successful applications can be seen in the most recent decade focusing on specific specialty, specific cohort of patients, rather considering all problem together. In this study, Comorbidity patients were identified based on Charlson Chronic Comorbidity Indexes. These patients have growing complications over time; and consequently, require more attention.

Furthermore, from machine learning perspective, dealing with different comorbidity codes over time, itself, is a time-series problem. Each patient has a dynamic sequence of codes and time besides other variables. For each sequence, technically speaking, based on machine learning perspective, there are some important facts in connection with readmission, which must be considered:

- The readmission might happen because of other complication that was already existed in long time ago
- In some episodes (time intervals), there are missing codes which possibly impacted the care and consequently readmission
- Patients with quite more history in database have more records which create a potential sparsity in data space. Such sparsity, evidently, affects negatively the learning process of various classification or probabilistic models.
- Repetitive events in a sequence can *locally* form a bias for a learning method.

To address all these issues, we employed a simple, yet efficient strategy. A table was created by applying data mining techniques to have all comorbidity codes in one place for each record. A time window was applied and the selected codes then augmented to each patient profile vector. This vector was used as the input of auto encoders. After, pre-training auto encoders, the obtained latent variables were employed as input to a GAN model.

In summary, this study proposes a novel framework for predicting the risk of re-admission mostly suitable for patients with multiple complications.

## 2. Literature review

Predicting hospital readmission risk is of great interest to identify which patients would benefit most from care transition interventions, as well as to risk-adjust readmission rates for the purposes of hospital comparison [12]. Readmission risk assessment could be used to help target the delivery of these resource-intensive interventions to the patients at greatest risk. Ideally, models designed for this purpose would provide clinically relevant stratification of readmission risk and give information early enough during the hospitalization to trigger a care intervention, many of which involve discharge planning and begin well before hospital discharge. Models designed for these purposes should have good predictive ability; be deployable in large populations; use reliable data that can be easily obtained; and use variables that are clinically related to and validated in the populations in which use is intended [12], [13]. According to recent review conducted by [24], the utilization outcome of existing readmission prediction models include all-cause admissions such as [25], cardiovascular-related disease including pneumonia such as [26], medical/internal medicine conditions such as [27], surgical conditions such as [28] and mental health conditions such as [29]. There is no model developed for readmission risk for all comorbidity patients.

Furthermore, all those models are traditional statistical model based using clinical/medical records data. They are hypothesis driven and repetitively assess the predictive abilities of the same set of biomarkers as predictive features. The performance of the applied existing models was inconsistent and due to the poor performance, there is limited applicability to be used in the hospital [24]. The research by [30] attempts to develop a data-driven, electronic-medical record-wide (EMR-wide) feature selection approach and subsequent

machine learning to predict readmission probabilities. They designed a multistep modeling strategy using the Naïve Bayes algorithm with encouraging results and revealed the utility of such data-driven machine learning in predicting readmission for heart failure cohort.

The predictive analysis study includes two key components: feature learning and classification. The application of deep learning in these two areas has recently gained unprecedented popularity [13]. Deep learning classification from EPR is initially studied to predict disease progression. For example, [31] applied recurrent neural network in longitudinal time stamped EPR to predict diagnoses and medications for the subsequent visit by building a generic temporal predictive model that covers observed medical conditions and medication uses, followed by the development of specific heart failure prediction model. The other research by [32] utilized a long-short memory (LSTM) method to model disease progression and predict future risk. Recently more attention is received in using deep learning method to predict the risk of readmission. For example, these researches [33], [34] applied convolutional neural network methods to detect and combines predictive local clinical motifs to stratify the risk of readmission. Authors in [35] developed an artificial neural network model to predict all cause risk of 30-day hospital readmission and [36] developed a hybrid deep learning model that combines topic modelling and recurrent neural network (RNN) to embed clinical concepts in short-term local context and long term global context to predict readmission. The research by [37] further developed a scalable deep learning model using RNN for prediction across multiple centers without site-specific data harmonization which is validated in readmission task.
Aside from those researches, [38] compares various deep learning-based models for predicting early hospital admissions. They found that the performance of existing models is insufficient for practical applications as the models generally fit to homogeneous patient subgroups. This leads to attentions of challenging nature of EPR such as inherent noises, and missing value that make it extremely difficult to apply most existing matured models for prediction [22].

Another challenge in EPR data processing is the class imbalance problem. There literally significantly higher number of records for normal people than those whose suffering from a specific disease [39]. Therefore, it is necessary to develop the learning method which is more robust against the class imbalance problem. Such challenges lying in EPR prevent many deep learning methods from exerting their strength in predictive analytics.

Recent development of generative adversarial network (GAN) caught attention [40] and have been mainly used on image, video and text data to learn useful features with better understandings, and robustness for incomplete and imbalanced data. GAN simultaneously trains a deep generative model and a deep discriminative model, which captures the data distribution and distinguishes generated data from original data respectively, as a mini-max game. Although there are attempts to apply GANs in EPR directly to predict disease [22], [41], there is no research on readmission prediction benefited from GANs methods. In this research, a novel deep learning approach proposed integrating both autoencoders and GANs in a learning framework. This framework actually leverages the true potential of deep learning which is representation learning and classification through adversarial learning. This study ultimately reveals completive performance with state-of-the-are models while achieving the highest sensitivity.

## 3. EPR processing and feature representation

The quality of the data in the hospital is of crucial importance as the accuracy and fairness of the algorithms are closely linked to the data they are being fed. However, the quality of coding for comorbidities has been a challenge for analytics and hospital pay by results because lack of adequate information captured consistently in the source documents. This has led to comorbidity information missing in some spells with negative implications such as resource planning and monitoring, full understanding of the complexity of condition and the utility of predictive models.

Different with exiting studies, we started our predictive modelling from discovering comorbidity codes that were missed the coding process from EPR. The information of over 130K comorbidity patients over half a million records (over 2 million records considering all patients) were extracted, beginning in 2010 and going through end of 2017 in one of the largest of hospitals in Berkshire, UK.

To identify all the relevant codes for comorbidity patients, we created a reference table in our database for Charlson comorbidity index [42]. Then, a time-code table was collated which enabled us looking in the real whole journey of patients in EPR. This table also included other information associated with patient profile like sociodemographic variables. A binary label then was added to the table to indicate 30-day readmission. Thereafter, only the records of 30-day readmission was filtered and selected for further evaluation. It is worth noting, filtering 30-day readmission does not mean that the comorbidity codes were filtered. All the codes were included as clinicians usually expect the chronic disease will last for a very long time.

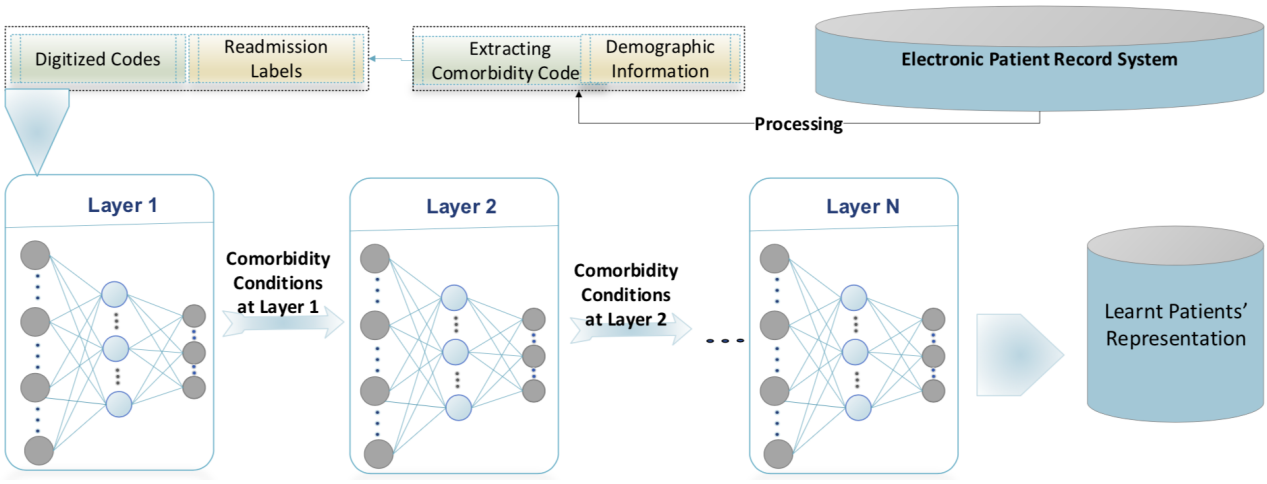


**Figure 1. Patients feature learning process using Stacked denoising-Auto-Encoders (SDAE)**

Dealing with missing values in other variables like ethnicity, gender and age, we endeavor to find such information from reference tables in other databases. Collating data from different databases and Tables, assess for any conflicts/duplication/alternatives, and appropriately join them to the resulting table were quite a tedious and challenging task. Finally, for filing out the remaining missing-values we employed mode statistic for categorical variables and K-nearest neighbor for continuous one. The list of variables used in this study is listed in Table 1.

**Table 1. Variables employed in this study**

| Group | Variables |
|---|---|
| Demographic | Age, Sex |
| Discharge | Month, Discharge type |
| Socioeconomic | Deprivation Index, Ethnicity, Marital status |
| Diagnostic codes | Charlson Comorbidity Indexes |
| Admission | Admission type, Admission source |
| Statistics | Average number of admissions per year, Average length of stay |

## 4. Patients feature learning using Stacked denoising-Auto-Encoders

To overcome the challenging nature of EPR especially for large data dimension, noise and sparseness, feature learning and extraction exploiting de-noising auto-encoders is a robust way to deal with such issues [43]. It is worth noting, the auto-encoder, itself, will help constructing the real data manifold. In this context, dealing with missing values and approximating them in the previous section can be partly addressed by auto encoders. In other words, we may not need to try comprehensively and precisely calculate the missing values. A good approximation employing simple statistics in large scale data perhaps is an efficient way; specifically, when a representation learning method is supposed to be employed on top of that.

A denoising autoencoder (DAE) is simply a neural network with one hidden layer that should be trained to reconstruct a clean version of input X from a corrupted/current version of x'. It is accomplished by a so-called encoder that is a deterministic mapping from an input vector x into hidden representation y.

$$f_\theta(\mathbf{x}) = s(\mathbf{W}\mathbf{x} + \mathbf{b}) \qquad (1)$$

where the parameter $\theta$ is $(\boldsymbol{W}, \boldsymbol{b})$, $\boldsymbol{W}$ is a weight matrix and $b$ is bias vector.

In stacking of DAE as demonstrated in Figure 1, the auto-encoder layers are placed on top of each other. Each layer is trained independently ('greedily') and then is stacked on top of previous one. In denoising autoencoders, the loss function is to minimizing the reconstruction loss between a clean X and its reconstruction from Y [44]. A decoder is then used to map the latent representation $y$ into a reconstructed ('repaired') vector such as g:

$$g_{\theta'}(\mathbf{y}) = s(\mathbf{W}'\boldsymbol{y}+\mathbf{b}') \qquad (2)$$

Training process starts with per-training the first hidden layer fed the training samples as input, training the second hidden layer with the outputs flowing from the first hidden layer, and so on. The auto-encoders have widely used for feature reconstruction. Nevertheless, it could be used as dimensionality reduction [45]. It would

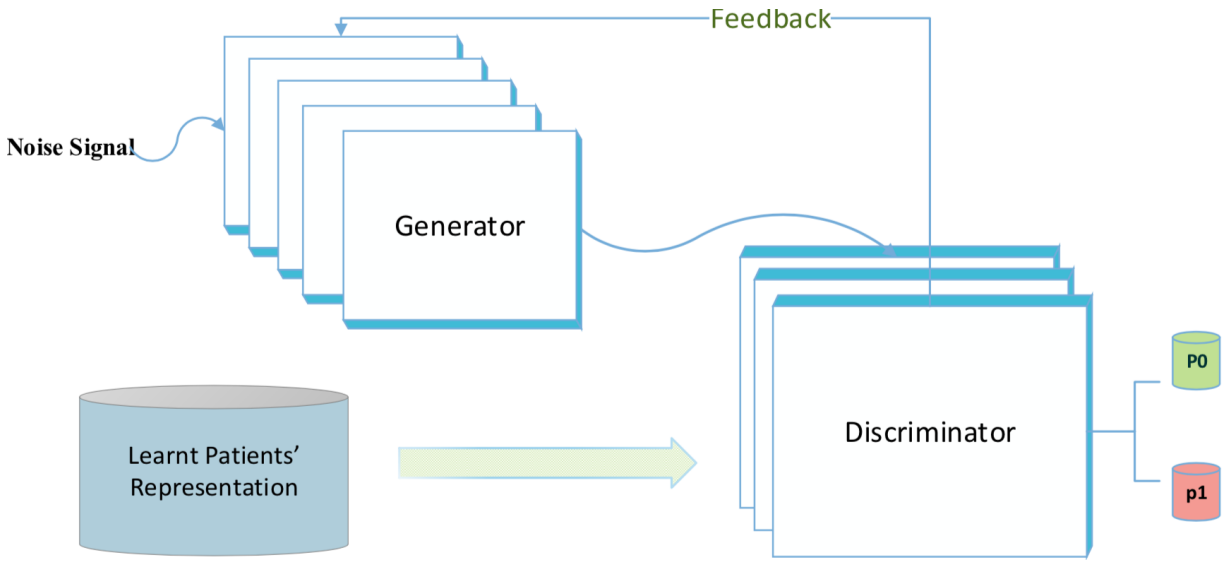


**Figure 2. Readmission risk prediction using Generative Adversarial Networks, p1 indicates the probability of readmission, p0 indicate the probability of no future readmission**

specifically effective for reducing the model complexity and extracting salient features when the input data is highly sparse. In our study, the input vector is quite sparse as there will be high dimensions of ICD codes but each patient will only have certain amount of ICD codes for their comorbidity conditions and the ICD codes' dimension value is 1 or 0 depending on patients' comorbidity. Learning highly non-linear and complicated patterns such as the relations among input features is one the prominent characteristics of SAE [46]. Another important feature of SAEs is the potential to learn the latent representation in data manifold. Authors in [47] showed that the learnt representation by autoencoders has connections to the intrinsic dimensionality of data. When the number of hidden layer nodes is restricted to be less than the number of original input dimension, a compressed representation of patient features is achieved.

The proposed model is demonstrated in Figure 1. Just after preparing the input tables in the last section, the SAE is applied to learn the higher level of representation to create the latent space. The latent variables instead of original features will be employed afterward. In this study, the auto encoders with the same structure discussed in our previous work for outpatient appointment prediction was employed [48]. Along with other studies, we empirically found a three-layer autoencoders performs more efficiently. We used many trial and errors to tune the parameters of the models upon validation data. Actually, only 1o% of whole data were employed as validation. After pre-training phase and obtain latent variables, different classifiers can be used over the extracted variables. We used GANs and other well-known classifiers to evaluate the final performance.

Furthermore, this approach has a specific advantage specifically for GANs model. The GANs generally works upon continuous data space. The latent variables are all continuous variables which are theoretically in line with such assumptions of GANs. On the other hand, considering lower dimensionality, the generator in GANs needs to learn the distribution of lower number of compact features which is evidently more efficient.

## 5. Generative Adversarial Networks

Adversarial models [40] are specific class of generative models [49] formalized as a competitive process between two players with distinct objectives. These players are represented by two neural network models with different architectures: the generator and discriminator (as shown in Figure 2). The discriminator role in GAN is like an artist that draws real images whereas the generator attempts to create fake versions from scratch. In this scenario, discriminator tries to produce better images from real world by analyzing not only the observed variables but also the noise images generated from outside source. Statistically speaking, the discriminator captures the conditional distribution of data given evidences while the generator is trying to learn the intrinsic distribution of data just from a random noise as well as the feedback from discriminator. Considering readmission, a statistical or probabilistic model may only need to look for similar cohort of patients in terms of similarity over a few variables; or

**Table 2. Comparing prediction results with other methods**

| Methods | Measures | | | |
|---|---|---|---|---|
| | F1-Score | Sensitivity | Specificity | AUC-ROC |
| SDAE-SVM | 0.34 | 0.24 | 0.60 | 0.55 |
| SDAE-Random Forest | 0.54 | 0.45 | 0.68 | 0.67 |
| SDAE-GAN | **0.66** | **0.55** | 0.83 | 0.65 |
| Fully-Connected Networks | 0.49 | 0.35 | 0.81 | 0.63 |
| Fully-Connected Networks [35]-500 features | 0.32 | 0.22 | 0.61 | **0.77** |
| CSDNN [34] - GHWs Dataset | 0.44 | 0.26 | **0.89** | 0.70 |
| - OPR Dataset | 0.64 | 0.49 | 0.87 | 0.73 |

* a bold number indicates the highest figure in its column.

probabilistic models mainly exploit frequency distributions to approximate probability of occurrence like [50]. Despite that, an adversarial model seeks for complex interdependencies among all sampled variables. Interestingly, unlike most models, the training process in GAN does not include a global loss function to be minimized. Instead, these models are supposed to reach an equilibrium point where no competitors could improve itself.

The learning in GAN is performed in two phases. Firstly, a noise which is produced from normal distribution will be fed into generator. The generator, then, attempts to produce an input for discriminator which resembles the real data. In this scenario, the discriminator has two responsibilities. It should learn how to distinguish between real and noise data while simultaneously realize the probabilities of occurring readmission based on highlighted comorbidities. In our case, the discriminator is basically a three-class classifier differentiating among admission-readmission classes if it realized the input as a real data and the noise class otherwise. In our implementation, we used softmax with three-class output. Thus, the generator produces three probabilities, sum of them equals 1, for three classes. The probability value close to zero means the data is fake/noise. It has been demonstrated [51] such discriminator extrapolates better on the test data than a basic classifier since it deals with more data patterns than a regular classifier does in a completely supervised manner. The Softmax [52] with three class outputs, Kernel size 5, was used in our implementation. Our implementation is based on the original implementation of DCGAN [22], [51]. The main purpose of using generator is to further robustness of the discriminator in the training process.

The generator, just after receiving feedback from the discriminator, would be optimized. It means that the generator attempts to learn the structural patterns [51] of training data to produce some samples as close to training samples as possible. The new input of generator would feed into the discriminator again. This optimization process continues to finally the discriminator failed to distinguish real and fake data. It happens when the generator is highly learnt various patterns while at the same time, the discriminator learns to differentiate between two data classes i.e., the readmission/admission.

## 6. Experimentation

Ultimately after preprocessing step and removing duplicating and repetitive records, there remained about 133K distinct inpatients with 465K records out of which approximately 243K records were readmissions (the target class). As both classes should be included in modeling, all records were pre-processed. Some patients have many readmissions even 100 and more. These repetitive readmissions make the machine learning models highly biased toward these patients. Hence, an indicative flag was created pinpointing only the first record of a sequence of repetitive readmissions. to this context, a sequence could be of any length from 1 (singular readmissions or the negative class, both were considered as sequences of length 1) to infinity. Then, those records marked by this flag were included in the learning process.

This strategy created a highly balanced dataset comprising of 47K and 38K records for both classes i.e., non-readmission and readmission, respectively. For a consistent analysis, this procedure applied to all data and subsequently in later phase, the training samples were randomly separated from testing and validating samples. The categorical features were dummy coded and continuous features were normalized and centered. A total of 1043 features were finally produced. From the resulting tables in data processing stage, a matrix and a binary readmission label vectors were created. As cross validations were computationally extensive, conventional strategy was followed to split the dataset

into three parts dividing 70% as training data, 20% test, and 10% for parameter tuning of the model. This 10% of samples were not employed anymore in other phases like testing the final model.

Experimentally a three-layer SDAE were found to perform equally or better than deeper networks. This fact was already shown in our previous research for learning patients' representation [48] and an excellent work by [19] called Deep Patient. The hidden layer employs half of the neurons of the previous neurons.

Table 2 demonstrates the performance of the performance of proposed approach with well-known classifiers. The first three models in the table depict how good those three distinct classification methods i.e., the SVM, Random Forest and GAN can learn from the latent representation which was learnt through SDAE. For all these models, the output of SDAE with 16 latent variables were employed as their input. The SVM was employed with linear kernel performed slightly better than poly kernel with size of 2. The SVM is computationally very extensive when the kernel size grows up. The Random Forest was highly fine-tuned concerning several parameters (Depth, Pruning, Number of Trees, minimum number of leaves). Finally, a version with 64 trees with pruning enabled and minimum number of leaves of 10 were selected. The SVM and Random Forest were implemented in MatLab 2019.

Besides these models, a fully connected neural networks were trained upon the input of SDAE. In that experiment, the output of SDAE were not employed; instead, a deep neural network with similar structure proposed in [35] was employed to learn the representation and produce the probability for each class. Note in that model the input dimensionality was 1667 that was the number of features. Then, in the middle layer, the output was halved and finally one output in the last layer as expected for a binary model. In our model the input dimensionality of SDAE and this model was 1043 and the middle layer had 521 neurons and corresponding drop-out layers.

It is noteworthy that the AUC of obtained model is significantly lower than similar model in [35] while other measures seem to be higher. Looking more closely into AUC, it can be seen the Random Forest with SDAE obtained relatively higher AUC figure than other models.

The results of CSDNN at the bottom of the table are detailed just to show its achievement upon private hospitals datasets. That were not implemented or employed in this study. The CSDNN which is convolutional network based-model is among the best models proposed in literature by [34]. Here is detailed only their best reported results for 30-Day Readmission Prediction upon two private datasets of theirs: i.e., GHWs and OPR. The proposed method exhibits competitive performance with higher sensitivity. It implies there is still rooms for improvement in future.

It is interesting to note that similar approach with [35] we got significantly higher performance. This, in fact, shows the data dependency of learning algorithms. In such cases, comparing the achievement of other authors can be fair only if we have the same configuration and the same data and same variables. These criteria can be hardly met in different researches though. Given all into account, different methods were implemented and evaluated while the results of these few researches were reported for giving us an instinct of feasible performance.

Overall, according to Table 2, the SDAE-GAN obtained the highest sensitivity and F1-score which is perhaps the contribution of GAN network. with learnt representation, the Random Forest classifier performs dramatically better than SVM while achieved highest AUC amongst our experiments.

## 7. Conclusion and future works

In this study, a novel deep learning model called SDAE-GAN exploiting stacked denoising autoencoders and generative adversarial networks was proposed. The proposed approach was actually an end-to-end deep model which incorporate the comorbidity codes, and multiple groups of relevant variables including sociodemographic, socioeconomic and statistics. To construct the input of model, a massive table containing the comorbidity codes of each patient was created by data mining and record level processing. The Charlson comorbidity codes were used as a basis for chronic disease. The model input was produced after some preprocessing and digitization. The SDAE, then, was employed to learn salient features appropriate for the following GANs model. Finally, a GANs model was manipulated to assess the probability of readmissions through the learnt features and its neural network structure.

The SDAE-GAN was evaluated upon 133K patients with comorbidities. The experimental results unveiled a competitive performance with current state-of-the-art approaches and its superior performance over some well-known machine learning classifiers. This study revealed the potential of deep learning method in predicting the risk of readmissions for comorbid conditions. Some authors attempted to reduce the complexity of readmission prediction by separating the patients' cohort [36] thereby obtaining different performance in different specialties. Therefore, they usually utilize more deal of prior knowledge and feature engineering to build models. Nevertheless, deep learning provides a way to avoid exhaustive analysis by suggesting an end-to-end model even without feature engineering and prior knowledge. Nonetheless, we

should consider that every model has its own pros and cons. That is why various approaches and different intelligent models has been proposed in the literature. In an ongoing research, we are going to apply our method -perhaps with different internal architecture-over all ICD-10 codes. It brings us an insight about whether more prior knowledge which significantly increase the dimensionality and noise (for the model input) could improve the performance of the model or not.

Quality of codes could directly affect the performance of the model. It is possible that some chronic comorbidity codes about the patient in current spell were ignored or missed. The coding quality grows over time. However, determining the real comorbidity close to reality at least by employing what has been stored in EPR, could be a separate research that potentially reduces readmissions risks. Such research, not only potentially reduces the risk of readmissions - by directly targeting the real comorbid conditions- but also provide a more reliable input for data-driven predictive models.

As other health conditions than chronic morbidities can lead to readmissions, studying about potential problems (based on diagnostic codes and procedures) which has most probably contributed to readmissions could add value to this research. Furthermore, severity of disease is different in different geographical regions [42]. Therefore, considering the severity, updating, and validating the chronic comorbidity indexes could be another valuable relevant line of research. Since, on one hand, it could reveal the potential risks directly related to a specific health center. On the other hand, it would provide more insight to target the follow-ups and reducing readmissions.

## 8. Acknowledgement

This research was sponsored by Informatics Department, Royal Berkshire NHS Foundation Trust and Informatics Research Centre, Henley Business School, University of Reading. We would like to thank Eghosa Bazuaye (Director of Trust) and Prof Keiichi Nakata (Director of Informatics Research Centre) for their support in this research.

## 9. References

[1] J. Basu, R. Avila, and R. Ricciardi, "Hospital readmission rates in US States: are readmissions higher where more patients with multiple chronic conditions cluster?," *Health Serv. Res.*, vol. 51, no. 3, pp. 1135–1151, 2016.

[2] J. S. Weissman, J. Z. Ayanian, S. Chasan-Taber, M. J. Sherwood, C. Roth, and A. M. Epstein, "Hospital readmissions and quality of care," *Med. Care*, pp. 490–501, 1999.

[3] H. J. Jiang, D. Stryer, B. Friedman, and R. Andrews, "Multiple hospitalizations for patients with diabetes," *Diabetes Care*, vol. 26, no. 5, pp. 1421–1426, 2003.

[4] B. Friedman and J. Basu, "The rate and cost of hospital readmissions for preventable conditions," *Med. Care Res. Rev.*, vol. 61, no. 2, pp. 225–240, 2004.

[5] S. F. Jencks, M. V Williams, and E. A. Coleman, "Rehospitalizations among patients in the Medicare fee-for-service program," *N. Engl. J. Med.*, vol. 360, no. 14, pp. 1418–1428, 2009.

[6] J. Donzé, S. Lipsitz, D. W. Bates, and J. L. Schnipper, "Causes and patterns of readmissions in patients with common comorbidities: retrospective cohort study," *Bmj*, vol. 347, p. f7171, 2013.

[7] H. H. Hijazi, M. S. Alyahya, H. M. Hammouri, and H. A. Alshraideh, "Risk assessment of comorbidities on 30-day avoidable hospital readmissions among internal medicine patients," *J. Eval. Clin. Pract.*, vol. 23, no. 2, pp. 391–401, 2017.

[8] K. Dharmarajan *et al.*, "Diagnoses and timing of 30-day readmissions after hospitalization for heart failure, acute myocardial infarction, or pneumonia," *Jama*, vol. 309, no. 4, pp. 355–363, 2013.

[9] D. Zekry *et al.*, "Prospective comparison of 6 comorbidity indices as predictors of 1-year post-hospital discharge institutionalization, readmission, and mortality in elderly individuals," *J. Am. Med. Dir. Assoc.*, vol. 13, no. 3, pp. 272–278, 2012.

[10] C. van Walraven *et al.*, "Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community," *Cmaj*, vol. 182, no. 6, pp. 551–557, 2010.

[11] J. Librero, S. Peiró, and R. Ordiñana, "Chronic comorbidity and outcomes of hospital care: length of stay, mortality, and readmission at 30 and 365 days," *J. Clin. Epidemiol.*, vol. 52, no. 3, pp. 171–179, 1999.

[12] D. Kansagara *et al.*, "Risk prediction models for hospital readmission: a systematic review," *Jama*, vol. 306, no. 15, pp. 1688–1698, 2011.

[13] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, 2017.

[14] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, 2017.

[15] M. Dashtban, M. Balafar, and P. Suravajhala, "Gene selection for tumor classification using a novel bio-inspired multi-objective approach," *Genomics*, vol. 110, no. 1, pp. 10–17, 2018.

[16] P. B. Jensen, L. J. Jensen, and S. Brunak, "Mining electronic health records: towards better research applications and clinical care," *Nat. Rev. Genet.*, vol. 13, no. 6, p. 395, 2012.

[17] N. G. Weiskopf, G. Hripcsak, S. Swaminathan, and C. Weng, "Defining and measuring completeness of electronic health records for secondary use," *J. Biomed. Inform.*, vol. 46, no. 5, pp. 830–836, 2013.

[18] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," *J. Am. Med. Informatics Assoc.*, vol. 20, no. 1, pp. 144–151, 2013.

[19] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep patient: an unsupervised representation to predict the future of patients from the electronic health records," *Sci. Rep.*, vol. 6, p. 26094, 2016.
[20] J. Wu, J. Roy, and W. F. Stewart, "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches," *Med. Care*, pp. S106–S113, 2010.
[21] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, 2014.
[22] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *2017 IEEE International Conference on Data Mining (ICDM)*, 2017, pp. 787–792.
[23] I. Goodfellow *et al.*, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.
[24] H. Zhou, P. R. Della, P. Roberts, L. Goh, and S. S. Dhaliwal, "Utility of models to predict 28-day or 30-day unplanned hospital readmissions: an updated systematic review," *BMJ Open*, vol. 6, no. 6, p. e011060, 2016.
[25] S. A. Choudhry, J. Li, D. Davis, C. Erdmann, R. Sikka, and B. Sutariya, "A public-private partnership develops and externally validates a 30-day hospital readmission risk prediction model," *Online J. Public Health Inform.*, vol. 5, no. 2, p. 219, 2013.
[26] C. Hebert *et al.*, "Diagnosis-specific readmission risk prediction using electronic health data: a retrospective cohort study," *BMC Med. Inform. Decis. Mak.*, vol. 14, no. 1, p. 65, 2014.
[27] J. Billings, I. Blunt, A. Steventon, T. Georghiou, G. Lewis, and M. Bardsley, "Development of a predictive model to identify inpatients at risk of re-admission within 30 days of discharge (PARR-30)," *BMJ Open*, vol. 2, no. 4, p. e001667, 2012.
[28] D. J. Taber *et al.*, "Inclusion of dynamic clinical data improves the predictive performance of a 30-day readmission risk model in kidney transplantation," *Transplantation*, vol. 99, no. 2, p. 324, 2015.
[29] S. N. Vigod *et al.*, "READMIT: a clinical risk index to predict 30-day readmission after discharge from acute psychiatric units," *J. Psychiatr. Res.*, vol. 61, pp. 205–213, 2015.
[30] K. Shameer *et al.*, "Predictive modeling of hospital readmission rates using electronic medical record-wide machine learning: a case-study using Mount Sinai heart failure cohort," in *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2017*, 2017, pp. 276–287.
[31] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512.
[32] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Deepcare: A deep dynamic memory model for predictive medicine," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2016, pp. 30–41.
[33] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. Ben Abdallah, and A. Kronzer, "Cost-sensitive Deep Learning for Early Readmission Prediction at A Major Hospital," *Canada Proc. BIOKDD*, no. 17, 2017.
[34] H. Wang, Z. Cui, Y. Chen, M. Avidan, A. Ben Abdallah, and A. Kronzer, "Predicting hospital readmission via cost-sensitive deep learning," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 15, no. 6, pp. 1968–1978, 2018.
[35] M. Jamei, A. Nisnevich, E. Wetchler, S. Sudat, and E. Liu, "Predicting all-cause risk of 30-day hospital readmission using artificial neural networks," *PLoS One*, vol. 12, no. 7, p. e0181173, 2017.
[36] C. Xiao, T. Ma, A. B. Dieng, D. M. Blei, and F. Wang, "Readmission prediction via deep contextual embedding of clinical concepts," *PLoS One*, vol. 13, no. 4, p. e0195024, 2018.
[37] A. Rajkomar *et al.*, "Scalable and accurate deep learning with electronic health records," *NPJ Digit. Med.*, vol. 1, no. 1, p. 18, 2018.
[38] J. Futoma, J. Morris, and J. Lucas, "A comparison of models for predicting early hospital readmissions," *J. Biomed. Inform.*, vol. 56, pp. 229–238, 2015.
[39] U. Hwang, S. Choi, and S. Yoon, "Disease prediction from electronic health records using generative adversarial networks," *arXiv Prepr. arXiv1711.04126*, 2017.
[40] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*, vol. 1. MIT press Cambridge, 2016.
[41] U. Hwang, D. Jung, and S. Yoon, "HexaGAN: Generative Adversarial Nets for Real World Classification," *arXiv Prepr. arXiv1902.09913*, 2019.
[42] H. Quan *et al.*, "Updating and validating the Charlson comorbidity index and score for risk adjustment in hospital discharge abstracts using data from 6 countries," *Am. J. Epidemiol.*, vol. 173, no. 6, pp. 676–682, 2011.
[43] T. Pham, T. Tran, D. Phung, and S. Venkatesh, "Predicting healthcare trajectories from medical records: A deep learning approach," *J. Biomed. Inform.*, vol. 69, pp. 218–229, 2017.
[44] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, no. Dec, pp. 3371–3408, 2010.
[45] L. Zamparo and Z. Zhang, "Deep autoencoders for dimensionality reduction of high-content screening data," *arXiv Prepr. arXiv1501.01348*, 2015.
[46] H.-I. Suk, S.-W. Lee, D. Shen, and A. D. N. Initiative, "Latent feature representation with stacked auto-encoder for AD/MCI diagnosis," *Brain Struct. Funct.*, vol. 220, no. 2, pp. 841–859, 2015.
[47] Y. Wang, H. Yao, and S. Zhao, "Auto-encoder based dimensionality reduction," *Neurocomputing*, vol. 184, pp. 232–242, 2016.
[48] M. Dashtban and W. Li, "Deep Learning for Predicting Non-attendance in Hospital Outpatient Appointments," in *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.
[49] J. M. Bernardo *et al.*, "Generative or discriminative? getting the best of both worlds," *Bayesian Stat.*, vol. 8, no. 3, pp. 3–24, 2007.
[50] A. R. Kroeger, J. Morrison, and A. H. Smith, "Predicting unplanned readmissions to a pediatric cardiac intensive care unit using predischarge Pediatric Early

Warning Scores," *Congenit. Heart Dis.*, vol. 13, no. 1, pp. 98–104, 2018.
[51] W.-S. Lai, J.-B. Huang, and M.-H. Yang, "Semi-supervised learning for optical flow with generative adversarial networks," in *Advances in Neural Information Processing Systems*, 2017, pp. 354–364.
[52] M. Lin, "Softmax gan," *arXiv Prepr. arXiv1704.06191*, 2017.