January 2020

Representation Learning With Autoencoders For Electronic Health Records

Najibesadat Sadatijafarkalaei
Wayne State University

REPRESENTATION LEARNING WITH AUTOENCODERS FOR ELECTRONIC HEALTH RECORDS

by

NAJIBESADAT SADATIJAFARKALAEI

THESIS

Submitted to the Graduate School,

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

MASTER OF SCIENCE

2020

MAJOR: COMPUTER SCIENCE

Approved By:

____

Advisor Date

DEDICATION

I would like to dedicate my work to my great family for their endless love, encouragement and support.

ACKNOWLEDGEMENTS

I would first like to express my gratitude to my advisor Dr. Dongxiao Zhu for his continuous support, motivation and academic advises throughout this dissertation project. Also I would like to thank to Dr. Ratna Babu Chinnam for his support and guidance in different steps of this research. I am grateful to Dr. Suzan Arslanturk and Dr. Ming Dong for serving on my dissertation committee and providing valuable insights and suggestions. I appreciate Dr. Philip Levi's support from department of emergency medicine and cardiovascular research institute at Medical School of Wayne State University for helping us to access cardiovascular datasets. Finally, I must express my gratitude to Milad (my husband) and my family for their love, encouragement and endless support. Without their help this accomplishment could have never been possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

Healthcare is transitioning to a new paradigm under the emergence of large biomedical datasets in various domains of health. In fact, the explosive access to large Electronic Health Records (EHR) and advanced analytics is providing a great opportunity in recent years for improving the quality of healthcare [30, 49]. The availability of patient-centric data brings about new opportunities in healthcare and enable data scientists to pursue new research avenues in the realm of personalized medicine using data-driven approaches.

Since EHRs are complex, sparse, heterogeneous and time dependent, leveraging them for personalized medicine is challenging and complicated to interpret. Representation learning (feature learning) provides an opportunity to overcome this problem by transforming medical features to a higher level abstraction, which can provide more robust features. On the other hand, labeling of clinical data is expensive, difficult and time consuming in general. However, there are many instances where unlabeled data may be abundant. Representation learning through unsupervised approaches is a very effective way to extract robust features from both labeled and unlabeled data which can enhance the performance of supervised learning on the top of labeled data.

There are many methods in the literature developed in order to overcome challenges in pattern recognition of electronic health records. These methods so called shallow learning like as Principle Component Analysis (PCA), Independent component analysis (ICA) and Manifold learning could not solve many challenges faced in EHR analytics. In addition, the traditional feature selection approaches like as tree-based algorithms [30, 57] or linear approaches (e.g Lasso regression [64, 59]) could improve the clinical prediction tasks and provide interpretable results but were not enough for many purposes . In the other hand, deep representation or deep learning provides several promising performance in learning from complex data like as electronic health records. In comparison with shallow feature mapping, the key advantage of feature representation using deep learning is the ability to provide more abstract features representation at higher levels by non-linear transformation using deep hidden layers [3].

Deep representation leaning has shown success in several domains such as computer vision, natural language processing, audio recognition and signal processing [13]. The choice of data representations has a significant impact on the performance of different machine learning tasks [3]. There has been increasing interest in using representation learning, which tries to learn higher level abstraction of data representations that are crucial as input to improve the performance of prediction models [13].

In clinical domain, building an accurate prediction model is challenging since EHRs include high dimensional heterogeneous structured and unstructured data components including demographic information, laboratory test results, diagnoses, medication information, clinical test notes, and medical images. Therefore, representation learning is needed to provide better information as input of prediction tasks in clinical domain such as precision medicine that focuses on the use of patient-centered data to recommend a treatment based on personalized health attributes. [26, 42]. Representation learning can overcome many challenges (e.g. label scarcity [31]) in facing EHRs while the choice of feature representation is significant to achieve better accuracy [3].

Recently several research have been applied deep feature representation as unsupervised approach on clinical data in the domain of precision medicine and health informatics [71] in order to provide more accurate prediction results on different targets such as risk of disease, disease phenotyping, treatment recommendation and medical events [41]. In spite of performing many representation learning research, there is significant opportunity to investigate about the choice of representation to provide a comprehensive guidance based on real work experimental studies.

In this research, we perform exploratory analysis [16] to show the impact of deep feature representation on prediction performance and investigate the choice of deep learning approach across small and large datasets. For this goal, we proposed an integrated approach as illustrated in Figure 1 including three consecutive steps. In step 1, we do preprocessing on clinical data including data integration from different sources of data (diagnosis, test results, patient information, etc.), cleaning the data (handling missing values and data quality) and transferring text features to vectors. In step 2, we trained deep neural network for feature representation based on different deep Auto-Encoders (AE). This step is unsupervised learning step and can be utilized by both labeled and unlabeled data specially in the cases that labeled data are scarce and expensive to be collected. After training the deep network for feature representation based on training set, the trained network is applied to represent (transform) the patient features (for both training and evaluation sets) as input of supervised learning step.

Step 3 is the supervised learning step where the target of interest is predicted using represented features. Since we applied different AEs in step 2, the best deep architecture is selected based on performance of supervised learning in step 3. While the proposed approach utilizes deep learning for feature representation to improve the prediction accuracy, the main goal is to compare and evaluate the choice of feature representation (deep network architecture) through steps 2 and 3 in case of small and large EHR datasets.

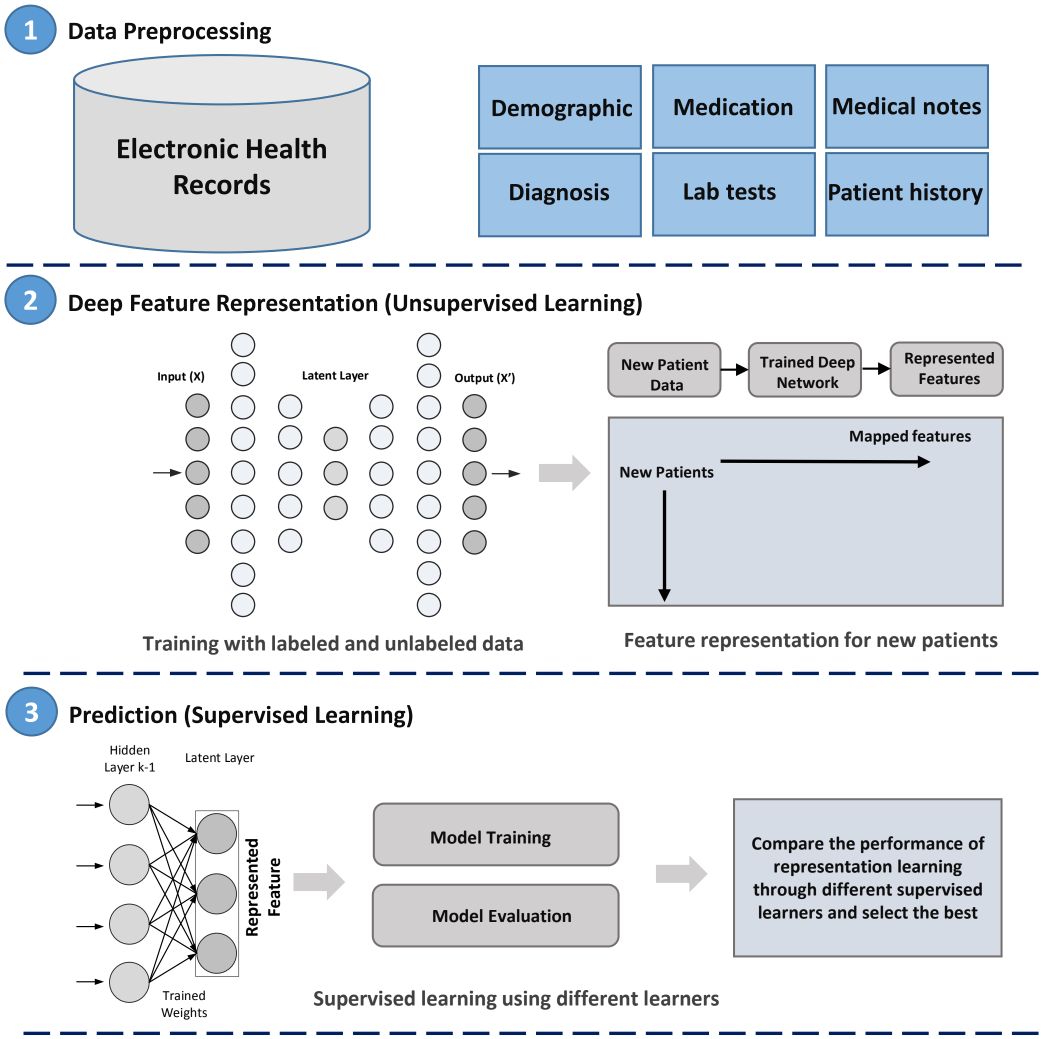

1
Data Preprocessing
Electronic Health Records
Demographic
Medication
Medical notes
Diagnosis
Lab tests
Patient history
2
Deep Feature Representation (Unsupervised Learning)
Input (X)
Latent Layer
Output (X')
Training with labeled and unlabeled data
New Patient Data
Trained Deep Network
Represented Features
Mapped features
New Patients
Feature representation for new patients
3
Prediction (Supervised Learning)
Hidden Layer k-1
Latent Layer
Represented Feature
Trained Weights
Model Training
Model Evaluation
Compare the performance of representation learning through different supervised learners and select the best
Supervised learning using different learners

Figure 1: The three consecutive steps of proposed approach

With this goal, we apply the proposed approach on two different healthcare informatics case studies using real clinical data. The first case study is related to prediction of risk disease where we use small-size EHRs data for specific subgroup of patient to predict heart failure risk. In the second experimental study, we apply the proposed approach on two large datasets to predict the patient length of stay in Intensive Care Unit (ICU). We use eICU collaborative research database, a freely available multi-center database for critical care research.

In general, we develop a predictive approach using deep learning and data representation for EHRs. In our method, we apply four deep architectures for feature representation in higher levels abstraction: Stacked sparse autoencoders, Deep belief network, Adversarial autoencoders and Variational autoencoders. Our contributions in this paper lie into two folds: 1) Improving predictive modeling by deep feature representation on EHRs where we apply various deep networks including advanced generative autoencoders (AAE, VAE) and regular autoencoders (SSAE, DBN). 2) It is one of the first comparative studies to investigate the choice of deep representation among small and large datasets, and provide practical guidelines.

In the rest of this thesis, in chapter 2, we review the shallow feature learning methods and importance of deep feature learning approaches in health informatics. In chapter 3, we explain the proposed prediction approach in detail and chapter 4 reports two clinical experimental case studies and implementation results. Finally, chapter 5 finishes with a discussion of results and future works.

CHAPTER 2 RELATED WORKS

Since principal component analysis (PCA) is developed [46], feature learning (feature extraction or representation learning) has been researched for more than a century to overcome the challenges of high dimensionality [73]. In this period, many shallow methods including linear and non-linear feature learning approaches have been developed until deep learning is introduced as an exciting new trend of machine learning in recent years. In the feature representation learning domain, generally, the algorithms can be divided into two categories: shallow learning and deep learning, based on the level of their hierarchical abstraction from input features. Before 2006, several researchers (e.g. [28]) made significant efforts to train deep multilayer network (more than five hidden layers) but their efforts was not successful because of lack of large scale data and existence of many hyper parameters in deep network. Therefore, shallow learning approaches overcome the feature learning and representation domain. Principal component analysis (PCA) [46] and linear discriminant analysis (LDA) [17] are two most popular shallow learning algorithms. The first one is a an unsupervised approach and the second one is a supervised method.

Hinton and Salakhutdinov [20] introduced a fast algorithm based on greedy layer-wise pre-training approach for deep neural networks training, and initiated a new research area of deep learning. Accordingly, their work was continued by other researches with similar idea. Deep learning demonstrated a great performance in feature representation through supervised and unsupervised learning approaches especially in the medical and health science domains. In this section, we first review the shallow feature learning methods and then we describe the importance of deep feature learning and review both supervised and unsupervised deep feature representation approaches.

2.1 Shallow Feature Learning

The most popular shallow feature representation approach is Principle Component Analysis. PCA produces linear combinations of features, called principal components, which are orthogonal to each other, and can explain variation of features, and may achieve lower dimensionality. PCA is applied in several healthcare applications. Martis et al. [37] used PCA for classification of ECG signals for automated diagnosis of cardiac health. Yeung et al. [70] applied PCA to project gene expressions into lower dimension to cluster genes. In the other study [34], authors proposed a systematic approach based on PCA to detect the differential gene pathways which are associated with the phenotypes.

In addition to standard PCA, several PCA-based approaches have been developed and applied in bioinformatics studies to improve the performance of PCA. For example, supervised PCA is proposed in [1] in order to diagnose and treat cancer more accurately using DNA microarray data. Zou et al. [75] proposed Sparse PCA using lasso (elastic net) regression for gene expression arrays. In the other research, Nyamundanda et al. [45] applied Probabilistic PCA to analyze the structure of metabolomic data.

Instead of PCA-based approaches, several other shallow feature learning methods have been developed with application to healthcare. Independent Component Analysis (ICA) [69] and Linear Discriminant Analysis (LDA) [68] are good examples of linear approaches.

Non-linear dimensionality reduction first developed by Kernel version of some linear dimensionality reduction algorithms. These kernel based approaches map the original feature space to a higher dimensions using a nonlinear function and then apply linear dimensionality reduction on the high dimensional feature space [73]. Two efficient kernel dimensionality reduction methods are the kernel version of PCA and LDA. The fisrt one is kernel PCA (KPCA) [53] and the second one is generalized discriminant analysis (GDA) [2]. There are many studies in bioengineering domain specially in ECG signal classification that applied kernel based dimensionally reduction [6, 24].

Among other nonlinear dimensionality reduction or so-called manifold learning approaches, isometric feature mapping (Isomap) [58], locally linear embedding (LLE) [52] and stochastic neighbor embedding (SNE) [21] are the most popular models. Several manifold learning algorithms have been developed with the goal of keeping local information between instances in the lower dimensional space.

While Kernel-based approaches map the original feature space into a higher dimensions and then project them to a reduced space through a linear function, manifold learning methods directly learn nonlinear representation of original features. For example, Isomap uses the geometric distances between all pairs of instances and estimates the intrinsic geometry of a data manifold and provides low-dimensional embedding of features [58] while LLE can learn the whole structure of nonlinear manifold and provides the local symmetries of linear reconstructions [52]. Finally, stochastic neighbor embedding (SNE) as another nonlinear dimension reduction approach formulates the neighborhood relationship among features in a low dimensional space [21]. SNE utilizes Gaussian distribution in the low dimensional space, Maaten and Hinton [35] extended SNE to t-SNE which uses a heavy-tailed Student-t distribution with one-degree of freedom to compute the similarity among the embeddings.

In the biomedical domain, Li et al. [29] applied LLE on the gene expression data to map them to low dimensional space in order to improve the accuracy of classification tasks. There are many other applications of manifold learning in the medical domain which are addressed in a review study by Mateus et al. [38].

In general, The key benefit of using shallow feature learning is the capability to interpret the represented features. Shallow feature learning approaches are efficient computationally and the learning process is straight forward but they have not demonstrated great performance in high dimensional and complex data such as temporal/spatial data or image data because they cannot be stacked to provide deeper and more abstract representations [3].

2.2 Deep Feature Learning

Deep neural networks have been illustrated to contribute promising capabilities in learning complex patterns in data, and have achieved remarkable success in several domains such as computer vision, natural language processing, speech recognition and etc. Recently, many efforts have been made in the field of biomedical and health informatics to improve the performance of machine learning tasks.

Deep feature representation has been employed in several areas using EHRs (e.g. diagnosis and medication data), genomics data, medical text and imaging data with various purposes including risk factors selection, disease phenotyping and disease risks prediction or classification [41]. Feature representation using deep learning can be achieved either by supervised deep learning predictive models (e.g., deep feed-forward neural network and convolutional nets) or by unsupervised deep learning approaches (e.g., deep autoencoders). In this section, we review the related studies for both approaches in biomedical and healthcare applications.

2.2.1 Deep Supervised Feature Learning

The supervised deep predictive approaches extract features through learning weights and biases with considering target variable in the cost function. As a good example, Li et al. [32] proposed a novel deep feature selection model for selecting significant features inputted in a deep neural network for multi-label data. The authors used elastic net regularization to select most important features. They added a one-to-one linear layer between the visible layer and the first hidden layer of a multi-layer perception (MLP) to rank features based on regularized weights in the input layer obtained after training. Finally, they applied their model to a genomics dataset. In another study, Cheng et al. [9] first represented the EHRs as a temporal matrix with two dimensions, time and event for all records and then used four-layer CNN for extracting phenotypes and applying prediction for two case studies: congestive heart failure and chronic obstructive pulmonary disease.

Choi et al. [10] proposed a predictive framework termed Doctor AI for medical events. The authors employed a recurrent neural network (RNN) on large-scale temporal EHR data to predict the diagnosis and medication categories for further visits. Zhao et al. [72] developed a brain tumor image segmentation method using convolutional neural networks (CNNs). They applied their approach on multimodal brain tumor image segmentation benchmark (BRATS) data and obtained advanced accuracy and robustness.

Feature learning through deep supervised predictive approach requires large scale labeled data for training while in many healthcare applications it is hard to collect enough labeled data. Unsupervised and semi-supervised feature learning approaches can overcome the label scarcity problem and provide better feature representation.

2.2.2 Deep Unsupervised Feature Learning

Many studies used unsupervised or semi-supervised deep feature learning for EHRs and applied predictive models on top of represented features. Miotto et al. [40] applied stack denoising autoencoders (SDA) for feature learning and representation of large scale electronic health records. They used EHRs of approximately $700,000$ individuals related to several diseases including schizophrenia, diabetes, and various cancers. Their model improved medical prediction, which could offer a machine learning framework for augmenting clinical decision systems.

Recently, Che at al. [8] proposed a semi-supervised framework for EHRs risk prediction and classification. They developed a modified generative adversarial network called ehrGAN for feature representation and used CNN for performing prediction task. In the other research, [43], authors proposed a novel feature selection model using deep stacked autoencoders. They performed their approach on a health informatics problem to identify the most important risk factors related to African-Americans who are in risk of heart failure. Wulsin et al. [66] developed an approach using deep belief nets for electroencephalography (EEG) anomaly detection to monitor brain function in critically ill patients.

Cao et al [7] trained deep belief network on several large datasets for prediction of protein tertiary structure for protein quality assessment based on different perspectives, such as physio-chemical and structural attributes. In the other study [14], a healthcare recommender system developed based on variational autoencoders and collaborative filtering. Authors used VAE to learn better relationships between items and users in collaborative filtering.

Our predictive approach for cardiovascular risk level (LVMI) and length of stay (LOS) in ICUs can be considered in the last group. Readers for more comprehensive review about deep feature learning applications in healthcare can refer to recent review articles provided by Ravi et al. [50], Litjens et al. [33], Miotto et al. [41], Xiao et al. [67] Shickel et al. [54] and Purushotham et al. [48].

Although deep feature learning (supervised, unsupervised and semi-supervised) generally provides better representation rather than shallow approaches, but it is hard to interpret them and computationally expensive to train with several hyperparameters. Therefore, the right choice of deep network can reduce significant effort in training process. In this way, we try to provide empirical insights for choice of deep representation approach across small and large datasets which has not been studied in the literature.

CHAPTER 3 DEEP REPRESENTATION LEARNING

Many machine learning tasks can be improved depending on how the input data is represented. For example, the operation of inserting a digit into the right position in a sorted list considered as an $O(n)$ operation in the case that the list is represented in a linked list, but if the list represented as a red-black tree, it changes to an $O(logn)$ operation. This example shows how data representation can enhance different information processing tasks including machine learning. In other words, a good feature representation approach can make machine learning tasks easier and more accurate [18]. Therefore the choice of representation learning can be considered as important step of machine learning tasks which can improve their performance significantly.

Similar to supervised deep networks (e.g. feed-froward deep architecture), unsupervised deep learning algorithms include a main objective for training but also they learn a feature representation as additional output. This represented features can be used on the other tasks. Accordingly, multiple tasks including supervised and unsupervised can be learned with together based on shared represented layers. Since representation learning can provide unsupervised and semi-supervised learning, it becomes interesting. In many application such as healthcare, labeled data are insufficient and expensive which lead to over-fitted trained supervised models. On the other side, there are large amount of unlabeled data for training and feature representation that provide this opportunity to overcome this over-fitting problem by learning from unlabeled data. Particularly, we can utilize the represented unlabeled data to improve the supervised learning tasks [18].

Unsupervised learning considered a key step in revolution of deep learning and it enables scientists to train deep supervised network without need of special networks such as recurrence or convolution architectures. This unsupervised learning that trained before supervised learning task called layer-wise unsupervised training [18]. Greedy layer-wise unsupervised pre-training approach depends on a single layer representation learning algorithm such as an Restricted Boltzmann Machine (RBM), a single-layer Auto-Eencoder (AE) or any other model that can learn latent representations. In this approach, each layer is pre-trained based on unsupervised learning and the output of that layer is used for input of the next hidden layer as a new represented data with a new distribution whose the pattern can be learned easier.

Greedy layer-wise training algorithms based on unsupervised learning have long been studies to solve the problem of jointly training the layers of a deep architecture for a supervised goal. The discovery of layer-wise approach in 2006 by Hinton [20] started the journey to develop a good initialization for a joint learning algorithm over all the layers which could be applied successfully to train fully connected architectures as well.

In this chapter of this thesis, we review 4 important unsupervised deep architecture for representation learning: Stacked Sparse Autoencoder, Deep Belief Network, Variational Autoencoder and Adversarial Autoencoder. We try to provide a sufficient explanation for each of them including their architecture, objective function and their pros and cons.

3.1 Stacked Sparse Autoencoder

An autoencoder network is an unsupervised learning methodology which number of output layer's neurons are equal to the number of input layer's neurons. AE tries to reconstruct input data (x) in output ($x^{'}$) layer by encoding and decoding process [22, 62]. AEs consist of an encoder, which converts the input to a latent representation, and a decoder, that remodels the input from this representation. Autoencoders are trained to minimize the reconstruction errors. Stacked Autoencoders is a deep network consisting of multiple layers of autoencoders (Figure 2) which can be trained in layer wised approach [4]. After training of deep network, middle layer illustrates the highest-level representation of original features [27].

The loss function for training an autoencoder can be defined as following:

$$Loss(x,x^{'}) = \|x - x^{'}\| = \|x - f(W^{'}(f(Wx+b)) + b^{'})\|, \tag{3.1}$$

where f is the activation function and W, $W^{'}$, b and $b^{'}$ are the parameters of the hidden layers.

The above loss function is reliable when the number of hidden units in the latent layer being small, but even in the case of large hidden units (even greater than the number of input features), which is called sparse representation or stacked sparse autoencoder (SSAE), we can still explore reliable architecture, by imposing sparsity constraints on the network [44]. In Sparse autoencoders, we formulate the loss function with regularizing activations (not weights of the network) and we encourage the learners to train encoding and decoding based on activating a small number of neurons. We can impose this sparsity constraint by adding $L1$ regularization or KL-Divergence (E.q 3.2) to the loss function [44].

$$Loss(x,x^{'}) + \lambda \sum_{i} \mid a_i^{(h)} \mid \quad or \quad Loss(x,x^{'}) + \lambda \sum_{j} KL(\rho \| \bar{\rho}_j). \tag{3.2}$$

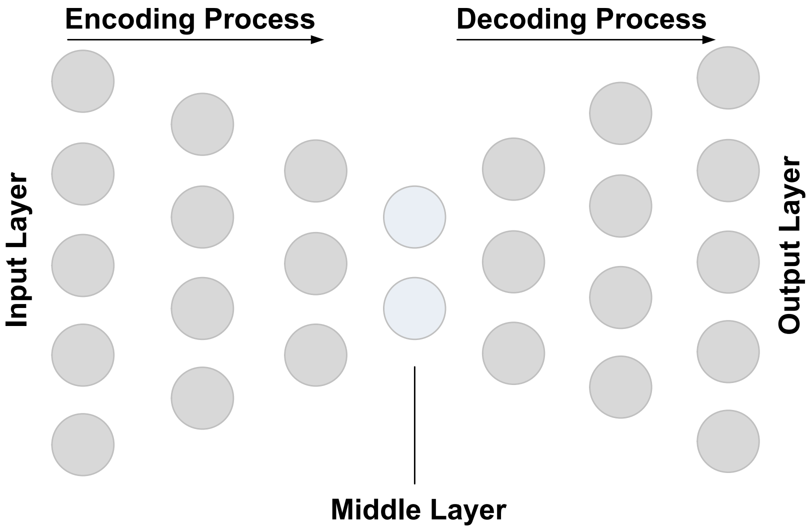


Figure 2: Stacked Autoencoders (SAE) architecture

3.2 Deep Belief Network

Deep Belief Networks are probabilistic generative models that are composed by stacking of multiple RBMs to provide better representation rather than single RBM. Hinton and Salakhutdinov [22] proposed the layer-wise training procedure which deep belief network can be trained layer by layer in unsupervised learning approach. They used the joint probability distribution between visible and latent layers as follows:

$$P(x, h^1, ..., h^l) = \prod_{k=0}^{l-2} P(h^k|h^{k+1})P(h^{l-1}h^l), \tag{3.3}$$

where, x= h^0, $P(h^k|h^{k+1})$ is a conditional probability of visible units in level k given the state of hidden units of the RBM in level $k+1$, and $P(h^{l-1}, h^l)$ defined as visible-hidden joint distribution in the last level of RBM. This structure has been illustrated in Figure 3.

In the layer-wised training approach, the input layer or so called visible units is trained as a RBM and the output is transformed into the hidden layer based on optimizing of log-likelihood as below [20]:

$$\log p(x) = KL(Q(h^{(1)}|x)\|p(h^{(1)}|x)) + H_{Q(h^{(1)}|x)} + \sum_h Q(h^{(1)}|x)(\log p(h^{(1)})) + \log p(h^{(1)}). \tag{3.4}$$

$KL(Q(h^{(1)}|x)\|p(h^{(1)}|x))$ is the KL divergence between $Q(h^{(1)}|x)$ of the first RBM and $p(h^{(1)}|x)$. Then the represented hidden units in the first layer are considered as input layer (visible units) for the second layer of DBN and this process continues until training of whole network. More comprehensive detail about the training process of DBN is provided in [20] and [4].

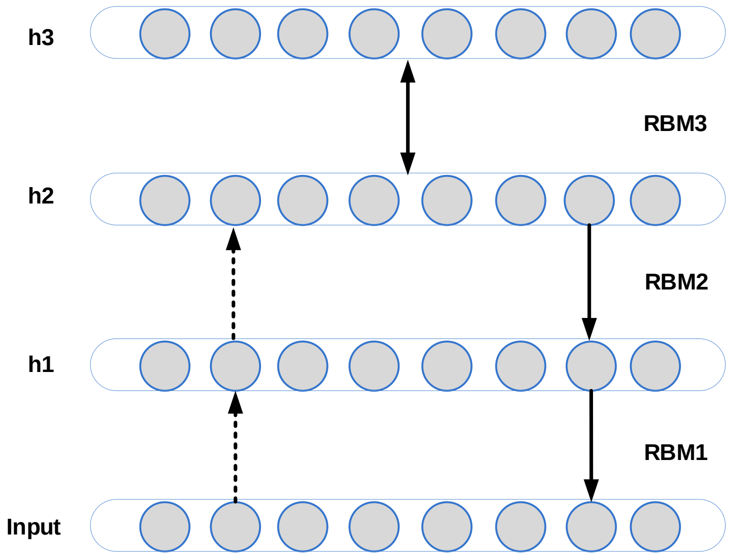


Figure 3: Deep Belief Network (DBN) architecture

3.3 Variational Autoencoder

Variational Autoencoder (VAEs) is one of the most popular approaches to representation learning developed in recent years. Variational autoencoders are probabilistic generative models and have the same architecture as autoencoders, but consider specific assumptions about the distribution of middle/latent layer variables. Variational autoencoders learn the true distribution of input features from latent variables distribution using Bayesian approach and present a theoretical framework for the reconstruction and regularization purposes [56]:

$$p(x) = \int p(x, z)dz = \int p(x|z)p(z)dz. \tag{3.5}$$

In Eq. (3.5), $p(x|z)$ is the probability function of the observed data and the output of the decoder network by considering noise terms. In this equation, z is the latent representation and $p(z)$ is the representation prior with an arbitrary distribution such as standard normal distribution or a discrete distribution like as Bernoulli distribution. There exist two problems for solving above equation: defining the latent variables (z) and marginalizing over z. The key intention behind the variational autoencoder is to try to sample values of z that are likely to have generated x and compute $p(x)$ for these values. To make tractable above integral, an approach is to maximize its variational lower bound using the Kullback-Leibler divergence (KL divergence or D) as follows: [15]:

$$\begin{aligned} &E_{q_\phi(z|x)}[\log^{p(x|z)}] - D(q(z|x) \parallel p(z)) = \\ &log^{p(x)} - D[q(z|x) \parallel p(z|x)]\ . \end{aligned} \tag{3.6}$$

We can apply Bayes rule to $p(z|x)$ and reformulate Eq. (3.6):

$$log^{p(x)} - D[q(z|x) \parallel P(z|x)] =$$

$$E_{z \sim q}[log^{p(x|z)}] - D[q(z|x) \parallel p(z)], \tag{3.7}$$

while q is encoding x to z and p is decoding z to reconstruct input x. The structure of variational autoencoder has been illustrated in Figure 4.

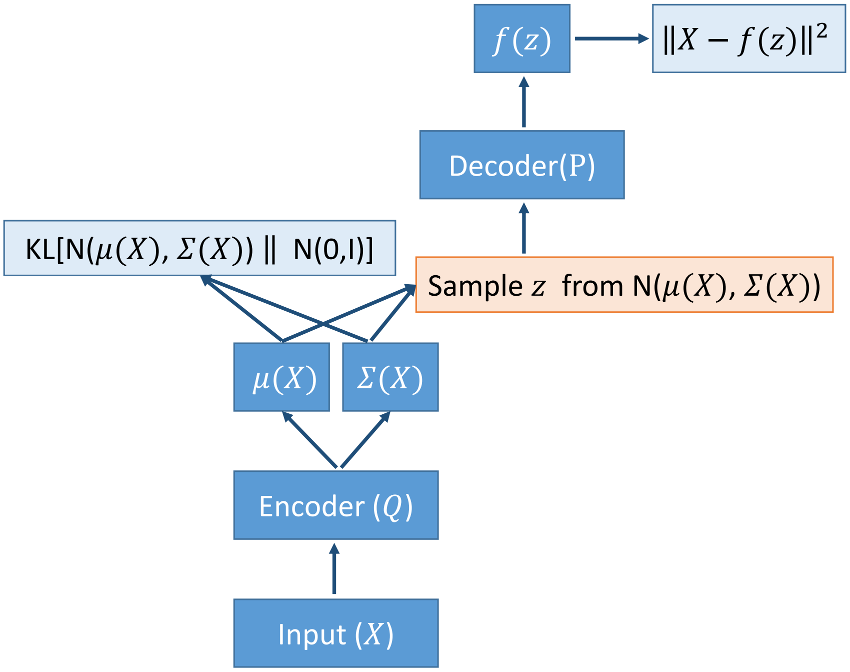
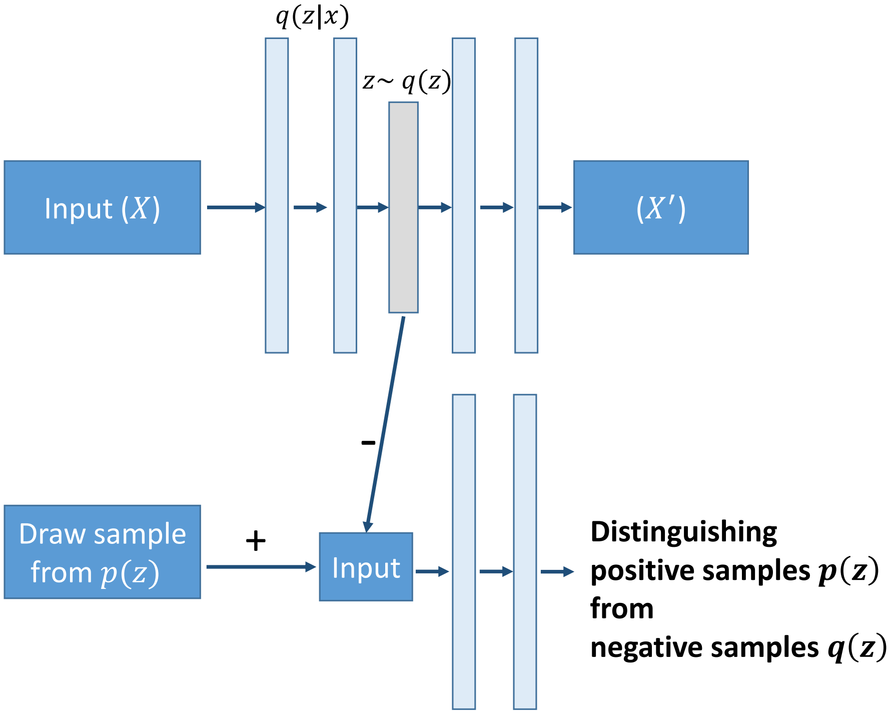

$f(z)$
$\|X - f(z)\|^2$
Decoder(P)
KL[N($\mu(X)$, $\Sigma(X)$) ‖ N(0,I)]
Sample z from N($\mu(X)$, $\Sigma(X)$)
$\mu(X)$
$\Sigma(X)$
Encoder (Q)
Input (X)

Figure 4: Variational autoencoder network, where $P(X|z)$ is Gaussian distribution

3.4 Adversarial Autoencoder

Adversarial autoencoder (AAE) is a probabilistic autoencoder based on generative adversarial networks (GAN) [19] which propose a minmax game among two neural network models: generative model (G) and discriminative model (D). The discriminator model, $D(x)$, is a neural network that estimate the probability of a point x in data space came from data distribution (true distribution which our model is training to learn) rather than coming from generative model [36]. At the same time, the generator model, $G(z)$, tries to map sample points z from the prior distribution $p(z)$ to the data space. $G(z)$ is trained by maximum confusing of discriminator in trusting that samples it produces; originated from the data distribution. The generator is trained by using the gradient of $D(x)$ related to x, and using that to improve its parameters. The solution of this game can be represented as below:

$$\min_G \max_D E_{x\sim p_{data}}[\log D(x)] + E_{z\sim p(z)}[\log(1 - D(G(z)))]. \tag{3.8}$$



$q(z|x)$
$z\sim q(z)$
Input (X)
(X')
-
+
Draw sample from $p(z)$
Input
Distinguishing positive samples $p(z)$ from negative samples $q(z)$

Figure 5: Adversarial autoencoder network, where the top row is a standard autoencoder and the bottom row shows a second network trained to discriminatively classify whether a sample arises from the latent layer or from a arbitrary distribution

The adversarial autoencoder uses similar idea of GAN in training true distribution of data space by matching the aggregated posterior of latent variables to an arbitrary prior distribution in the reconstruction and the regularization phases [36]. In the other word, The adversarial autoencoder is an autoencoder that is regularized by coordinating the aggregated posterior, q(z), to an arbitrary prior, p(z). Simultaneously, the autoencoder attempts to minimize the reconstruction error. The architecture of an adversarial autiencoder is shown in Figure 5.

The summary of deep networks applied in this study has been described in table 1.

Table 1: Summary and comparison of deep networks used in this study

Architectures Description	Key Points
1. Stacked Sparse Autoencoder • Proposed in [22, 44] with the goal of dimensionality reduction • AE tries to reconstruct input data in output layer by encoding and decoding process	**Pros** • Sparse autoencoder is appropriate for small data through regularization. • Sparse autoencoder can be fine tuned easily by itself using ordinary back-propagation approach **Cons** • The pre-training step is needed • Vanishing errors may cause problem in training step
2. Deep Belief Network • Introduced in [20] constructed by stacking of several RBMs • DBNs are graphical models that can be trained based on greedy-layer wised approach • Only the connection between top layers is undirected	**Pros** • Take the advantages of energy-based loss function instead of ordinary one **Cons** • Training process is computationally expensive • Fine tuning of DBNs seems to be difficult
3. Variational Autoencoder • Proposed in [25] to learn the true distribution of input features from latent space distribution using Bayesian approach • VAEs apply a KL divergence term to impose a prior on the latent layer	**Pros** • VAEs are flexible generative model • VAE is a principled approach to generative models **Cons** • Approximation of true posterior is limited • VAEs Can have high variance gradients
4. Adversarial Autoencoder • Proposed in [36] to impose the structure of input data on the latent layer of an autoencoder • Adversarial autoencoders are generative autoencoders that use adversarial training to match the distribution of an arbitrary prior on the latent space	**Pros** • Flexible representation to impose arbitrary distributions on the latent layer. • It can capture any distribution for generation sample, both continuous and discrete **Cons** • It is challengeable to train because of the GAN objective • It is not scalable to higher number of latent variables

CHAPTER 4 METHODOLOGY AND EXPERIMENTAL RESULTS

In this project, we propose a comprehensive evaluation study using an integrated predictive framework for deep representation learning. We use this framework to compare different deep networks to solve healthcare informatics problems in predictive modeling. Our methodology follows the work flow shown in Figure 6 that includes three consecutive steps.

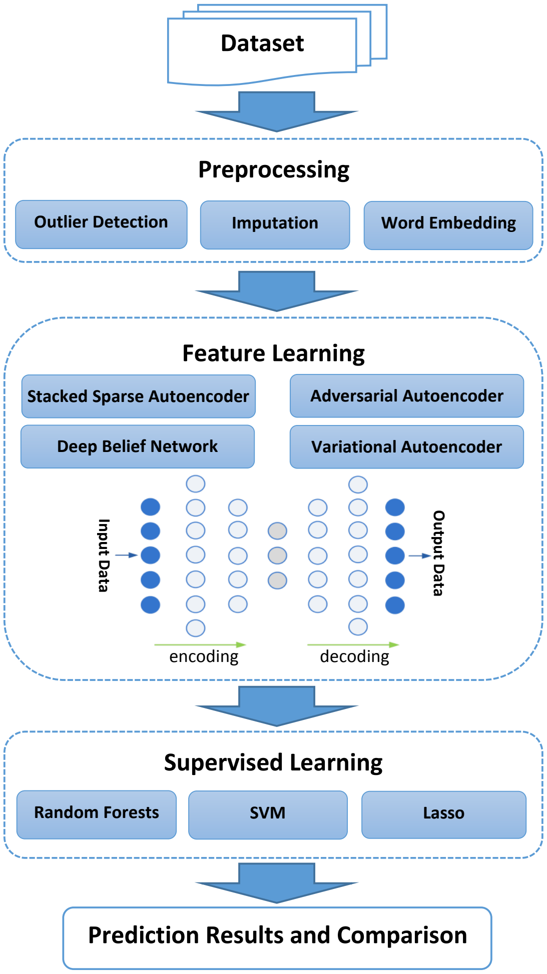
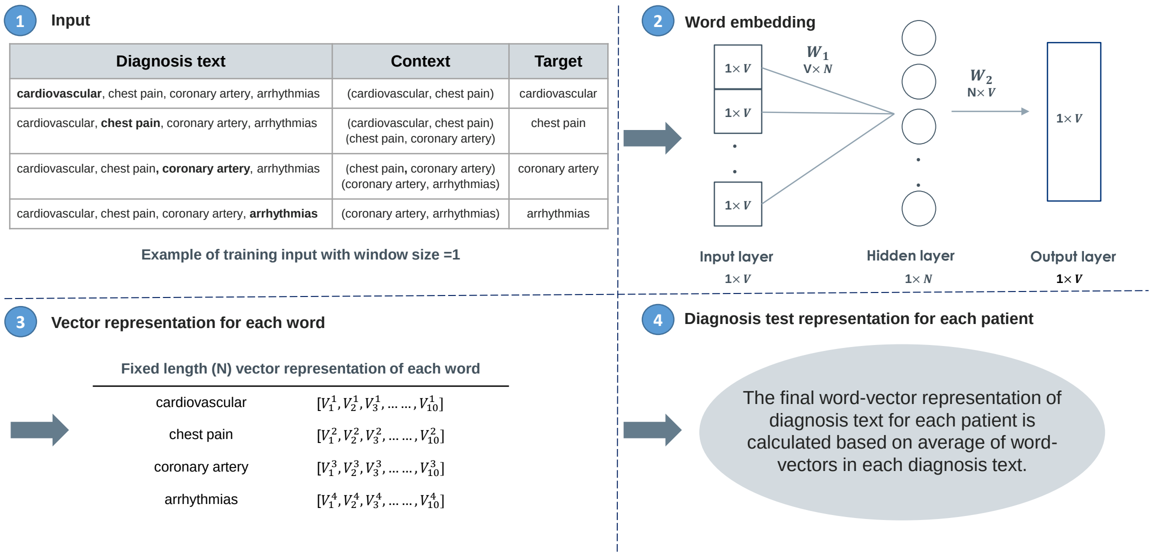


Figure 6: The technical workflow of the proposed approach

4.1 Step−1: Preprocessing and Word Embedding

In the first step, we use the preprocessing methods such as outlier detection and imputation for missing values existed in the dataset. We also transform text variables to vectors using word embedding techniques. Discovering efficient representations of text features have been a key challenge in a variety of biomedical and healthcare applications [11]. Word Embedding algorithms are developed to map text features (words) to vectors of real numbers. Word embeddings have been widely applied in Natural Language Processing (NLP) applications to provide vector representations of unstructured data. Word embedding can capture semantic properties and relationship between words using word co-occurrence matrix and shallow neural networks techniques. Word embeddings have been popular used as feature input to machine learning tasks, which provide better representation from raw text data [65].

There has been an increasing amount of research using word embeddings in biomedical domain [65]. For instance, recently some studies have been applied word embedding techniques to learn vector representations of diagnosis codes and procedure information in EHR, and improve the performance of various medical prediction tasks [12, 65]. There are different approaches for word embedding (word representation) including skip-gram [39], continuous bag of words [39] and Glove [47]. In general, these approaches try to predict the probability of word given its context. So the vector represented for two words with similar context will be similar. Figure 7 illustrates our word embedding approach (based on continues bag of words) to represent the medical diagnosis codes as N-dimensions vectors for each patient. We implemented the word embedding using gensim library in python [51].



Diagnosis text	Context	Target
cardiovascular, chest pain, coronary artery, arrhythmias	(cardiovascular, chest pain)	cardiovascular
cardiovascular, **chest pain**, coronary artery, arrhythmias	(cardiovascular, chest pain) (chest pain, coronary artery)	chest pain
cardiovascular, chest pain**, coronary artery**, arrhythmias	(chest pain**,** coronary artery) (coronary artery, arrhythmias)	coronary artery
cardiovascular, chest pain, coronary artery, **arrhythmias**	(coronary artery, arrhythmias)	arrhythmias

Example of training input with window size =1

Fixed length (N) vector representation of each word	
cardiovascular	$[V_1^1, V_2^1, V_3^1, \dots\dots, V_{10}^1]$
chest pain	$[V_1^2, V_2^2, V_3^2, \dots\dots, V_{10}^2]$
coronary artery	$[V_1^3, V_2^3, V_3^3, \dots\dots, V_{10}^3]$
arrhythmias	$[V_1^4, V_2^4, V_3^4, \dots\dots, V_{10}^4]$

The final word-vector representation of diagnosis text for each patient is calculated based on average of word-vectors in each diagnosis text.

Figure 7: Word embedding for vector representation of text features

As shown in Figure7, the input of word embedding approach is all words of a text feature (column) such as diagnosis codes. The table in step 1 is demonstrating how context and target is defined in training set based on window size of 1. Step 2 is word embedding method (continues bag of words) that captures context and target from training set and train a shallow neural network. Step 3 is the vector representation of each word which obtained from the hidden layer for trained network in step 2, and finally in step 4, the vector for diagnosis text related to each patient is calculated based on average of vectors of that text's words.

4.2 Step−2: Feature Representation using Deep Learning

In the second step, all features will be represented in higher-level abstraction by four different deep autoencoder networks separately: 1) Stacked sparse Autoencoder (SSAE), 2) Deep Belief Network (DBN), 3) Adversarial Autoencoder (AAE) and 4) Variational Autoencoder (VAE). The performance of each network could be various in the different cases and it is necessary to consider hyper-parameters tuning such as learning rate, batch size, number of epochs, and number of hidden layers and hidden units for precise training to avoid over-fitting.

4.3 Step−3: Supervised Learning

In this step, we apply supervised learning models on the top of represented dataset for all four feature extraction approaches. Once the features are extracted, these representations from main dataset are entered in a linear and non-linear supervised regression models such as Random Forests, SVM and LASSO for prediction. Finally we evaluate and compare the performance of feature learning step by RMSE based on the prediction models.

4.4 Experimental Study

In our experimental study, we implement our methodology on three different EHRs datasets. First, we use a small dataset related to cardiovascular disease with high dimensional features, then we apply our method on two large datasets from eICU collaborative research database. This study design (considering small and large datasets) helps us to discover the performance of our method in different scenarios and compare the choice of representation learning for each one.

4.4.1 Experiment 1- Small Dataset

In the first experimental study, we used data related to a group of African-American patients diagnosed with cardiovascular disease at high risk of heart attack. This data received from emergency department at Detroit Medical Center (DMC). Cardiovascular disease is one of the main causes for death in the United State which has different risk level among racial subgroups. For instance, the risk of cardiovascular disease in African-Americans is higher than white patients. In addition, Hypertension is significantly prevalent among African-Americans and impacts directly to the risk of stroke and heart failure.

The size of dataset used in this case study is small related to 91 patients with 172 attributes after cleaning and preprocessing step. This data includes several attributes such as demographic information, patient medical history, laboratory test results and cardiovascular related measurements. However, we used Left Ventricular Mass Index (LVMI) as the target of prediction task which is an important risk factor in cardiovascular disease and has significant cost of measurement.

We implemented Deep Belief Network, Variational autoencoder and Adversarial autoencoder by using TensorFlow and Theano libraries and executed Stacked sparse autoencoder by Keras library with tensorflow backend in Python. All authoencoders except DBN are applied with 5 hidden layers (two hidden layers of encoders and decoders and one middle layer). We performed deep belief network with 3 hidden layers.

For each deep architecture, we applied parameter tuning for major parameters such as learning rate, activation functions and batch size to select the best parameters. We performed different deep networks which differ in the number of neurons in hidden layers for all autoencoders type and then select the best performance of autoencoders across all networks.

For the supervised learning step we consider three well-known supervised classifiers: Random Forests [5], Lasso Regression [59] and Support Vector Machine (SVM) [55]. We used Root Mean Squared Error (RMSE) as our evaluation measure for performance validation in testing process. We evaluated the prediction performance in each combination of deep represented data and supervised learners (e.g. VAE for features representation and Random Forest for prediction task) using the average results of 5-folds cross validation process (for each fold we considered 80% of the data for training, 10% for test and 10% for validation set). We used the weights learned in training process to represent the data in the testing and validation processes. In addition, we performed the same approach for finding the performance of each supervised learners on original data (unrepresented data).

The comparison results have been illustrated in Table 2. Based on the RMSE results, in general our framework improves the performance of prediction task using deep representation learning. In addition, combination of stacked sparse autoencoders for feature representation and Random Forests for supervised learning achieved the least RMSE in the case of small dataset.

Table 2: Performance comparison among represented data and original features (DMC dataset)

Approach	RF	Lasso	SVM
SSAE	6.89	**9.53**	**9.31**
DBN	7.91	9.81	10.02
AAE	8.49	9.89	10.06
VAE	9.65	10.17	9.95
Original	11.08	13.86	12.16

4.4.2 Case study 2 (Large Datasets): eICU dataset

In the second case study, we used the eICU collaborative research database: a large, publicly available database provided by the MIT Laboratory in partnership with the Philips eICU Research Institute [23]. Medical doctors predict intensive care units (ICUs) length of stay for planning ICU capacity as an expensive unit in the hospital and identifying unexpectedly long ICU length of stay in special cases to better monitoring [61]. The care provided by ICUs is complicated and the related costs are high, so ICUs are particularly interested in evaluating, planning and improving their performance [60].

The most popular approach for prediction of length of stay in ICUs is developed based on acute physiology score of APACHE (Acute Physiology and Chronic Health Evaluation) which lead to poor prediction performance in several cases [61]. APACHE introduced in 1978 for developing of severity-of-illness classification system and proposing a measure for describing different groups in ICUs and assessing their care [63]. APACHE approaches use multivariate linear regression procedure based on acute physiology score and some other variables such as age and chronic health conditions [74] to predict length of stay.

The data in the eICU database includes patients who were admitted to intensive care units during 2014 and 2015. Among different patients, we choose cardiovascular and Neurological patients admitted in the Cardiac-eICU and Neuro-eICU respectively. We integrated several features including hospital and administration data, demographics information, diagnosis and laboratory test data, drugs information, monitored invasive vital sign data and clinical patient history data. After cleaning and preprocessing step, we finalized more than 150 features for each dataset with approximately 7000 and 8000 records belonging to Cardiac-eICU and Neuro-eICU units, respectively. In this case study, we conduct the same approach as we did for the first case study and our purpose is to predict the patient length of stay (days) in these two ICU units based on high-dimensional features.

Table 3 and 4 demonstrate RMSE results for different types of autoencoders and supervised learners in Cardiac-ICU and Neuro-ICU data respectively. Although, AAE shows a great performance, VAE's results are impressive and provides perfect prediction for length of stay. As illustrated, SSAE and DBN could not improve the model accuracy as much as AAE and VAE.

We also considered the prediction results of APACHE approach reported in eICU collaborative research database and calculated the RMSE for both Cardiac and Neuro patients. The results indicated weak performance of APACHE approach, for instance, the RMSE for Cardiac and Neuro datasets were 8.04 and 8.12 respectively. Therefore, different machine learning approaches using original data or represented data outperformed APACHE approach significantly.

Table 3: Performance comparison between represented data and baseline (Cardiac-ICU)

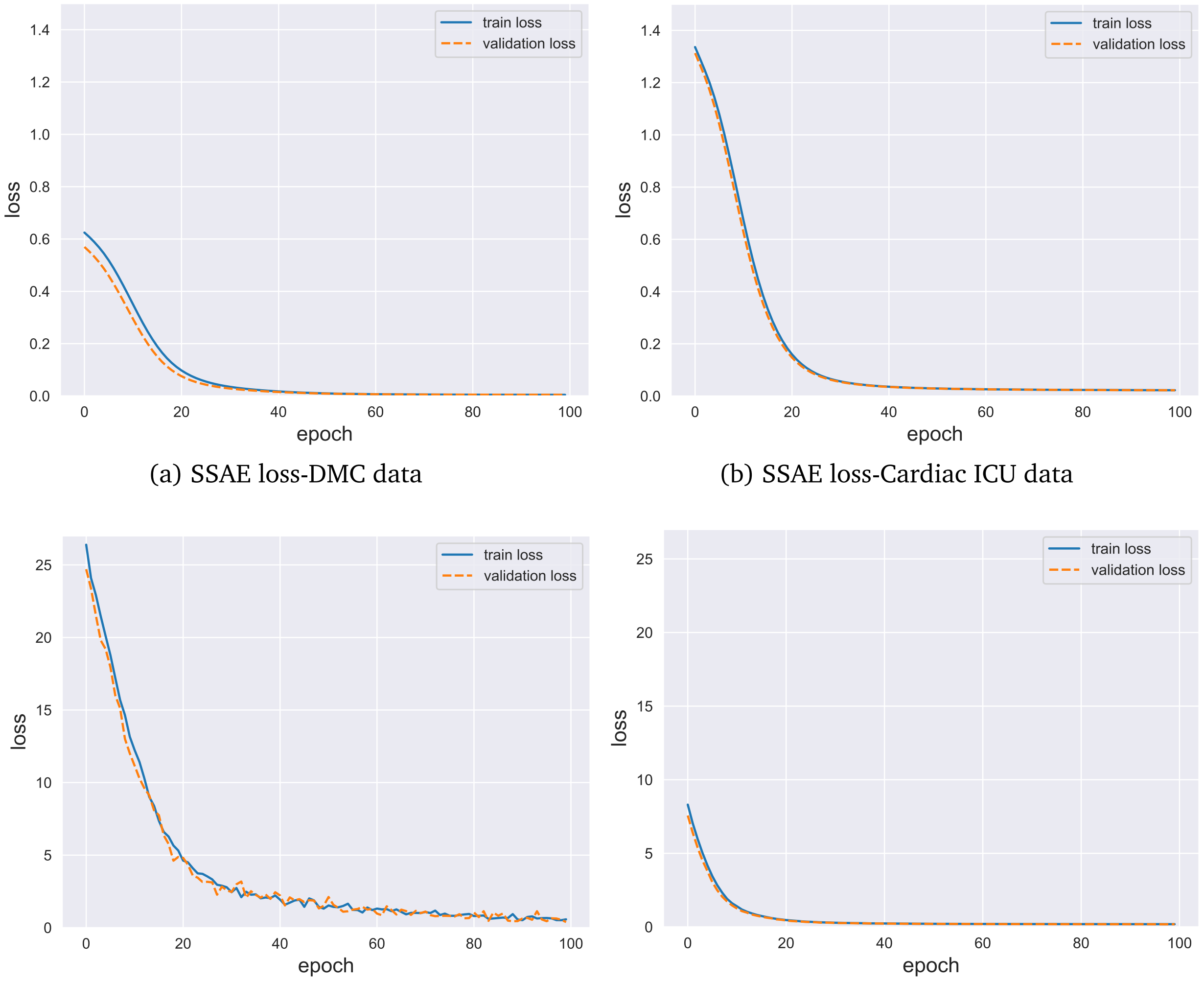
Approach	RF	Lasso	SVM
SSAE	1.59	4.08	2.47
DBN	0.97	3.78	2.29
AAE	0.57	3.57	1.99
VAE	0.20	**2.83**	**1.87**
Original	1.65	4.14	2.52

Table 4: Performance comparison between represented data and baseline (Neuro-ICU)

Approach	RF	Lasso	SVM
SSAE	1.31	2.76	3.34
DBN	0.95	2.53	2.06
AAE	0.72	2.28	2.45
VAE	0.14	**2.05**	**1.94**
Original	1.38	2.95	3.51

To better understanding of SSAE and VAE performance in the representation learning, we analyze their training and validation loss for small and large datasets separately. As illustrated in Figures 8a, 8b, 8c and 8d, the training and validation loss are end up to be roughly the same and also their values are converging (good fitting). Since we used regularization in both SSAE and VAE, it leads to have less amount of loss in validation set rather than training. Based on Figures 8a and 8b, the SSAE loss (which is based on

MAE loss function) in small dataset (DMC data) is less than SSAE loss of large dataset (Cardiac-ICU) across 100 epochs, hence the SSAE achieves better representation for small dataset. In the other side (Figures 8c and 8d), VAE loss (which is based on MSE of reconstruction error + average of KL loss) in large dataset is lower than VAE loss in small dataset, therefore provides better representation learning.



(a) SSAE loss-DMC data

(b) SSAE loss-Cardiac ICU data

(c) VAE loss-DMC data

(d) VAE loss-Cardiac ICU data

Figure 8: Training and validation loss of SSAE and VAE in training process for small and large datasets

CHAPTER 5 CONCLUSION AND FUTURE WORKS

In this research, we proposed a comparative study for evaluation of deep feature representation in applications to Electronic Health Records (EHRs). Our Deep Integrated Prediction approach discovers the complexity and dependencies in the EHRs using unsupervised learning (feature representation) which improves the clinical prediction performance significantly. The proposed approach consists three steps: data preprocessing, deep feature representation and supervised learning. We applied our approach on two different experimental studies while the first one is related to a small dataset of African-Americans in high risk of cardiovascular disease and second one includes two larger datasets from eICU collaborative research database. In the first case study we try to predict the heart failure risk level using EMR and in the second one, the goal is to predict length of stay in ICU units based on personalized patient attributes such as demographics data, diagnosis history, medication information and laboratory test results.

In both case studies, we used four different deep architectures (SSAE, DBN, AAE and VAE) for representation learning of EMRs. We considered different training parameters in each network (including number of hidden units, batch size, number of epochs and learning rate). Then, we performed three well-known linear and non-linear supervised learners (Random Forests, Lasso Regression and SVM) on the top of represented features and original features.

According to our results, the choice of deep representation achieves different performance in prediction tasks among small and large datasets. In the first experimental study, regular autoencoders (SSAE, DBN) had a better accuracy in comparison with advanced generative autoencoders (AAE, VAE) and in the second study with large datasets (eICU database), Variational Auto-encoder outperforms the other deep architectures significantly. In general, use of representation learning improves the accuracy of prediction tasks for both small and large datasets while the choice of deep architecture leads to different performance. The generative networks like as AAE and VAE try to find true distribution of original variables based on distribution of samples from latent variables (middle layer) and provides better representation in case of larger datasets.

Empirically, our results demonstrate that: 1) Medical feature representation can improve the performance of prediction and 2) Choice of representation can lead to different performances which should be selected appropriately. This choice of representation might be related to the number of instances (n) and number of features (p) in the dataset. For future works, here are some directions which can extend this research:

- It is necessary to compare the performance of generative models (e.g. AAE and VAE) with some other approaches (e.g. SSAE and DBN) in different scenarios. As a guideline, we need to consider four scenarios: a) large n and large p, b) large n and small p, c) small n and large p, d) small n and small p and generate some insights about choice of representation between advanced generative models and otherwise.

- The proposed approach can be applied on different domains and datasets. For example, choice of representation learning for images or time series data would lead to interesting results.

- Although representation learning can improve the performance of supervised learning tasks, it makes them hard to interpret the prediction model (the black box). In other words, after feature representation, it will be difficult to discover the most important features and their relationship to the desired target variable (model interpretability). Therefore, further research can focus on finding interpretive approach for deep representation and provide a trade-off between accuracy and interpretability of deep representation.

REFERENCES

[1] E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

[2] G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach. *Neural computation*, 12(10):2385–2404, 2000.

[3] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

[4] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *Advances in neural information processing systems*, pages 153–160, 2007.

[5] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[6] G. Camps-Valls, J. L. Rojo-Álvarez, M. Martínez-Ramón, et al. *Kernel methods in bioengineering, signal and image processing*. Idea Group Pub., 2007.

[7] R. Cao, D. Bhattacharya, J. Hou, and J. Cheng. Deepqa: improving the estimation of single protein model quality with deep belief networks. *BMC bioinformatics*, 17(1):495, 2016.

[8] 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 *Data Mining (ICDM), 2017 IEEE International Conference on*, pages 787–792. IEEE, 2017.

[9] 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.

[10] E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare Conference*, pages 301–318, 2016.

[11] E. Choi, M. T. Bahadori, E. Searles, C. Coffey, M. Thompson, J. Bost, J. Tejedor-Sojo, and J. Sun. Multi-layer representation learning for medical concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504. ACM, 2016.

[12] E. Choi, A. Schuetz, W. F. Stewart, and J. Sun. Medical concept representation learning from electronic health records and its application on heart failure prediction. *arXiv preprint arXiv:1602.03686*, 2016.

[13] L. Cui, X. Xie, and Z. Shen. Prediction task guided representation learning of medical codes in ehr. *Journal of biomedical informatics*, 84:1–10, 2018.

[14] X. Deng and F. Huangfu. Collaborative variational deep learning for healthcare recommendation. *IEEE Access*, 7:55679–55688, 2019.

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

[16] S. S. Fazeli, S. Venkatachalam, R. B. Chinnam, and A. Murat. Two-stage stochastic choice modeling approach for electric vehicle charging station network design in urban communities. *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[17] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[18] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[20] G. E. Hinton, S. Osindero, and Y.-W. Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[21] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *Advances in neural information processing systems*, pages 857–864, 2003.

[22] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[23] A. E. Johnson, T. J. Pollard, L. A. Celi, and R. G. Mark. Analyzing the eicu collaborative research database. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 631–631. ACM, 2017.

[24] M. Kallas, C. Francis, L. Kanaan, D. Merheb, P. Honeine, and H. Amoud. Multi-class svm classification combined with kernel pca feature extraction of ecg signals. In *2012 19th International Conference on Telecommunications (ICT)*, pages 1–5. IEEE, 2012.

[25] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[26] I. Landi, B. S. Glicksberg, H.-C. Lee, S. Cherng, G. Landi, M. Danieletto, J. T. Dudley, C. Furlanello, and R. Miotto. Deep representation learning of electronic health records to unlock patient stratification at scale. *arXiv preprint arXiv:2003.06516*, 2020.

[27] D. Learning. Computer science department. *Stanford University. http://ufldl. stanford. edu/tutorial*, 20:21–22, 2013.

[28] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[29] B. Li, C.-H. Zheng, D.-S. Huang, L. Zhang, and K. Han. Gene expression data classification using locally linear discriminant embedding. *Computers in Biology and Medicine*, 40(10):802–810, 2010.

[30] X. Li, D. Zhu, M. Dong, M. Z. Nezhad, A. Janke, and P. D. Levy. Sdt: A tree method for detecting patient subgroups with personalized risk factors. *AMIA Summits on Translational Science Proceedings*, 2017.

[31] X. Li, D. Zhu, and P. Levy. Leveraging auxiliary measures: a deep multi-task neural network for predictive modeling in clinical research. *BMC medical informatics and decision making*, 18(4):45–53, 2018.

[32] Y. Li, C.-Y. Chen, and W. W. Wasserman. Deep feature selection: theory and application to identify enhancers and promoters. *Journal of Computational Biology*, 23(5):322–336, 2016.

[33] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[34] S. Ma and M. R. Kosorok. Identification of differential gene pathways with principal component analysis. *Bioinformatics*, 25(7):882–889, 2009.

[35] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[36] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey. Adversarial autoencoders. *arXiv preprint arXiv:1511.05644*, 2015.

[37] R. J. Martis, U. R. Acharya, K. Mandana, A. K. Ray, and C. Chakraborty. Application of principal component analysis to ecg signals for automated diagnosis of cardiac health. *Expert Systems with Applications*, 39(14):11792–11800, 2012.

[38] D. Mateus, C. Wachinger, S. Atasoy, L. Schwarz, and N. Navab. Learning manifolds: design analysis for medical applications. In *Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis*, pages 374–402. IGI Global, 2012.

[39] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[40] 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. *Scientific reports*, 6:26094, 2016.

[41] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics*, page bbx044, 2017.

[42] M. Z. Nezhad, N. Sadati, K. Yang, and D. Zhu. A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer. *Expert Systems with Applications*, 115:16–26, 2019.

[43] M. Z. Nezhad, D. Zhu, X. Li, K. Yang, and P. Levy. Safs: A deep feature selection approach for precision medicine. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 501–506. IEEE, 2016.

[44] A. Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.

[45] G. Nyamundanda, L. Brennan, and I. C. Gormley. Probabilistic principal component analysis for metabolomic data. *BMC bioinformatics*, 11(1):571, 2010.

[46] K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[47] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[48] S. Purushotham, C. Meng, Z. Che, and Y. Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of biomedical informatics*, 83:112–134, 2018.

[49] S. Rajendran, A. Ansaripour, M. Kris Srinivasan, and M. J. Chandra. Stochastic goal programming approach to determine the side effects to be labeled on pharmaceutical drugs. *IISE Transactions on Healthcare Systems Engineering*, 9(1):83–94, 2019.

[50] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang. Deep learning for health informatics. *IEEE journal of biomedical and health informatics*, 21(1):4–21, 2017.

[51] R. Rehuvek and P. Sojka. GensimâĂŤstatistical semantics in python. *Retrieved from genism. org*, 2011.

[52] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[53] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

[54] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2018.

[55] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[56] P. Tabacof, J. Tavares, and E. Valle. Adversarial images for variational autoencoders. *arXiv preprint arXiv:1612.00155*, 2016.

[57] E. Taghizadeh. Utilizing artificial neural networks to predict demand for weather-sensitive products at retail stores. *arXiv preprint arXiv:1711.08325*, 2017.

[58] J. B. Tenenbaum, V. De Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[59] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[60] I. W. Verburg, N. F. de Keizer, E. de Jonge, and N. Peek. Comparison of regression methods for modeling intensive care length of stay. *PloS one*, 9(10):e109684, 2014.

[61] I. W. M. Verburg, A. Atashi, S. Eslami, R. Holman, A. Abu-Hanna, E. de Jonge, N. Peek, and N. F. de Keizer. Which models can i use to predict adult icu length of stay? a systematic review. *Critical care medicine*, 45(2):e222–e231, 2017.

[62] 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. *Journal of Machine Learning Research*, 11(Dec):3371–3408, 2010.

[63] D. P. Wagner and E. A. Draper. Acute physiology and chronic health evaluation (apache ii) and medicare reimbursement. *Health care financing review*,

1984(Suppl):91, 1984.

[64] L. Wang, M. Dong, E. Towner, and D. Zhu. Prioritization of multi-level risk factors for obesity. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1065–1072. IEEE, 2019.

[65] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu. A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20, 2018.

[66] D. Wulsin, J. Blanco, R. Mani, and B. Litt. Semi-supervised anomaly detection for eeg waveforms using deep belief nets. In *2010 Ninth International Conference on Machine Learning and Applications*, pages 436–441. IEEE, 2010.

[67] C. Xiao, E. Choi, and J. Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 2018.

[68] P. Xu, G. N. Brock, and R. S. Parrish. Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Computational Statistics & Data Analysis*, 53(5):1674–1687, 2009.

[69] F. Yao, J. Coquery, and K.-A. Lê Cao. Independent principal component analysis for biologically meaningful dimension reduction of large biological data sets. *BMC bioinformatics*, 13(1):24, 2012.

[70] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo. Validating clustering for gene expression data. *Bioinformatics*, 17(4):309–318, 2001.

[71] J. Zhao, P. Papapetrou, L. Asker, and H. Boström. Learning from heterogeneous temporal data in electronic health records. *Journal of biomedical informatics*, 65:105–

119, 2017.

[72] L. Zhao and K. Jia. Multiscale cnns for brain tumor segmentation and diagnosis. *Computational and mathematical methods in medicine*, 2016, 2016.

[73] G. Zhong, X. Ling, and L.-N. Wang. From shallow feature learning to deep learning: Benefits from the width and depth of deep architectures. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(1):e1255, 2019.

[74] J. E. Zimmerman, A. A. Kramer, D. S. McNair, and F. M. Malila. Acute physiology and chronic health evaluation (apache) iv: hospital mortality assessment for todayâĂŹs critically ill patients. *Critical care medicine*, 34(5):1297–1310, 2006.

[75] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

ABSTRACT

REPRESENTATION LEARNING WITH AUTOENCODERS FOR ELECTRONIC HEALTH RECORDS

by

NAJIBESADAT SADATIJAFARKALAEI

August 2020

Advisor: Dr. Dongxiao Zhu

Major: Computer Science

Degree: Master of Science

Increasing volume of Electronic Health Records (EHR) in recent years provides great opportunities for data scientists to collaborate on different aspects of healthcare research by applying advanced analytics to these EHR clinical data. A key requirement however is obtaining meaningful insights from high dimensional, sparse and complex clinical data. Data science approaches typically address this challenge by performing feature learning in order to build more reliable and informative feature representations from clinical data followed by supervised learning. In this research, we propose a predictive modeling approach based on deep feature representations and word embedding techniques. Our method uses different deep architectures (stacked sparse autoencoders, deep belief network, adversarial autoencoders and variational autoencoders) for feature representation in higher-level abstraction to obtain effective and robust features from EHRs, and then build prediction models on top of them. Our approach is particularly useful when the unlabeled data is abundant whereas labeled data is scarce. We investigate the performance of representation learning through a supervised learning approach. Our focus is to present a comparative study to evaluate the performance of different deep architectures through supervised learn-

ing and provide insights for the choice of deep feature representation techniques. Our experiments demonstrate that for small data sets, stacked sparse autoencoder demonstrates a superior generality performance in prediction due to sparsity regularization whereas variational autoencoders outperform the competing approaches for large data sets due to its capability of learning the representation distribution.

AUTOBIOGRAPHICAL STATEMENT

Najibesadat Sadatijafarkalaei is a Ph.D candidate in Industrial and Systems Engineering department at Wayne State University, Detroit, Michigan. She is also pursuing her second Master's degree in Computer Science at the same university. She received her first Master's degree in the Industrial and Systems Engineering from AmirKabir University of Technology in Tehran, Iran, 2013. Najibesadat got her bachelors' degree in Industrial Engineering from Sharif University of Technology in Tehran, Iran, 2011. Her main research interests are Machine Learning, Healthcare Data Analytics, Applied Operation Research and Operation Management. Najibesadat's papers have been published in top conferences and journals and here is the link for her Google scholar: https://scholar.google.com/citations?user=6FsrzF4AAAAJ&hl=en