Prediction and Informative Risk Factor Selection of Bone Diseases

Hui Li, Xiaoyi Li, Murali Ramanathan, and Aidong Zhang, Fellow, IEEE

Abstract—With the booming of healthcare industry and the overwhelming amount of electronic health records (EHRs) shared by healthcare institutions and practitioners, we take advantage of EHR data to develop an effective disease risk management model that not only models the progression of the disease, but also predicts the risk of the disease for early disease control or prevention. Existing models for answering these questions usually fall into two categories: the expert knowledge based model or the handcrafted feature set based model. To fully utilize the whole EHR data, we will build a framework to construct an integrated representation of features from all available risk factors in the EHR data and use these integrated features to effectively predict osteoporosis and bone fractures. We will also develop a framework for informative risk factor selection of bone diseases. A pair of models for two contrast cohorts (e.g., diseased patients vs. non-diseased patients) will be established to discriminate their characteristics and find the most informative risk factors. Several empirical results on a real bone disease data set show that the proposed framework can successfully predict bone diseases and select informative risk factors that are beneficial and useful to guide clinical decisions.

Index Terms—Electronic health records (EHRs); risk factor analysis; integrated feature extraction; risk factor selection; disease memory; osteoporosis; bone fracture.

1 INTRODUCTION

Risk factor (RF) analysis based on patients’ electronic health records (EHRs) is a crucial task of epidemiology and public health. Usually, people treat variables in EHR data as numerous potential risk factors (RFs) that need to be considered simultaneously for assessing disease determinants and predicting the progression of the disease, for the purpose of disease control or prevention. More importantly, some common diseases may be clinically silent but can cause significant mortality and morbidity after onset. Unless early prevented or treated, these diseases will affect the quality of life, and increase the burden of healthcare costs. With the success of RF analysis and disease prediction based on an intelligent computational model, unnecessary tests can be avoided. The information can assist in evaluating the risk of the occurrence of disease, monitor the disease progression, and facilitate early prevention measures. In this paper, we focus on the study of osteoporosis and bone fracture prediction.

Over the past few decades, osteoporosis has been recognized as an established and well-defined disease that affects more than 75 million people in the United States, Europe and Japan, and it causes more than 8.9 million fractures annually worldwide [7]. It’s reported that 20-25% of people with a hip fracture are unable to return to independent living and 12-20% die within one year. In 2003, the World Health Organization (WHO) embarked on a project to integrate information on RFs and bone mineral density (BMD) to better predict the fracture risk in men and women worldwide [7]. Osteoporosis in the vertebrae can cause serious problems for women such as bone fracture. The diagnosis of osteoporosis is usually based on the assessment of BMD measured by dual energy X-ray absorptiometry (DXA). Different from osteoporosis measured by BMD, bone fracture risk is determined by the bone loss rate and various factors such as demographic attributes, family history and life style. Some studies have stratified their analysis of fracture risk into those who are fast or slow bone losers. With a faster rate of bone loss, people have a higher risk of fracture [7].

Osteoporosis and bone fracture are complicated diseases. As shown in Fig. ??, they are associated with various potential RFs such as demographic attributes, patients’ clinical records regarding disease diagnoses and treatments, family history, diet, and life style. Different representations might entangle the different explanatory reasons of variation behind various RFs and diseases. Some of the fundamental questions have been attracting researchers’ interest in this area, for example, how to perform feature extraction and select the integrated significant features? Also, what are appropriate approaches for manifold feature extraction and maintaining the real and intricate relationships between a disease and its potential RFs? A good representation has an advantage for capturing un-
derlying factors with shared statistical strength for predicting bone diseases. A representation-learning model discovers explanatory factors behind shared intermediate representation with the combination of knowledge from both input data and output specific tasks. The rich interactions among numerous potential RFs or between RFs and a disease can complicate our final prediction tasks. Besides, the other types of questions we aim to address in this paper are: what are the informative RFs from the whole list of RFs? Whether patients can change some modifiable RFs for delaying the onset and progression of bone diseases. The proposed approach will show some good properties for answering these questions.

Traditionally, the assessment of the relationship between a disease and a potential RF is achieved by finding statistically significant associations using the regression model such as linear regression, logistic regression, Poisson regression, and Cox regression [7], [8], [9], [10]. Although these regression models are theoretically acceptable for analyzing the risk dependence of several variables, they pay little attention to the nature of the RFs and the disease itself. Sometimes, less than ten significant RFs are fed into those models, which are not intelligent enough for predicting a complicated disease such as osteoporosis. Other data mining studies under this objective are association rules [11], decision tree [12] and Artificial Neural Network (ANN) [13]. For these methods, it’s ineffective to build a comprehensive model that can guide medical decision-making if there are a number of potential RFs to be studied simultaneously. Usually limited RFs are selected based on the knowledge of physicians since handling a lot of features is computationally expensive. Feature selection techniques are well applied to select limited number of RFs before feeding to a classifier. However, the feature selection problem is known to be NP-hard [14] and more importantly, those abandoned RFs might still contain valuable information. Furthermore, the performance of ANN depends on a good setting of meta parameters and so parameter-tuning is an inevitable issue. Under these scenarios, most of these traditional data mining approaches may not be effective.

Mining the causality relationship between RFs and a specific disease has attracted considerable research attention in recent years. In [7], [8], [10], limited RFs are used to construct a Bayesian network and those RFs are assumed conditionally independent of one another. It is noteworthy to mention that the random forest decision tree has been investigated for identifying RFs? The data in this work is processed using FRAX [15]. Although this is a popular fracture risk assessment tool developed by WHO, it may not be appropriate to directly adopt the results from this prediction tool for evaluating the validity of an algorithm since FRAX sometimes overestimates or underestimates the fracture risk [16]. The prediction results from FRAX need to be further interpreted with caution and properly re-evaluated. Some hybrid data mining approaches might also be used to combine classical classification methods with feature selection techniques for the purpose of improving the performance or minimizing the computational expense for a large data set [17], but they are limited by the challenge of explaining selected features.

The existing methods for predicting osteoporosis and bone fracture are all based on expert knowledge or handcrafted features. Besides, both approaches are time-consuming, brittle, and incomplete. To solve these problems, we propose a RF learning model including learning an abstract representation for predicting the bone related diseases and selecting the most influential RFs that cause the disease progression as shown in Fig. 2. In this generalized RF learning pipeline, we apply all variables of EHR data as RFs into the Risk Factor Learning module which includes three tasks: (1) RF extraction is used to produce integrated features which show combinations of multiple nonlinear RF transformations with the goal of yielding more abstract and salient RF representations, (2) RF selection aims at choosing a subset of RFs from a pool of candidates showing that informativeness enables statistically significant improvements in disease prediction and (3) Expert RF is extracted based on the domain expert knowledge to validate the performance.
of both RF extraction and RF selection. For such a framework, we are facing three main challenges:

- The performance of the follow-up analysis will be highly dependent on how well the integrated features capture the underlying causes of the disease and the predictive power using those integrated features. To obtain the latent variables from large amounts of entangled RFs, we are actually facing the problem of learning an extraction and representation which can best disentangle the salient integrated features from original complex data.
- It is difficult to discriminate the different roles of seemingly independent features for both healthy individuals and diseased individuals. Selecting the informative RFs are beneficial and useful to guide clinical decision. Besides, these informative RFs can save budget and time for physicians to predict health conditions. Therefore, our model should handle with some problems such as what are the most important RFs that contribute to the development of diseases? How many RFs do we need to achieve a good predictive performance?
- EHR data are diverse, multi-dimensional, large in size, with missing and noisy values, and without ground truth in nature. These properties make existing methods inapplicable because of the lack of extensive learning power and overall model expressiveness. A state-of-the-art model should expressiveness. A state-of-the-art model should be carefully designed for handling those questions simultaneously.

In this paper, we propose a novel approach for the study of bone diseases in two aspects: bone disease prediction and disease RF selection according to the significance. For clear understanding, we define Disease Memory (DM) as a model trained by a specific group of samples aiming to memorize the underlying characteristics for this group. In addition, we apply all samples in our model and train a general model which captures the characteristics for both diseased patients and non-diseased patients to predict an unknown sample, denoted by the comprehensive disease memory (CDM) model. Our model is separately trained using diseased samples and non-diseased samples to distinguish their different properties. Bone disease memory (BDM) is a type of DM model which is trained by diseased samples and so it only memorizes the characteristics of those patients who suffer from bone diseases. Similarly, the non-disease memory (NDM) is a model which is trained by the non-diseased samples and memorizes their attributes. We individually train them because we want to find informative RFs which can be used to distinguish the diseased individuals from non-diseased ones. In other words, different DM models increase the flexibility for exploring different tasks. DM serves as an important embedded module in our framework that has the following nice properties. First, diseased patients and healthy patients are modeled together to establish a CDM which captures the salience of all RFs by a limited number of integrated features for predicting bone diseases. Second, diseased patients and healthy patients are modeled separately based on their unique characteristics to find the RFs that cause the disease. Third, our model is robust in the presence of missing and noisy data. Last but not least, the model doesn’t require all samples are labeled, instead, it can be trained in a semi-supervised fashion. These nice properties are achieved by our proposed model - a deep graphical model focusing on the bone disease. Recently, many efforts have been devoted to develop learning algorithms for the deep graphical model with impressive results obtained in various areas such as computer vision and natural language processing [?]. The intuition and extensive learning power of these models are suitable for our task. To the best of our knowledge, our method is the first work on risk factor analysis using a deep learning method which can handle high-dimensional and imbalanced data and interpret hidden reasons behind bone diseases.

2 Overview of Our System

In this section, we define our problem by showing a pipeline for the whole framework. Generally speaking, our proposed system contains a two-task framework, as shown in Fig. ?? The upper component of Fig. ?? shows the roadmap for the first task: the bone disease prediction based on integrated features. The bottom component of Fig. ?? shows the roadmap for the second task: informative RF selection. Given patients’ information, our system can not only predict the risk of osteoporosis and bone fractures, but also rank the informative RFs and explain the semantics of each RF. The description of each component is given as follows.

Task 1 – The Bone Disease Prediction Component

In this component, we feed the original data set to the comprehensive disease memory (CDM) which is a trained model of the intermediate representation of the original RFs. The training procedure of CDM includes two steps: pre-training and fine-tuning. In
the pre-training step, we train CDM in an unsupervised fashion. This pre-training procedure aims at capturing the characteristics among all RFs. In the fine-tuning step, we focus on the training using two types of labeled information (osteoporosis and bone loss rate). We use a greedy layered learning algorithm to train a two-layer deep belief network (DBN) which is the underlying structure of CDM. All RFs in the original data are projected onto a new space with the lower dimensionality by restricting the number of units in the output layer of DBN. Therefore, the CDM module extracts the integrated risk features from the original data set. These lower-dimensional integrated risk features are a new representation of the original higher-dimensional RFs, which will be evaluated by a two-phase prediction module. In Phase 1, we predict the risk of osteoporosis for all test samples. The osteoporotic bones are labeled as the positive output and the normal bone as the negative output. Because the osteoporotic patients tend to have more severe bone fractures, in Phase 2, we further predict the risk of bone loss rate for all positive samples from Phase 1. The high bone loss rate, as the positive output, reveals the higher possibility to have bone fractures, and the low bone loss rate is defined as the negative output of Phase 2.

Task 2 – The Informative RF selection Component. Although the integrated features generated in the first component can be used to effectively predict bone diseases, it is difficult to directly relate the semantics of the integrated features to individual patients. Thus, in this component, we propose to select the most meaningful and significant RFs. Instead of using all samples in the training procedure, we first split the original data set into two parts: diseased samples and non-diseased samples. In the training procedure, we separately train the bone disease memory (BDM) model using the diseased samples and the non-diseased samples in the original data set. Therefore, the CDM module extracts the integrated risk features from the original data set. These lower-dimensional integrated risk features are a new representation of the original higher-dimensional RFs, which will be evaluated by a two-phase prediction module. In Phase 1, we predict the risk of osteoporosis for all test samples. The osteoporotic bones are labeled as the positive output and the normal bone as the negative output. Because the osteoporotic patients tend to have more severe bone fractures, in Phase 2, we further predict the risk of bone loss rate for all positive samples from Phase 1. The high bone loss rate, as the positive output, reveals the higher possibility to have bone fractures, and the low bone loss rate is defined as the negative output of Phase 2.

3 Methodology

In this section, we first briefly describe the evolution of the energy models as the preliminaries to our proposed method. Then we introduce single-layer and multi-layer learning approaches to construct our different disease memories. Finally, we propose our model focusing on the prediction and informative RF selection for bone diseases.

3.1 Preliminaries

3.1.1 Hopfield Net

A Hopfield network is a form of recurrent artificial neural network invented by John Hopfield. It serves as the content-addressable memory systems with the binary threshold nodes where each unit (node in the graph simulating the artificial neuron) can be updated using the following rule:

\[
S_i = \begin{cases} 
1 & \text{if } \sum_j W_{i,j} S_j > \theta_i, \\
-1 & \text{otherwise} 
\end{cases}
\]  

(1)

where \(W_{i,j}\) is the strength of the connection weight from unit \(j\) to unit \(i\), \(S_j\) is the state of unit \(j\), \(\theta_i\) is the threshold of unit \(i\). Based on Eq. (1) the energy of Hopfield Net is defined as,

\[
E = -\frac{1}{2} \sum_{i,j} W_{i,j} S_i S_j + \sum_i \theta_i S_i.
\]  

(2)

The difference in the global energy that results from a single unit \(i\) being 0 (off) versus 1 (on), denoted as \(\Delta E_i\), is given as follows:

\[
\Delta E_i = \sum_j w_{i,j} s_j + \theta_i.
\]  

(3)

Eq. (3) ensures that when units are randomly chosen to update, the energy \(E\) will either lower in value or stay the same. Furthermore, repeatedly updating the network will eventually converge to a state which is a local minima in the energy function (which is considered to be a Lyapunov function [7]). Thus, if a state is a local minimum in the energy function, it is a stable state for the network. Note that this energy function belongs to a general class of models in physics, under the name of Ising models. This in turn is a special case of Markov networks, since the associated probability measure, the Gibbs measure, has the Markov property.

3.1.2 Boltzmann Machines

Boltzmann machines (BM) can be seen as the stochastic, generative counterpart of Hopfield nets [7]. They are one of the first examples of a neural network capable of learning internal representations, and are able to represent and (given sufficient time) solve difficult combinatoric problems. The global energy in
a Boltzmann machine is identical in form to that of a Hopfield network, with the difference that the partial derivative with respect to each unit (Eq.(??)) can be expressed as the difference of energies of two states:

$$\Delta E_i = E_{i=off} - E_{i=on}. \quad (4)$$

If we want to train the network so that it will converge to a global state according to a data distribution that we have over these states, we need to set the weights making the global states with the highest probabilities which will get the lowest energies. The units in the BM are divided into “visible” units, $V$, and “hidden” units, $h$. The visible units are those which receive information from the data. The distribution over the data set is denoted as $P^+(V)$. After the distribution over global states converges and marginalizes over the hidden units, we get the estimated distribution $P^-(V)$ that is the distribution of our model. Then the difference can be measured using KL-divergence [?], and partial gradient of this difference will be used to update the network. But the computation time grows exponentially with the machine’s size, and with the magnitude of the connection strengths.

3.2 Single-Layer Learning for the Latent Reasons Underlying Observed RFs

To have a good RF representation of latent reasons for the data, we propose to use Restricted Boltzmann Machine (RBM) [?]. A RBM is a generative stochastic graphical model that can learn a probability distribution over its set of inputs, with the restriction that their visible units and hidden units must form a fully connected bipartite graph. Specifically, it has a single layer of hidden units that are not connected to each other and have undirected, symmetrical connections to a layer of visible units. We show a shallow RBM in Fig. ??(a). The model defines the following energy function: $E : \{0, 1\}^{D+F} \to \mathbb{R}$:

$$E(v, h; \theta) = -\sum_{i=1}^{D} v_i W_{ij} h_j - \sum_{i=1}^{D} b_i v_i - \sum_{j=1}^{F} a_j h_j, \quad (5)$$

where $\theta = \{a, b, W\}$ are the model parameters. $D$ and $F$ are the number of visible units and hidden units. The joint distribution over the visible and hidden units is defined by:

$$P(v, h; \theta) = \frac{1}{Z(\theta)} \exp(-E(v, h; \theta)), \quad (6)$$

where $Z(\theta)$ is the partition function that plays the role of a normalizing constant for the energy function.

Exact maximum likelihood learning is intractable in RBM. In practice, efficient learning is performed using Contrastive Divergence (CD) [?]. In particular, each hidden unit activation is penalized in the form:

$$\sum_{j=1}^{F} KL(\rho | h_j),$$

where $F$ is the total number of hidden units, $h_j$ is the activation of unit $j$ and $\rho$ is a predefined sparsity parameter, typically a small value close to zero (we use 0.05 in our model). So the overall cost of a sparse RBM used in our model is:

$$E(v, h; \theta) = -\sum_{i=1}^{D} v_i W_{ij} h_j - \sum_{i=1}^{F} b_i v_i - \sum_{j=1}^{F} a_j h_j + \beta \sum_{j=1}^{F} KL(\rho | h_j) + \lambda \|W\|,$n

where $\|W\|$ is the regularizer, $\beta$ and $\lambda$ are hyper-parameters.

The advantage of RBM is that it investigates an expressive representation of the input RFs. Each hidden unit in RBM is able to encode at least one high-order interaction among the input variables. Given a specific number of latent reasons in the input, RBM requires less hidden units to represent the problem complexity. Under this scenario, RFs can be analyzed by a RBM model with an efficient CD learning algorithm. In this paper, we use RBM for an unsupervised greedy layer-wise pre-training. Specifically, each sample describes a state of visible units in the model. The goal of learning is to minimize the overall energy so that the data distribution can be better captured by the single-layer model.

3.3 Multi-Layer Learning for Mining Abstractive Reasons

The new representations learned by a shallow RBM (one layer RBM) can model some directed hidden causalities behind the RFs. But there are more abstractive reasons behind them (i.e. the reasons of the reasons). To sufficiently model reasons in different abstractive levels, we can stack more layers into the shallow RBM to form a deep graphical model, namely, a DBN [?]. DBN is a probabilistic generative model that is composed of multiple layers of stochastic, latent variables. The latent variables typically have binary values and are often called hidden units. The

1. We tried different settings for both $\beta$ and $\lambda$ and found our model is not very sensitive to the input parameters. We fixed $\beta$ to 0.1 and $\lambda$ to 0.0001.
3.4 Integrated Risk Features for Bone Disease Prediction

Our goal is to disentangle the salient integrated features from the complex EHR data for the bone disease prediction. We propose to define a learning model based on the given data set for two types of bone disease prediction, osteoporosis and bone loss rate. Our general idea is shown in Fig. ??, where a good RF representation for predicting osteoporosis and bone loss rate is achieved by learning a set of intermediate representation using a DBN structure at bottom and a classifier is appended on it. This multi-learning model can capture the characteristics from both observed input (bottom-up learning) and labeled information (top-down learning). The internal model, which memorizes the trained parameters using the whole training data and preserves the information for both normal and abnormal patients, is termed as the comprehensive disease memory (CDM). That is, the learned representation model CDM discovers good intermediate representations that can be shared across two prediction tasks with the combination of knowledge from both input layer with the original training data and output layer with two types of class labels.

As shown in Algorithm ??, the training procedure for CDM concentrates on two specific prediction tasks (osteoporosis and bone loss rate) with all RFs as the input and model parameters as the output. It includes a pre-training stage and a fine-tuning stage. In the first stage, the unsupervised pre-training stage, we apply the layer-wise CD learning procedure for putting the parameter values in the appropriate range for further supervised training. It guides the learning towards basins of attraction of minima that support a good RF generalization from the training data set [?]. So the result of the pre-training procedure establishes an initialization point of the fine-tuning procedure inside a region of parameter space in which the parameters are henceforth restricted. In the second stage, the fine-tuning (FT) stage, we take advantage of label information to train our model in a supervised fashion. In this way, the prediction errors for both prediction tasks will be minimized. Specifically, we use parameters from the pre-training stage to calculate the prediction results for each sample and then back propagate the errors between the predicted result and the ground truth about osteoporosis from top to bottom to update model parameters to a better state. Since we have another type of labeled information, we then repeat the fine-tuning stage by calculating errors between the predicted result and another ground truth about bone loss rate. After the two-stage training procedure, our CDM is well trained and can be used to predict osteoporosis and bone loss rate simultaneously.

In Algorithm 1, lines 2 to 4 reflect a layer-wise Contrastive Divergence (CD) learning procedure where \( z \) is a predetermined hyper-parameter that controls how...
many Gibbs rounds for each sampling are completed and \( t_{+1} \) is the state of upper layer. In our experiments, we choose \( z = 1 \). The pre-training phase stops when all layers are exhausted. Lines 5 to 15 show a standard gradient update procedure (fine-tuning). We update model parameters from top to bottom by a simple two-step procedure. First, we update model parameters \( M \) by the gradient descent on the cost \( c \) for the training set. Second, we use early stopping as a type of regularization in our method to avoid overfitting. We compare the cost of the current step \( c_t \) with the previous step \( c_{t-1} \) for the validation set (holdout set) and halt the procedure when the validation error stops decreasing or starts to increase within 5 minibatches. Since we have ground truth \( g_1 \) and \( g_2 \) representing osteoporosis and bone loss rate, we implement the second fine-tuning procedure using \( g_2 \) after the stage using \( g_1 \). Moreover, we randomly dropout 30% hidden units for each layer for the purpose of alleviating the counter effect between different label information during the fine-tuning stage.

The main advantage of the DBN in the above training procedure is that it tends to display more expressive and invariant results than the single layer network and also reduces the size of the representation. This approach obtains a filter-like representation if we treat unit weights as the filters [2]. We want to filter the insignificant RFs and thus find out the robust and integrated features which are the fusion of both observed RFs and hidden reasons for predicting bone diseases.

The disease risk prediction requires building a predictive model such as a classifier for a specific disease condition using the integrated features in CDM. The new representation CDM of RFs extracted by a two-layer DBN can be served as the input of several traditional classifiers, as shown in Fig. ?? To incorporate labeled samples, we propose to add a regression layer on top of DBN to get classification results, which can be used to update the overall model using back propagation. Based on the proposed model in Fig. ??, physicians and researchers can assess the risk of a patient in developing osteoporosis or bone fracture. Then the proper intervention and care plan can be designed accordingly for the purpose of prevention or disease control.

### 3.5 Informative Risk Factor Selection

In the previous section, we have proposed CDM to model both diseased patients and healthy patients together and established a comprehensive disease memory which captures the salience of all RFs by a limited number of integrated features for predicting osteoporosis and bone loss rate. However, the informative RF selection aims to capture the differences between the diseased patients and non-diseased patients. Therefore, our CDM model cannot be applied to this task since it models over all patients. In this section, we propose to model the diseased patients and healthy patients separately based on their unique characteristics and identify the RFs that cause the disease (osteoporosis). Two variants of disease memory will be introduced to conduct the informative RF selection for bone diseases.

**Bone Disease Memory (BDM).** We term the bone disease memory (BDM) model as a variant of DM that is totally different from CDM model. The difference mainly lies in the input data during the training and testing stage. The ultimate goal of BDM is to monitor those RFs which cause people to get osteoporosis. Therefore, we have a crucial step for splitting data set as shown at the bottom block in Fig. ?? During the training stage, the top block of Fig. ?? shows a hierarchical latent DBN structure that is well trained by applying diseased RFs, as shown by dashed arrows.
in this figure. An interesting property of DBNs is the capability of reconstructing RFs [?]. Therefore, RFs reconstructed using BDM are reflections of the diseased individuals. We try to minimize the errors between both sides for a well-trained BDM. Based on the property of DBNs, if there is a large error between the original RF and the reconstructed one, this RF is likely to be a noisy RF and should not be further considered. After such a noisy RF selection process, we can measure the reconstructed error for each RF to find a possible informative RF in the testing stage. However, in the testing stage, we will use non-diseased RFs as the input, as shown by solid arrows in Fig. ?? This effort monitors the differences between the original RFs and the reconstructed RFs. Therefore, we intend to track a large error between both sides during the testing stage. The larger the error is, the more possible it is the informative RF. Under this scenario, we measure the reconstructed error for each RF for filtering the noisy RF in the training stage and finding the possible informative RF in the testing stage. We rank the total reconstructed error to select the top-N informative RFs by following distance metrics:

Reconstructed Error in Training Stage: $d^{(k)}_{\text{train}} = \sqrt{\frac{\sum_{i=1}^{n}(RRF^{(k)}_i - ORF^{(k)}_i)^2}{n}}$, where we use Root Mean Square Error (RMSE) to calculate the $k$th RF distance between the reconstructed RF $RRF^{(k)}_i$ and the original RF $ORF^{(k)}_i$ for $n$ training samples, and $n$ is the total number of training samples.

Reconstructed Error in Testing Stage: $d^{(k)}_{\text{test}} = \sqrt{\frac{\sum_{j=1}^{m}(RRF^{(k)}_j - ORF^{(k)}_j)^2}{m}}$, where we use RMSE to calculate the $k$th RF distance between the reconstructed RF $RRF^{(k)}_j$ and the original RF $ORF^{(k)}_j$ for $m$ test samples, and $m$ is the total number of testing samples. For a new incoming sample, we still use above formula but change $m$ to 1.

Total Reconstructed Error: $d^{(k)}_{\text{total}} = |d^{(k)}_{\text{test}} - d^{(k)}_{\text{train}}|$, where $d^{(k)}_{\text{total}}$ represents the total error of both stages for the $k$th RF by calculating its absolute distance. Note that only the RFs with large reconstructed error in the testing stage and small error in the training stage (i.e. not the noisy RF) are regarded as the informative RFs. We rank $d^{(k)}_{\text{total}}$ by decreasing order and yield a top-N informative RF list by selecting the first $N$th terms.

Non-Disease Memory (NDM). Similarly, we term the non-disease memory (NDM) model as a model which is trained by the non-diseased individuals so as to focus on the characteristics of those patients who have healthy bone. The structure of NDM is similar to BDM. However, the input data for training and testing the NDM model are swapped. The training procedure for NDM is achieved by using all non-diseased RFs as input data, instead of diseased RFs in Fig. ?? During the testing stage, we replace non-diseased RFs in Fig. ?? with diseased RFs and aim at observing if an osteoporotic diseased individual will get back to normal. This procedure can be severed as a cross-validation to evaluate the informative RFs provided by BDM. Since only the informative RFs produce a large total reconstructed error if we successfully remove the unreliable data, the informative RFs predicted by either BDM or NDM should be consistent. We apply distance metrics in accordance with BDM when calculating the total reconstructed error.

4 Experiments

4.1 Data Set

The Study of Osteoporotic Fractures (SOF) is the largest and most comprehensive study of RFs for bone diseases which includes 9704 Caucasian women aged 65 years and older. It contains 20 years of prospective data about osteoporosis, bone fractures, breast cancer, and so on. Potential RFs and confounders were classified into 20 categories such as demographics, family history, lifestyle, and medical history [?]. As shown in Fig. ??, there are missing values for both RF space and label space, denoted as empty shapes.

A number of potential RFs are grouped and organized at the first and second visits which include 672 variables scattered into 20 categories as the input of our model. The rest of the visits contain time-series dual-energy x-ray absorptiometry (DXA) scan results on bone mineral density (BMD) variation, which will be extracted and processed as the label for our data set. Based on WHO standard, T-score of less than -1 indicates the osteopenia condition that is the precursor to osteoporosis, which is used as the first type of label. The second type of label is the annual rate of BMD variation. We use at least two BMD values in the data set to calculate the bone loss rate and define the high bone loss rate with greater than 0.84% bone loss in each year [?]. Notice that this is a partially labeled data set since some patients just come during the first and second visit and never take a DXA scan in the following visits like example Patient3 shown in Fig. ??.

4.2 Evaluation Metrics

The error rate on a test dataset is commonly used as the evaluation method of the classification performance. Nevertheless, for most skewed medical data sets, the error rate could be still low when misclassifying entire minority sample to the class of majority. Thus, two alternative measurements are used in this paper. First, Receiver Operator Characteristic (ROC) curves are plotted to generally capture how the number of correctly classified abnormal cases varies with the number of incorrectly classifying normal

2. T-score of -1 corresponds to BMD of 0.82, if the reference BMD is 0.942 and the reference standard deviation is 0.122.
cases as abnormal cases. Since in most medical problems, we usually care about the fraction of examples classified as abnormal cases that are truly abnormal, the measurements, Precision-Recall (PR) curves, are also plotted to show this property. We present the confusion matrix in Table ?? and several derivative quality measures in Table ??.

![Image](https://via.placeholder.com/150)

**Fig. 7:** Illustration of missing values for the SOF dataset shown in non-shaded shapes for both RF space and label space. Two types of label information $L_1$ and $L_2$ with binary values are shown.

**TABLE 1: Confusion matrix.**

<table>
<thead>
<tr>
<th>Predicted Class</th>
<th>Actual Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP</td>
</tr>
<tr>
<td>Negative</td>
<td>FN</td>
</tr>
</tbody>
</table>

**TABLE 2: Metrics definition.**

\[
\begin{align*}
\text{True Positive Rate} & = \frac{TP}{TP + FN} \\
\text{False Positive Rate} & = \frac{FP}{FP + TN} \\
\text{Precision} & = \frac{TP}{TP + FP} \\
\text{Recall} & = \frac{TP}{TP + FN} \\
\text{Error Rate} & = \frac{FP + FN}{TP + FN + FP + TN + FN + FP + FN}
\end{align*}
\]

### 4.3 Experiments and Results for Integrated Risk Features Extraction

#### 4.3.1 Experiment Setup

To show the excellent predictive power using integrated features extracted by our CDM model, we manually choose RFs based on the expert opinion [?], [?], [?] as the baseline approach shown in Table ???. For a fair comparison, we fix the number of the output dimensions to be equal to the expert selected RFs. Specifically, we fix the number of units in the output layer to be 11, where each unit in this layer represents a new integrated feature describing complex relationships among all 672 input factors, rather than a set of typical RFs selected by experts shown in Table ??.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Numeric</td>
<td>Between 65 - 84</td>
</tr>
<tr>
<td>Weight</td>
<td>Numeric</td>
<td>BMI = weight/height</td>
</tr>
<tr>
<td>Height</td>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>BMI</td>
<td>Numeric</td>
<td></td>
</tr>
<tr>
<td>Parent tall</td>
<td>Boolean</td>
<td>Hip fracture in the patient’s mother or father</td>
</tr>
<tr>
<td>Smoke</td>
<td>Boolean</td>
<td>Excess alcohol</td>
</tr>
<tr>
<td>Excess alcohol</td>
<td>Boolean</td>
<td>Daily</td>
</tr>
</tbody>
</table>
| Rheumatoid arthritis | Boolean | It is noteworthy that holding out portions of the dataset is a manner similar to cross-validation. In Algorithm 1, unsupervised pre-training of the CDM model employs both labeled and unlabeled training samples while the supervised fine-tuning phase is conducted by a 5-fold cross-validation on the labeled training examples. Specifically, we divide the whole data set into 5 parts, in which 3 parts are used to train the model, and the fourth part is applied as the holdout set for mitigating the impact of over-fitting, and the fifth part is used to run a classification test. In the next run, the parts used for training, holding out, and testing are changed. Thus, each run on testing sample outputs a vector in the range [0,1] indicating the belief score to a class, yielding 5 independent vectors in total after a 5-fold cross-validation. When plotting ROC and PR curves, 5 vectors are concatenated into a vector with their equal-sized label vectors. In this way, AUC score is indeed the averaged score over 5-fold cross-validation runs.

<table>
<thead>
<tr>
<th>TABLE 3: Typical risk factors from the expert opinion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Factors From:</td>
</tr>
<tr>
<td>Expert knowledge</td>
</tr>
<tr>
<td>Shallow RBM without FT</td>
</tr>
<tr>
<td>Shallow RBM with FT</td>
</tr>
<tr>
<td>DBN without FT</td>
</tr>
<tr>
<td>DBN with FT</td>
</tr>
</tbody>
</table>

#### 4.3.2 Performance study for osteoporosis prediction

The overall results for the SOF data after Phase1 are shown in Table ???. The area under curve (AUC) of
ROC curve for each classifier (denoted as “LR-ROC”, “SVM-ROC”) and the AUC of PR curve (denoted as “LR-PR”, “SVM-PR”) are shown in Table ???. AUC indicates the performance of a classifier: the larger the better (an AUC of 1.0 indicates a perfect performance). The classification results using expert knowledge are also shown for the performance comparison as the baseline. From Table ??, we observe that a “shallow RBM without FT” method gets a sense of how the data is distributed which represents the basic characteristics of the data itself. Although the performances are not always higher than the expert model, this is a completely unsupervised process without borrowing knowledge from any types of labeled information. Achieving such a comparable performance is not easy since the expert model is trained in a supervised way. Further improvements may be possible by more thorough experiments under the help of label for finishing a two-stage fine-tuning that is used to better satisfy our prediction tasks. Next we transform from an unsupervised task to a semi-supervised task. Table ?? also shows the classification results which boost the performance of all classifiers because of the two-stage fine-tuning shown as “shallow RBM with FT”. Especially, the AUC of PR of our model significantly outperforms the expert system. Since the capacity for the RBM model with one hidden layer is usually small, it indicates a need for a more expressive model over the complex data. To satisfy this need, we add a new layer of non-linear perceptron at the bottom of RBM, which forms a DBN as shown in Fig. ?? (b). This new added layer greatly enlarges the overall model expressiveness. More importantly, the deeper structure is able to extract more abstractive reasons. As we expected, unsupervised pre-training of a deeper structure yields a better performance than the shallow RBM model (denoted as “DBN without FT”), and the model further improves its behavior after the two-stage fine-tuning shown as “DBN with FT” in Table ??.

4.3.3 Performance study for bone loss rate prediction

In this section, we show the bone loss rate prediction using the abnormal cases after Phase1. High bone loss rate is an important predictor of higher fracture risk. Moreover, it’s reported that RFs that account for high and low bone loss rates are different [3]. Our integrated risk features are good at detecting this property since they integrate the characteristics of data itself and nicely tuned under the help of two kinds of labels. We compare the results between expert knowledge based model and our DBN with fine-tuning model that yields the best performance for Phase1. The classification error rate is defined in Table ??.

Since our result is also fine-tuned by the bone loss rate, we can directly feed new integrated features into Phase2. Table ?? shows that our model outperforms the expert model when predicting bone loss rate. In this case, the expert model fails because the limited features are not sufficient to forecast the bone loss rate which may interact with other different RFs. This highlights the need for a more complex model to extract the precise attributes from amounts of potential RFs. Moreover, our CDM module takes into account the whole data set, not only keeping the 672 risk factors but also utilizing two types of label. The integrated risk features reserve the characteristics of bone loss rate after the second round fine-tuning, which assist in bone loss rate prediction.

4.4 Experiments and Results for Informative Risk Factor Selection

In this section, we will show experiments and results on informative RF selection. Based on the proposed method shown in Fig. ??, we show a case study which lists the top 20 informative RFs selected using BDM and NDM in Table ???. Description for each variable can be found from the data provider website [3].

In this study osteoporosis appears to be associated with several known RFs that are well described in the literature. Based on the universal rule used by FRAX [2] that is a popular fracture risk assessment tool developed by WHO, some of the selected RFs have already been used to evaluate fracture risk of patients such as age, fracture history, family history, BMD and extra alcohol intake. Besides, most informative RFs we reported in Table ?? have been reviewed and endorsed by bone health institutions and medical researchers. For example, some physical and lifestyle risk factors such as dizziness(DIZZY), vital status(CSHAVG), inability to rise from a chair (STDARM), daily exercises(50TMWT), are examined as important risk factors for osteoporotic fractures [3], [2], [2]. Blood pressure is a secondary risk factor in that blood pressure pills may increase the risk [2]. Breast cancer has been examined as a risk factor by National Institutes of Health (NIH) which says that women with breast cancer are at increased risk for developing osteoporosis due to a drop in estrogen levels, chemotherapy and the production of osteoclasts [2]. Of greatest interest is that some physical performances such as steadiness/step in turn (TURNUM, STEADY, STEPUP), aid used for pace tests(GAID) are perhaps the identifiable informative risk factors and can be easily incorporated into routine clinical practice. Based on these results, some environmental/behavioral risk factors are modifiable and preventions and therapeutic interventions are needed to reduce osteoporosis and fracture risks.

| Table 5: Classification error rates of expert knowledge model and our model |
|--------------------------|------------------|
| Expert                   | LR-Error | SVM-Error |
| DBN with FT              | 0.107    | 0.094     |

 achievement of a comparable performance is not easy since the expert model is trained in a supervised way. Further improvements may be possible by more thorough experiments under the help of label for finishing a two-stage fine-tuning that is used to better satisfy our prediction tasks. Next we transform from an unsupervised task to a semi-supervised task. Table ?? also shows the classification results which boost the performance of all classifiers because of the two-stage fine-tuning shown as “shallow RBM with FT”. Especially, the AUC of PR of our model significantly outperforms the expert system. Since the capacity for the RBM model with one hidden layer is usually small, it indicates a need for a more expressive model over the complex data. To satisfy this need, we add a new layer of non-linear perceptron at the bottom of RBM, which forms a DBN as shown in Fig. ?? (b). This new added layer greatly enlarges the overall model expressiveness. More importantly, the deeper structure is able to extract more abstractive reasons. As we expected, unsupervised pre-training of a deeper structure yields a better performance than the shallow RBM model (denoted as “DBN without FT”), and the model further improves its behavior after the two-stage fine-tuning shown as “DBN with FT” in Table ??.

4.3.3 Performance study for bone loss rate prediction

In this section, we show the bone loss rate prediction using the abnormal cases after Phase1. High bone loss rate is an important predictor of higher fracture risk. Moreover, it’s reported that RFs that account for high and low bone loss rates are different [3]. Our integrated risk features are good at detecting this property since they integrate the characteristics of data itself and nicely tuned under the help of two kinds of labels. We compare the results between expert knowledge based model and our DBN with fine-tuning model that yields the best performance for Phase1. The classification error rate is defined in Table ??.

Since our result is also fine-tuned by the bone loss rate, we can directly feed new integrated features into Phase2. Table ?? shows that our model outperforms the expert model when predicting bone loss rate. In this case, the expert model fails because the limited features are not sufficient to forecast the bone loss rate which may interact with other different RFs. This highlights the need for a more complex model to extract the precise attributes from amounts of potential RFs. Moreover, our CDM module takes into account the whole data set, not only keeping the 672 risk factors but also utilizing two types of label. The integrated risk features reserve the characteristics of bone loss rate after the second round fine-tuning, which assist in bone loss rate prediction.

4.4 Experiments and Results for Informative Risk Factor Selection

In this section, we will show experiments and results on informative RF selection. Based on the proposed method shown in Fig. ??, we show a case study which lists the top 20 informative RFs selected using BDM and NDM in Table ???. Description for each variable can be found from the data provider website [3].

In this study osteoporosis appears to be associated with several known RFs that are well described in the literature. Based on the universal rule used by FRAX [2] that is a popular fracture risk assessment tool developed by WHO, some of the selected RFs have already been used to evaluate fracture risk of patients such as age, fracture history, family history, BMD and extra alcohol intake. Besides, most informative RFs we reported in Table ?? have been reviewed and endorsed by bone health institutions and medical researchers. For example, some physical and lifestyle risk factors such as dizziness(DIZZY), vital status(CSHAVG), inability to rise from a chair (STDARM), daily exercises(50TMWT), are examined as important risk factors for osteoporotic fractures [3], [2], [2]. Blood pressure is a secondary risk factor in that blood pressure pills may increase the risk [2]. Breast cancer has been examined as a risk factor by National Institutes of Health (NIH) which says that women with breast cancer are at increased risk for developing osteoporosis due to a drop in estrogen levels, chemotherapy and the production of osteoclasts [2]. Of greatest interest is that some physical performances such as steadiness/step in turn (TURNUM, STEADY, STEPUP), aid used for pace tests(GAID) are perhaps the identifiable informative risk factors and can be easily incorporated into routine clinical practice. Based on these results, some environmental/behavioral risk factors are modifiable and preventions and therapeutic interventions are needed to reduce osteoporosis and fracture risks.
TABLE 6: Informative risk factors generated by BDM and NDM

<table>
<thead>
<tr>
<th>Category</th>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Age</td>
<td>The patient’s age at this visit</td>
</tr>
<tr>
<td>Fracture history</td>
<td>IFX14</td>
<td>Vertebral fractures</td>
</tr>
<tr>
<td></td>
<td>INTX</td>
<td>Intertrochanteric fractures</td>
</tr>
<tr>
<td></td>
<td>FACEF</td>
<td>Face fracture</td>
</tr>
<tr>
<td></td>
<td>ANYF</td>
<td>Follow-up time to 1st any fracture since current visit</td>
</tr>
<tr>
<td>Family history</td>
<td>MHIP80</td>
<td>Mom hip fracture after age 80</td>
</tr>
<tr>
<td>Exam</td>
<td>DSTBMC</td>
<td>Distal radius bone mass content ($gm/cm^2$)</td>
</tr>
<tr>
<td></td>
<td>PRXBM</td>
<td>Proximal radius bone mass density ($gm/cm^2$)</td>
</tr>
<tr>
<td>Physical</td>
<td>TURNUM</td>
<td>Number of steps in turn</td>
</tr>
<tr>
<td>performance</td>
<td>STEADY</td>
<td>Steadiness of turn</td>
</tr>
<tr>
<td></td>
<td>STEPPUP</td>
<td>Ability to step up one step</td>
</tr>
<tr>
<td></td>
<td>STDMRM</td>
<td>Does participant use arms to stand up?</td>
</tr>
<tr>
<td></td>
<td>GAID</td>
<td>Aid used for pace tests (i.e. crutch, cane, walker)</td>
</tr>
<tr>
<td>Exercise</td>
<td>50TMWT</td>
<td>Total number of times of activity/year at age 50</td>
</tr>
<tr>
<td>Life style</td>
<td>DR30</td>
<td>During the past 30 days, how often did you have 5 or more drinks during one day</td>
</tr>
<tr>
<td>Breast cancer</td>
<td>BRSTCA</td>
<td>Breast cancer status such as tumor behavior, staging of tumor and so on</td>
</tr>
<tr>
<td>Blood pressure</td>
<td>LISTSYS</td>
<td>Systolic blood pressure lying down (mmHg)</td>
</tr>
<tr>
<td></td>
<td>DIZZY</td>
<td>Dizziness upon standing up</td>
</tr>
<tr>
<td>Vision</td>
<td>CSHAVG</td>
<td>Average contrast sensitivity</td>
</tr>
</tbody>
</table>

On the other hand, learner may need to purchase data during the training stage and recruit people to answer hundreds and thousands questions. With a fixed budget and limited time, it might be impossible to acquire as many as possible features for all participants. So what are the most important questions the physicians need to know? How many features they need to achieve a good predictive performance. Using the proposed approach we select at most top 50 informative RFs, instead of using all of them, and feed them directly to the logistic regression classifier for the osteoporosis prediction. Fig. ?? shows the osteoporosis prediction AUC result for both ROC and PR curve as the number of informative RFs increases. As we can see, the proposed informative RF selection method exhibits great power of predicting osteoporosis in that the selected RFs are more significant than the rest of the RFs. Moreover, the best prediction performance is achieved using the proposed method when selecting the top 20 to top 25 informative RFs. And the AUC is even better than the expert knowledge model (AUC of ROC: 0.729; AUC of PR:0.458). The performance of the prediction result of top-N RFs selected by BDM and NDM is inferior to that of integrated RFs extracted by CDM (AUC of ROC: 0.878; AUC of PR:0.718) in that some information are discarded and those information might still make contribution to enhancing the predictive behaviors.

5 SENSITIVITY ANALYSIS AND PARAMETER SELECTION

5.1 Sensitivity to Skewed Class

We provide the data set statistics in Fig. ??, We collected patients’ BMD in baseline visit and next 10-years visit. The whole dataset can be split to five parts: normal BMD to normal BMD, normal BMD to osteoporotic BMD, osteoporotic BMD to normal BMD, osteoporotic BMD to osteoporotic BMD and missing BMD. We used 8074 patients with label for osteoporosis prediction in Section ??, in which 4349 patients (NormalToNormal and OsteoToNormal) have normal BMD and 3725 patients (NormalToOsteo and OsteoToOsteo) have osteoporotic BMD in 10-years later. Although class distribution is not far from the uniformly distribution, it is still necessary to discuss how well our model overcomes the skewed class problem. We conduct experiments on the partial data set that is highly imbalanced to examine our model since the imbalanced class problem may seriously degrade a classifier’s performance on the minority class. We manually remove OsteoToOsteo data, shown as the fourth bar in Fig. ?? As a result, the data set has a ratio of roughly 6:1 for two classes (4349 normal patients and 689 osteoporotic patients). Algorithm 1 includes two steps: pre-training and fine-tuning. During the pre-training phrase, we do not rely on any
label information, that is, an un-supervised learning fashion. If the training set is imbalanced, this pre-training will likely initialize the structure of model close to the majority class. The common solutions for balancing the data are under-sampling (ignoring data from the majority class), over-sampling (replicating data from the minority class) and informed under-sampling (selecting data according to some set of principles) [2], [7], [8]. One straightforward way is to independently sample several subsets from the majority class, with each subset having approximately the same number of examples as the minority class. In this way, we can create six roughly balanced datasets by replicating the minority class and partitioning the majority class into six subsets. Then, we can independently train six classifiers and count votes for the final decision shown in Fig. ??.

To test the effect of the class distributions on our proposed model, we compare the performances based on our proposed model shown in Fig. ?? with a DBN structure appended a logistic regression model on it. We still use AUC of ROC and PR curves as performance evaluation measures shown in Fig. ??.

Experimental results show that class imbalance is harmful for osteoporosis prediction, and dealing with this problem improves the AUC scores of both ROC and PR curve. We also observed that on models which add fine-tuning, it is especially problematic since imbalanced label information leads our model to a local minimum that ignores the minority cases.

### 5.2 Sensitivity to Noisy Data

Fig. ?? shows that there are missing values or noisy values in the risk factor space of the data set. This is a common problem in most clinical datasets. To handle with missing/noisy data, we follow up two steps: (1) manually removing those risk factor columns with more than 70% missing values during the data pre-processing procedure (249 of 672 risk factor columns with more than 70% missing values are deleted) and (2) using a column-wise mean to fill out the blank for the surviving columns. Then we rely on the good de-noising properties coming from the basic structure “RBM” of our model. The de-noising capability of RBM model has been widely examined by some computer vision tasks [9], [10]. Contrastive Divergence training is actually a stochastic sampling process, which randomly turns on the hidden unit based on the activation probability. This randomness cancels out the data noise in a certain level. Moreover, the data distribution is more consistent across all training samples, but the noise distribution differs for each sample. When we feed the model with enough samples, the sampling process will drive the model toward the data distribution because this distribution occurs more frequently. We may need to apply a more sophisticated method such as matrix factorization based collaborative filtering so as to maximize the use of the original data in future studies.

### 5.3 Parameter Selection

The number of hidden units is closely related to the representational power of our model. Ideally, we can represent any discrete distribution exactly when the number of hidden units is very large [11]. In our experiment, we examine the power of our model when increasing the number of binary units on the first hidden layer. Fig. ?? shows the performance of our CDM model under different number of hidden units. When the number of hidden units is small, the model is lack of capacity to capture data complexity which results in lower performance on both AUC of ROC and PR curves. As we increase the number of hidden units, our model shows a strictly improved modeling power. However, when the number is too large, we don’t have sufficient samples to train the network which results in a lower performance and stability. In our experiment, we choose 400 as the hidden unit number.
Despite the model parameters changing between updates, these changes should be small enough that only a few steps of Gibbs (in practice, often one step is used) are required to maintain samples from the equilibrium distribution of the Gibbs chain, i.e., the model distribution. The learning rate used to update weights is fixed to the value of 0.05 that is chosen from the validation set. And the number of iterations is set to 10 for efficiency since we observed that the model cost can reach into a relatively stable state within 5 to 10 iterations. We use mini-batch gradient for updating the parameters and the batch size is set to 20. After the model is trained, we simply feed it with the whole data and get the new integrated RFs and then run the same classification module to get results.

6 CONCLUSIONS

We developed a multi-tasking framework for osteoporosis that not only extracts the integrated features for progressive bone loss and bone fracture prediction but also selects the individual informative RFs that are valuable for both patients and medical researchers. Our framework finds a representation of RFs so that the salient integrated features can be disentangled from ill-organized EHR data. These integrated features constructed from original RFs will become the most effective features for bone disease prediction. We developed disease memory (DM), which categorizes and stores the underlying characteristics for a specific cohort. In essence, we trained an independent model based on a specific group of patients. For example, comprehensive disease memory (CDM) captures the characteristics for all patients to predict the disease. Bone disease memory (BDM) memorizes the characteristics of those individuals who suffer from bone diseases. Similarly, the non-disease memory (NDM) memorizes attributes for non-diseased individuals. A variety of DM models increase the flexibility for monitoring the disease for different groups of patients. Our extensive experimental results showed that the proposed method improves the prediction performance and has great potential to select the informative RFs for bone diseases. As a long-term impact, a bone disease analytic system will ultimately be deployed in bone disease monitoring and preventing settings which will offer much greater flexibility in tailoring the scheduling, intensity, duration and cost of the treatment.

Xiaoyi Li

Xiaoyi Li received the Master degree at Robin Li Data Mining & Machine Learning Laboratory with supervision by Prof. Aidong Zhang. His current work focuses on learning representation from data with multiple views - how to maximize the fusion of different views to better represent an object.

Murali Ramanathan

Murali Ramanathan is Professor of Pharmaceutical Sciences and Neurology at the State University of New York at Buffalo, New York, US in 2011. He is currently pursuing a PhD degree at Robin Li Data Mining & Machine Learning Laboratory lead by Prof. Aidong Zhang. His primary interests lie in data mining, machine learning and deep learning. His current work focuses on learning representation from data with multiple views – how to maximize the fusion of different views to better represent an object.

Aidong Zhang

Aidong Zhang is UB Distinguished Professor and Chair in the Department of Computer Science and Engineering at State University of New York at Buffalo. Her research interests include bioinformatics, data mining, multimedia and database systems, and content-based image retrieval. She is an author of over 250 research publications in these areas. She has chaired or served on over 100 program committees of international conferences and workshops, and currently serves several journal editorial boards. She has published two books Protein Interaction Networks: Computational Analysis (Cambridge University Press, 2009) and Advanced Analysis of Gene Expression Microarray Data (World Scientific Publishing Co., Inc. 2006). Dr. Zhang is a recipient of the National Science Foundation CAREER award and State University of New York (SUNY) Chancellor’s Research Recognition award. Dr. Zhang is an IEEE Fellow.