5-1-2019

Machine Learning Approach for Prediction of Bone Mineral Density and Fragility Fracture in Osteoporosis

Bibek Bhattarai
bibekbhattarai.brt@gmail.com

MACHINE LEARNING APPROACH FOR PREDICTION OF BONE MINERAL DENSITY AND FRAGILITY FRACTURE IN OSTEOPOROSIS

By

Bibek Bhattarai

Bachelor in Computer Engineering (B.E.)
Tribhuvan University, Kathmandu, Nepal
2012

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science in Computer Science

Department of Computer Science
Howard R. Hughes College of Engineering
The Graduate College

University of Nevada, Las Vegas
May 2019

UNLV | GRADUATE COLLEGE

Thesis Approval

The Graduate College
The University of Nevada, Las Vegas

April 17, 2019

This thesis prepared by

Bibek Bhattarai

entitled

Machine Learning Approach for Prediction of Bone Mineral Density and Fragility Fracture in Osteoporosis

is approved in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science
Department of Computer Science

Fatma Nasoz, Ph.D.
Examination Committee Chair

Kathryn Hausbeck Korgan, Ph.D.
Graduate College Dean

Laxmi Gewali, Ph.D.
Examination Committee Member

Justin Zhan, Ph.D.
Examination Committee Member

Qing Wu, Ph.D.
Graduate College Faculty Representative

Abstract

Osteoporosis is a prevailing bone disease, which weakens the bone and is one of the major factors of disability, especially in elderly persons. In this thesis, we developed various machine learning models to predict fracture risk of osteoporosis. These models were built to base their predictions on genotype and phenotype data of patients. We performed two different types of analysis: fracture risk prediction (a classification model) and bone mineral density (BMD) prediction (a regression model). For fracture risk prediction we implemented four different algorithms: logistic regression, random forest, gradient boosting, and multi-layer perceptron (MLP) based on different risk factors identified. We performed our experiments using 307 and 1103 Single Nucleotide Polymorphism (SNPs) with data from 5133 patients. For both 307 and 1103 SNPs the performance of MLP was the best with area under curve (AUC) of 0.970 and 0.981 respectively. Logistic regression had the worst performance among four models with AUC of 0.816 and 0.904. For BMD prediction we implemented linear regression, random forest, gradient boosting and MLP and as a performance metric we plotted mean squared error (MSE) versus number of iterations for both train and test set of data. The random forest performed the best in both cases with MSE of 0.004 and linear regression was the worst with MSE of 0.104 in the test data for both sets of SNPs.

Acknowledgements

"First and foremost, I would like to express my sincere gratitude to my advisor Dr. Fatma Nasoz for her dedicated guidance, continuous support and motivation during this work. I would like to thank her for providing me valuable suggestions and motivating me.

I would also like to thank Dr. Qing Wu for providing me the dataset for this thesis and providing me valuable suggestions during my work. I would also like to thank Dr. Laxmi Gewali and Dr. Justin Zhan for their continuous support, reviewing my work, and being part for my advisory committee. Furthermore, I express my sincere gratitude to Dr. Ajoy Datta for his continuous feedback and support during my Master's program.

I want to extend my gratitude to my family: my mother and my brother. I would also like to thank my dear friends Pradip Singh Maharjan, Ashish Tamrakar, Shristy Maharjan, Shuveksha Tuladhar, Sailuj Shakya, Abhusan Acchami, and Sameeksha Sapkota for their unconditional support and encouragement through these years.

I would also like to acknowledge Xiangxue Xiao for helping me throughout my thesis work. Lastly, I want to thank University of Nevada, Las Vegas (UNLV) for providing me an opportunity to pursue my Master's program in this wonderful environment."

Bibek Bhattarai

University of Nevada, Las Vegas

May 2019

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Osteoporosis is the prevailing bone disease in which the density and quality of bones are reduced literally leading to abnormality called a porous bone, which is compressible, like a sponge. It is generally characterized by low bone mineral density mass and micro-architectural deterioration of bone tissue [IAA14]. This disease develops without showing symptoms in its early stages. Osteoporosis weakens the bone and results in recent fractures in the bones. It is becoming a real public health problem because of its increasing frequencies over different countries [CDA+17]. Low Bone Mineral Density (BMD) has been considered as the strong risk factor for osteoporosis, and thus has been considered as key factor or indicator for its treatment and diagnosis. Genome wide association studies (GWAS) have identified BMD is highly heritable.

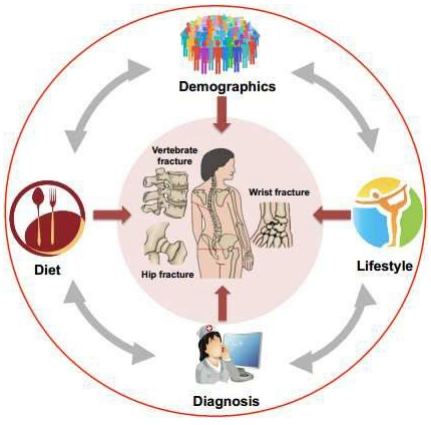

Figure 1.1: Risk Factors for Osteoporosis [CDA+17]

Osteoporosis prevention is complicated and in recent years the social burden of this disease has become large. Thus prevention and treatment of osteoporosis have become an urgent issue to be addressed, so modeling the relationships between the disease and its risk factors (potential ones) is

an important and crucial task. There are several potential risk factors associated with osteoporosis as shown in figure 1.1. But the potential risk factors are not limited to demographic attributes, family history, diet, and lifestyle [CDA+17]. Bone Mineral Density, which is one of the prime factors for bone fractures is heritable so different genetic features too contribute as risk factors for osteoporosis as well.

1.1 Genome Wide Association Study

In genetics, a genome wide association study (GWAS) is an observational study of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait. Typically, GWASs are hypothesis free methods for identification of associations between loci (genetic regions) and traits (including diseases) [EBI]. We know that genetic variation can cause differences in phenotypes between individuals. These variants and those tightly related to their region of the chromosome are thus present at a higher frequency in individuals with the trait (cases) than individual without traits (controls).

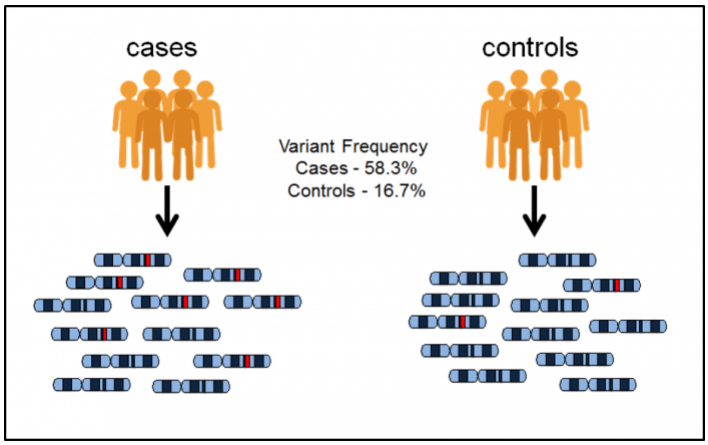

Figure 1.2: Typical allele distribution. [EBI]

GWASs typically focus on associations between single-nucleotide polymorphisms (SNPs) and traits like major human diseases. The variants associated with the disease can be found at a higher frequency in cases than controls.

1.2 Single Nucleotide Polymorphisms

Single Nucleotide polymorphisms (SNPs) are considered as the most common type of genetic variation among different individuals. Each SNP means a difference in a single nucleotide (building

block of DNA). For example, an SNP may replace nucleotide guanine (G) and nucleotide adenine (A) in a certain stretch of DNA. In on average, SNPs occur once in every 1,000 nucleotides. The variations occurred may be unique or may occur in many individuals and these variations are found in the DNA between genes. Most SNPs do not have an effect on health, but some studies have found that SNPs may assist to predict the risk associated with certain diseases [GHR]. So the SNPs may play a direct role in the disease that have been affecting the gene's function. In this thesis, we included different indentified SNPs that have significant association with fracture risk in osteoporosis.

1.3 Objective

The main objective of this thesis is to perform predictive analysis on the genotypes dataset using various machine learning algorithms. The main focus is to identify if risk of fracture exists or not and to predict the bone mineral density value for the available genotype dataset.

1.4 Outline

In chapter 1, we provided an introduction to osteoporosis and genome-wide association study. We also introduced to the risk factors associated with the disease. In chapter 2, we will be focusing on the related works previously conducted for the identification of different SNPs that are associated with high risk for fracture. Also in chapter 2 we will provide some background on the algorithms and terms associated with machine learning.

Chapter 3 will be focused on data descriptions and imputation of the dataset. In chapter 4, we will be presenting the experimental results obtained with different models.

Lastly, in chapter 5 we will summarize our results and offer an insight about future works that can prove to be more helpful in solving this problem in proximate future..

Chapter 2

Background and Preliminaries

2.1 Related Works

Bone Mineral Density (BMD) has been a widely used varaible for predicting fracture risk in Osteoporosis. And, recent studies show that BMD is heritable, and GWAS have identified common variants at different loci associated with the trait, including those that are significantly associated with fracture risk. In meta-analysis of lumbar spine and femoral neck BMD, it was identified that 63 SNPs were related highly for fracture risk, after all possible pairwise interactions of the 82 SNPs [ESE+12]. 307 conditionally independent SNPs that attained genome-wide significance at 203 loci were identified in a GWAS conducted within 142,487 individuals from UK. This research included 153 previously unreported loci [KMG+17]. In the recent publication in April 2018, 518 genome-wide significant loci (301 new) were identified, explaining 20% of its variance [MKY+19]. A recent article published by Kim shows the identification of 613 new loci associated with heel BMD for osteoporosis and fracture [Kim18]. The research conducted using data from UK Biobank identifies 1362 independent SNPs which are clustered into 899 loci.

A supervised Machine learning approach was used to identify the risk of osteoporosis using two algorithms: Naive Bayes' (NB) and Multi-layer Perceptron (MLP) [CDA+17]. 20 risk factors were identified based on information collected from 45 patients in Nigerian hospitals. The MLP accuracy was at 100% where NB method achieved the accuracy of just 71.4%. Hsueh-Wei et al. [CCK+13] used wrapper-based feature selection method was used along with three classification algorithms: multilayer feedforward neural network (MFNN), NB, and logistic regression. The performance of the MFNN model with wrapper-based approach was the best predictive model classifying osteoporosis outcome.

In an experiment that was conducted by Forgetta et al. [FKBMF+18], which used 341,449 individuals from UK biobank with speed of sound (SOS): a risk factor for osteoporosis fracture. The experiment was conducted to develop gnomically-predicted SOS (gSOS) by using various machine learning algorithms. Genotypes data was used, which resulted in a relevant prediction of SOS and fracture. This article focuses on analyzing the osteoporosis fracture with SOS which explained 4.8-fold more variance in SOS than FRAX (fracture risk assessment tool) clinical factors.

Tae et al. [YKK+13] conducted a research on osteoporotic data of 1674 Korean postmenopausal women osteoporotic data with low BMD at any site among total hip, femoral neck, or lumbar spine measurements. Among different algorithms implemented support vector machines (SVM) had higher area under curve (AUC) of the receiver operating characteristic (ROC). SVM, artificial neural network (ANN), and logistic regression (LR) were three algorithms implemented for creating the models.

2.2 Preliminaries

Before going into applications of machine learning algorithms, this chapter helps the reader to understand the concepts of machine learning and different implementation of machine learning algorithms. Mostly, we will be focusing on the supervised learning: task of inferring a function from labeled training data.

2.2.1 Machine Learning

Machine learning is the field of artificial intelligence and can can be defined as programming computers to optimize a performance criterion using some example data or past experience [Eth10]. Moreover, it is the study of algorithms and statistical models that computers use to perform a specific task without being explicitly programmed. Machine learning algorithms build a mathematical model of some sample data, also known as "training data", which is then used to predict or make decisions for some other data also known as "test" data. Several machine learning algorithms are widely used in various real life applications like spam filtering, stock prediction, image processing, anomaly detection, stock market prediction, fraud detection, medical diagnosis and many more.

Usually, machine learning is divided into two types: supervised and unsupervised. However, reinforcement learning, ensemble learning and neural network also have been considered as types of machine learning approaches [Dey16]. Supervised learning deals with mapping an input to output

labels or input to continuous output. Whereas, in unsupervised learning we wish to discover the new pattern or learn the inherent structure of the data without explicitly provided labels. Reinforcement learning is about attaining a complex objective or maximizing along a particular dimension over many steps. Algorithms in reinforcement learning can be expected to perform better in more complex, real-life environments. Ensemble learning is a learning paradigm where multiple learners are trained to solve the same problem collectively. This thesis deals mostly with supervised learning algorithms for constructing different models which we will discuss in next section. However, we've used "boosting" – an ensemble learning approach for predicting output and "backpropagation" – a neural network approach.

2.2.2 Supervised Learning

Supervised learning is an approach where we infer a function from labeled training data. The training data consists of a set of training examples where each example is a pair consisting of input features and a desired output value. In supervised learning the goal is to map the inputs x to output y, given a labeled set of training data

$$D = (x_i, y_i)_{i=1}^{N}$$

where N is the number of training examples. Depending upon the form of response or output variable the supervised learning can be further categorized into two: classification and regression. We deal with both a classification and a regression problem in this thesis which are discussed in sections 2.2.3 and 2.2.4.

2.2.3 Classification

Classification, a type of supervised learning, is the task of approximating a mapping function (f) from input variables (x) to discrete output variables (y), often called as labels or categories. In this thesis, predicting whether there is a fracture or not is the example of a classification problem. Here the problem is binary classification problems as there are only two classes (yes or no) for predicting fracture. Furthermore, classification can be multi-class (where the output labels are more than two output labels) or multi-label (where each sample set is assigned to target labels) [TOWb]. In present thesis we deal with binary classification where we predict whether the patient has risk of bone fracture or not.

2.2.4 Regression

Regression is similar to classification except the response or output variable is continuous. So, it is the task of approximating a mapping function (f) from input variables (x) to a continuous output variable (y). Since regression predicts a quantity, the performance of the model must be reported as an error in those predictions. In this thesis, we have used regression techniques for prediction the BMD value using different input features.

2.2.5 Ensemble Learning

The main principle behind ensemble learning is grouping weak learners to form a strong learner so that accuracy can be increased. For example, ensemble learning may combine several decision trees classifiers to produce better predictive performance than a single decision tree classifier. Ensemble learning helps to reduce the factors variance and bias, which cause the main differences in actual and predicted values. Bagging and boosting are the techniques that are used to decrease variance and increase robustness of the model. In this thesis we have both boosting and bagging concepts which are discussed next.

Boosting

In general, boosting is an ensemble approach for reducing bias, and variance in supervised learning. The idea of boosting is to train weak learners sequentially, each trying to correct its predecessor by adjusting weights to the samples that were previously misclassified [MED].

Bagging

Bootstrap aggregating, also called bagging, is another strong model for ensemble learning approach designed to improve the stability and accuracy of machine learning algorithms used in classification and regression problems. This method reduces variance and helps to avoid overfitting. Bagging is based on the bootstrap algorithm which draws random sample from given dataset with replacement. This method helps us to understand the mean and standard deviation from the dataset in a better view. An example for bootstrapping is shown in figure 2.1.

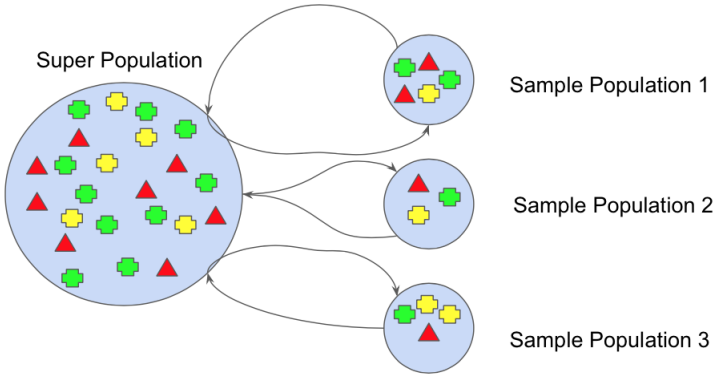

Figure 2.1: Bootstrapping from main population to sample population [BEC]

2.2.6 Neural Networks

Artificial neural network are the computing systems inspired by biological neural networks that constitute animal brains. Actually, neural networks is not an algorithm, but rather a framework for many machine learning algorithms to work together and process data inputs. These systems perform tasks by learning examples without being programmed with any specific rules. Neural Network is constructed from 3 types of layers: input (initial data for NN), hidden (intermediate layer between input and output layers where all the calculations are done) and output (result for the given inputs) layers [TOWa]. Figure 2.2 shows the basic architecture for neural networks. In this thesis, we used back-propagation algorithm for our data analysis and predicting results.

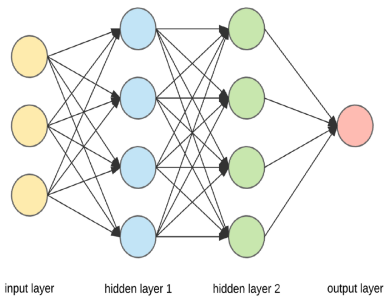

Figure 2.2: Neural Network Architecture[TOWa]

2.2.7 Model Selection

In this section we will discuss the different learning approaches we used for predictive analysis of the dataset. Depending upon the analysis we choose five predictive models: linear regression, logistic regression, random forest, gradient boosting, and backpropagation algorithms which are discussed in following section.

2.2.8 Linear Regression

Before going into linear regression, let us get familiar with regression. In statistical modeling, regression is a method for estimating the relationships among variables. It is a method of modeling a target value based on different independent predictors. Based on the number of independent variables and type of relationship between both dependent and independent variables, regression techniques differ mostly. Linear Regression – a type of regression analysis – is one of the most well known and understood algorithms in statistics and machine learning because the representation is so simple. Linear regression is used for continuous dependent variable. The representation in linear regression for a specific set of input values "x" and the predicted output "y" (continuous varaible) for that set of input values would be:

$$y = b_0 + b_1 x \tag{2.1}$$

where b_0 and b_1 are the parameters to estimate. Here, b_0 is also called the bias term and b_1 is the weight for input variable x. Figure 2.3 shows an example of linear regression.

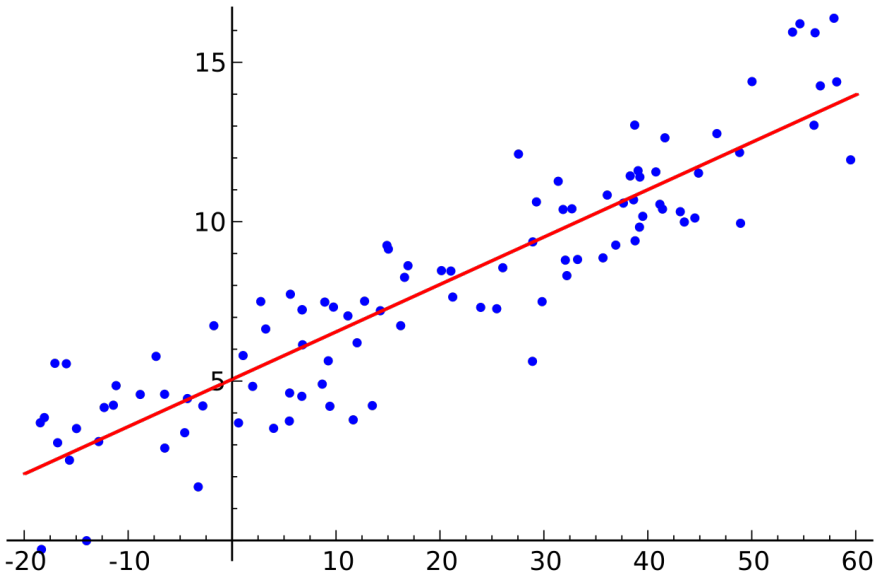

Figure 2.3: Linear regression [WIL]

2.2.9 Logistic Regression

Like other regression analyses, logistic regression is also a predictive analysis. This is also one of the most popular algorithms used for classification problems. Logistic regression is used when the target variable (dependent) variable has only two values, say 0 and 1 or Yes or No. Multinomial logistic regression is usually used for the case when dependent variables has three or more cases. Unlike linear regression that outputs continuous values, logistic regression uses sigmoid function to return a probability value which can be then mapped into number of discrete classes. The sigmoid function can be given by :

$$S(z) = \frac{1}{1 + e^{-z}} \tag{2.2}$$

where S(z) is the probability estimate (output between 0 and 1), z is input to the function (in the form $b_0 + b_1x$) and e base of natural log. This prediction function returns a probability value between 0 and 1. In order to map this to a class, we select a threshold value from which we will classify values to class 0 or class 1. The plot for the sigmoid function is shown in Figure 2.4. The decision boundary is given as:

$$P \geq 0.5, class = 1 \tag{2.3}$$

$$P < 0.5, class = 0 \tag{2.4}$$

Here the probability function P is defined as :

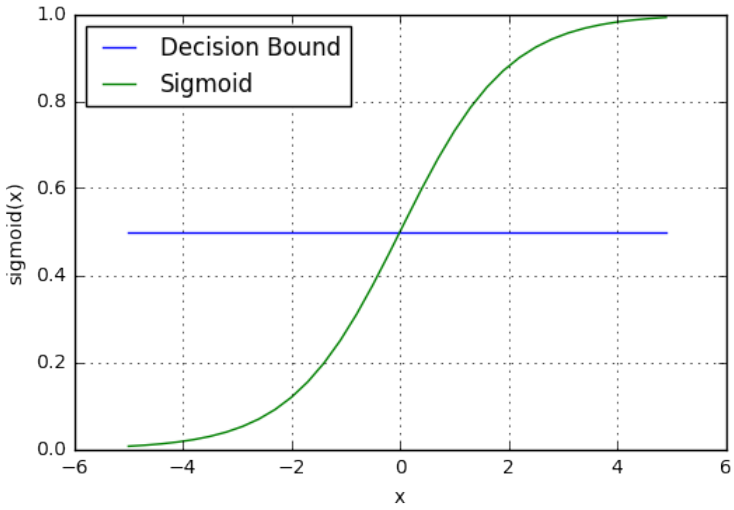

Figure 2.4: Sigmoid function Graphv [MLR]

$$P(class = 1) = S(b_0 + b_1 x) \tag{2.5}$$

Here b_0 and b_1 are the logistic regression parameters to estimate and are thus learned during the training process.

2.2.10 Random Forest

Random forest is an ensemble learning approach for classification and regression problems. It is one of the most used algorithms, because it can be used for both classification and regression problem as well as of its simplicity. Like from its name, it creates a forest and makes it random by training on different samples of data. It implements "bagging" where it builds different decision trees in ensemble. The general idea of bagging method is to combine different learning models (trees) so as to increase the overall results and performance [TOWc]. Decision trees are the foundation of random forest algorithm so before going into random forest let us get familiarize with decision tree concepts.

Decision tree is one of the most widely used methods for inductive inference over supervised data. It represents a procedure that classifies the categorical data [RJA+17]. A basic representation of decision tree can be seen in Figure 2.5 where it classifies whether weather is suitable to play tennis or not with decision "yes" or "no". The example starts with an outlook with three choices: sunny, overcast and, rain. If it is sunny we check if the humidity is high or normal. If it is high we make decision "no" for playing else "yes" for playing. If the outlook is overcast then we decide to play and if rain we check if wind is high or low. Decision tree represents a flowchart like tree structure,

where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node holds a class label. This involves breaking down of the training set into different subsamples.

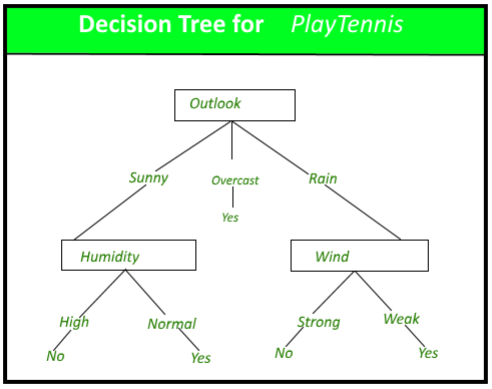

Figure 2.5: Decision tree [GEE]

Random forest utilizes the bootstrap concepts, which assert them simply re-running the same learning algorithm on different subsets of the data can result in highly correlated predictors, thus limiting the amount of variance reduction that is possible. Random forests tries to decorrelate the base learners by learning trees based on a randomly chosen subset of input variables, as well as a randomly chosen subset of data cases [Mur12]. An example of random forest is shown in figure 2.6.

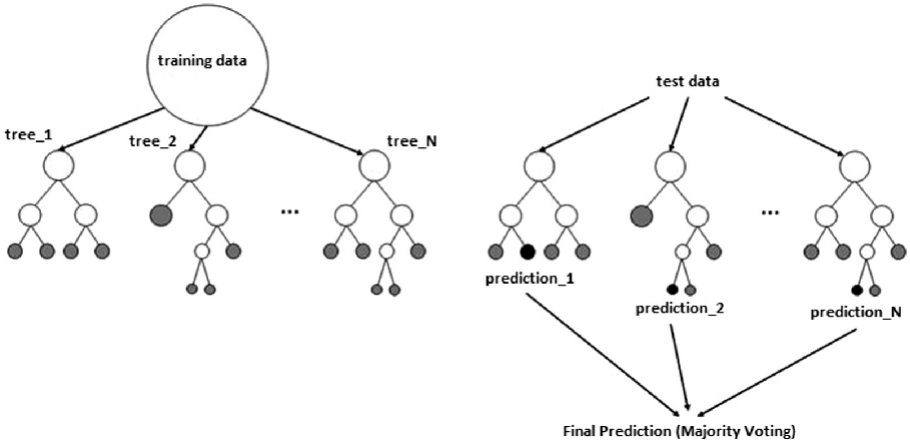

Figure 2.6: Random Forest [AEA+17]

In random forest algorithm, each new data point visits all the trees in the ensemble, which were grown using random samples from the training set. The function for aggregation will differ

depending upon the task (i.e. classification or regression). For regression task, it uses the average prediction values of each tree, whereas for classification, it uses the mode or most frequently predicted class by individual trees (also known as majority voting)[KDN].

2.2.11 Gradient Boosting

Gradient boosting, another ensemble approach, is a widely-used machine learning algorithm, due to its efficiency, accuracy, and interpretability. As discussed in section 2.2.5, gradient boosting trains the models in a sequential manner to create them strong learners from a weak one. The gradient boosting algorithm can be understood easily by understanding another boosting algorithm known as Adaptive boosting (AdaBoost). In AdaBoost, each tree is assigned an equal weight during training. After the evaluation of the first tree, those observations that are difficult to classify are given some extra weights and the weights for those observations that are easy to classify are lowered [TOWd]. As a result, the second tree is grown on the new weighted data. The main idea for AdaBoost improving the predictions made by the first tree. For the third tree, we compute the classification error from previous two trees and grow third tree to predict the revised residuals and so on. The final predictions is the weighted sum of the predictions made by the previous tree models.

Gradient boosting algorithms alos trains many models in additive, gradual, and sequential manner. Unlike AdaBoost, where it identifies the shortcomings by adjusting weights on data points, gradient boosting performs same by using gradients in loss function (y = mx + b + e, e being an error term). The loss function indicates how good the model's coefficients are fitting the data. Instead of a loss function that generally offers less control and which does not correspond with real world applications, gradient boosting allows one to optimize a user specified cost function. This is one of the biggest motivations of using gradient boosting [TOWd].

2.2.12 Feedforward Neural Networks

A feedforward neural network, also known as multilayer perceptron (MLP), is a series of logistic regression models stacked on top of each other, with the final layer being logistic or linear regression model, depending upon whether we are solving a classification or regression problem [Mur12]. Before going into multi-layer neurons lets us get some concepts of single neuron and its model. Neuron, also referred as "node" or "unit", is the basic unit of computation in a neural network. In a single neuron model, the node receives input from sources, the system does the calculations

where each input is complemented with weight (w) and produces the output. The node applies function f to the weighted input sum. This is shown in figure 2.7 where the network accepts two inputs x1 and x2 with weights w1 and w2, respectively. There is also the bias "b" which provide a trainable constant value for each node. The output is calculated as shown in figure. The function f is nonlinear and also called activation function.

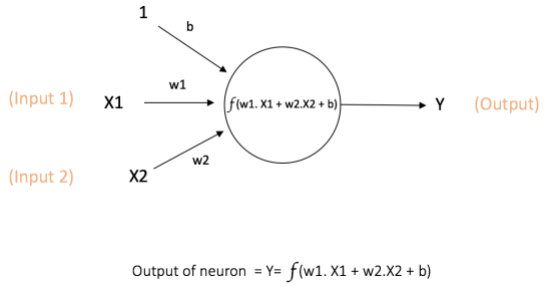

Figure 2.7: Single neuron[MED]

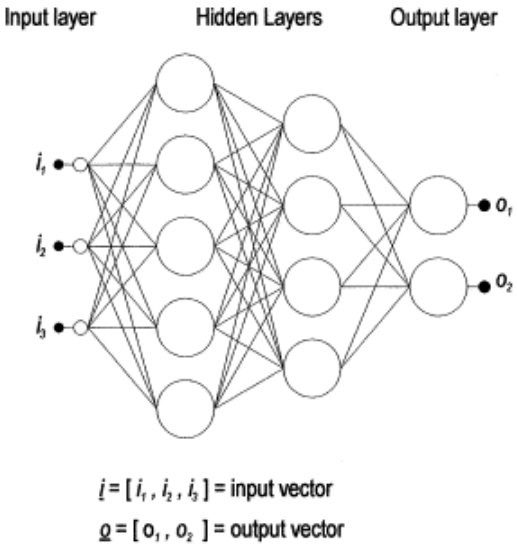

Figure 2.8: MLP with two hidden layers[GD98]

As in contrast to single neuron model, MLP is a model consisting of a system of simple interconnected nodes representing a non-linear mapping between an input vectors and output vectors[GD98]. The example of MLP with two hidden layers is as shown in figure 2.8 Each of the nodes are connected or assigned weights. The output signals are function of the sum of inputs to the node modified by activation function. MLP includes at least one hidden layer (except of one input and one output layer).

2.2.13 Evaluation Criteria

In machine learning, we use majority of data to train the model. And later we test the trained model with remaining portion of dataset to evaluate the performance of the created model. In this thesis, we used the following evaluation criteria for testing the performance of our model.

Confusion Matrix

A confusion matrix is an N * N matrix, where N is the number of classes (class labels), is a table (matrix) that is widely used to describe the performance of a classification model on a set of test data whose true values are known. The example of confusion matrix for binary classifier is shown in figure 2.9:

n=165	**Predicted: NO**	**Predicted: YES**
Actual: NO	50	10
Actual: YES	5	100

Figure 2.9: Confusion matrix for binary classifier [DAT]

There are two predicted classes "yes" or "no". The total number of examples is 165 and classifier predicted yes 110 times and no 55 times. For understanding the performance of the model let us get familiarize with some basic terms used in confusion matrix:

True Positive (TP): These are the cases in which the model predicted "yes" and actual value is also "yes".

False Positive (FP): These are the cases where the model predicted "yes" but the actual value was "no".

True Negative: (TN) These are the cases where the model predicted "no" and the actual value is also "no".

False Negative (FN): These are the cases where the model predicted "no" but the actual value was "yes"

From above example we can modify the confusion matrix as shown in figure 2.10:

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Figure 2.10: Confusion matrix for binary classifier II [DAT]

Recall

Recall, also known as sensitivity, attempts to answer the question: what proportion of actual positives were identified correctly? Mathematically, it can be represented as:

$$Recall = \frac{TP}{(TP + FN)} \tag{2.6}$$

In the above example, we have 100 TP and 5 FN so the recall, using above formula can be found as 100/105.

Precision

Precision attempts to answer the question: when it predicts "yes", how often it is correct? Mathematically, it can be represented as ratio of TP by sum of TP and FP.

$$Precision = \frac{TP}{(TP + FP)} \tag{2.7}$$

In the above example, using the above formula precision can be calculated as 100/110 that is 0.91.

True Negative Rate

True Negative Rate (TNR) also known as specifity is calculated as the number of correct negative predictions divided by total number of negative rate in the dataset. The best case for specificity is 1 and the worst case is 0. Mathematically, specifity can be represented as:

$$Specifity = \frac{TN}{(TN + FP)} \tag{2.8}$$

From above example, we can calculate TNR to be 50/60 or 5/6.

AUC-ROC curve

In a classification problem, we use AUC (Area Under the Curve) of ROC (Receiver Operating Characteristics) curve, to measure and visualize the performance for our model at various threshold settings. AUC-ROC curve is one of the most widely used and important evaluation metrics for checking any classification model's performance.

An example of AUC-ROC curve is shown in figure 2.11. ROC is a probability curve and AUC represents degree or measure of separability. Higher the AUC, better the model is at predicting. The ROC curve is plotted with True Positive Rate (TPR) against False Positive Rate (FPR) where TPR is on y-axis and FPR is on x-axis. TPR is also known as recall or sensitivity and FPR is given as:

$$FPR = 1 - Specifity \tag{2.9}$$

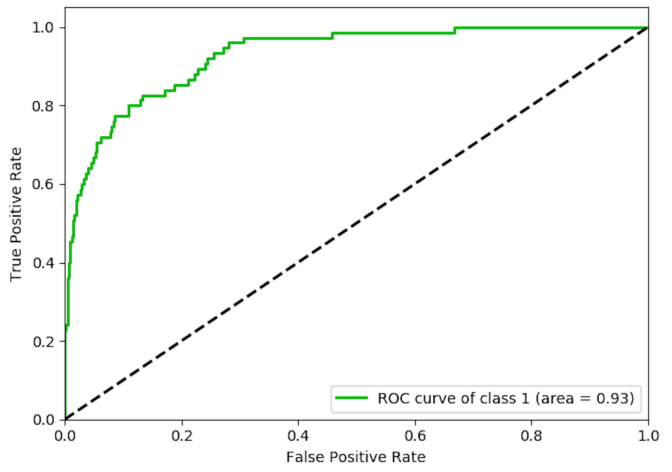

Figure 2.11: AUC$_{R}OCcurve$

Mean Squared Error

Mean squred error (MSE) is a metric used for regression analysis which tells how close a regression line is to set of points. In general, mean squared error (MSE) is the measures of the mean of the squares of the errors – the difference between the predicted values and true values. It is a risk function which is corresponding to the expected value of the squared error loss [MEM]. MSE is strictly positive because of square. If y_i is the true value for i^{th} point and yp_i be the estimated value for i^{th} instance, then mathematically MSE can be represented as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - yp_i)^2 \tag{2.10}$$

In this thesis, we used MSE as one of the evaluation criteria for analyzing the train and test dataset loss during the prediction of total spine BMD which is continuous values.

Chapter 3

Implementations

3.1 Data Description

The dataset implemented in this thesis is from MrOs (Osteoporotic Fractures in Men Study), a research study funded by the National Institutes of Health. This dataset contains phenotypes information as well as genotype informations for different SNPs. The dataset contains highly confidential information along with different numerical values for different features. Some brief description for the numerical values in the dataset are present in Table 3.1.

Variable	Calculated TypeType	Description
subjectId	Integer	De-identified Subject Id
ASCA	Decimal	Serum Calcium
B1THD	Decimal	Hologic 4500 Total Hip BMD
B1TLD	Decimal	4500 Total Spine BMD Values
BUAMEAN	Decimal	Mean of 3 BUA (Broad-band ultrasound attenuation) measures
FAANYHIP	Enum Integer (0 or 1)	Incident hip fracture
FAANYSLD	Enum Integer (0 or 1)	Incident proximal humerus fracture
FAANYWST	Enum Integer (0 or 1)	Incident wrist fracture
FAHIPFV1	Integer	Follow up time to first Incident Hip Fracture
FVDISPAR	Enum Integer (1 to 4)	Depth perception levels
AGE	Integer	Age of the patients
GRS_FN	Decimal	Femoral Neck Genetic Risk Score (GRS)
GRS_LS	Decimal	Lumber Spine Genetic Risk Score (GRS)
GSGRAVG	Decimal	Grip Strength
.....	.....	.....
.....	.....	.....

Table 3.1: Brief data description for the MrOs dataset

The dataset contains data from 5133 patients with two different types of BMD values and 3

different types of fractures cases. Out of 5133 , 5.98% of patients (i.e., 307) patients have fractures. The visualization for the fractured vs non fractured data can be seen in Figure 3.1.

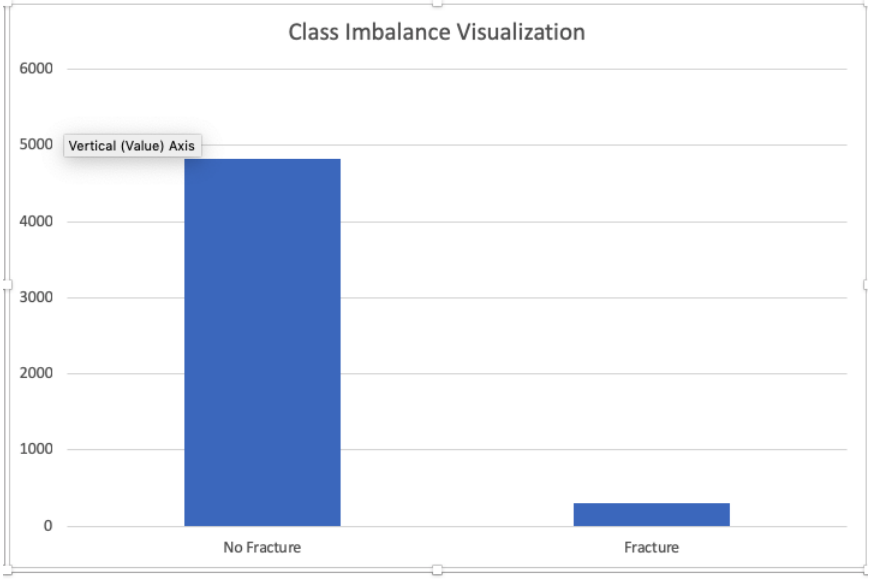

Figure 3.1: Data visualization for fracture analysis

3.2 Imputations

Initially, due to confidentiality requirements the dataset we had was in binary format. So, we needed to carryout pre-imputation process to make the file in variable coded file (vcf) format. After the pre-imputation process, the vcf files were uploaded in Michigan Imputation Server for genotype imputation service [IMP]. To this server, we can upload phased or un-phased GWAS genotypes data and can receive phased and imputed genomes, which would be used to calculate the GRS values later in post-imputation process. After the calculation of GRS values, we used other available and phenotype information of individuals to build different models.

3.3 Post Imputation Process for Risk Score Calculations

After the imputation was done, we received the vcf, info and vcf binary files from the Michigan Imputation Server. The info files contained different SNPs with their unique id (rsId), position and neighbors position (called as position+1) along with minor allele frequency (MAF), r-square values, allele and alternate alleles for each chromosomes. From this info file we extracted the 307 SNPs information. The extracted information is used to get all alleles pairing for each chromosomes which were used to generat ped files. From ped files we used the beta values for each SNPs to

calculate the genetic risk scores for each patient and for each chromosomes. The scores generated were weighted and unweighted GRS. In present work we've used the weighted GRS values for both femoral neck (FN) and lumber spine (LS). After the calculations were done for both FN and LS for each chromosome, we then summed up all the weighted GRS values and used them for analysis.

3.4 Data Normalization

Normalization is a technique mostly applied as a part of data preparation for machine learning models. The main goal of normalization is to change the values of numeric columns in the dataset to a common scale, without distorting differences in the range of values. Witho our dataset we performed mean variance normalization on the training dataset, where we obtain the normalized data by using following formula:

$$NormalizedData = \frac{x - \mu}{\sigma} \tag{3.1}$$

where,

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3.2}$$

is the mean and,

$$\sigma^2 = \sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} is the standard deviation. \tag{3.3}$$

3.5 Data Splitting

For each approach, we split the entire dataset into 80% as training set and 20% as test set. We specified a random number (random seed) while splitting the data, so as to ensure the same data split every time when the program is executed. We used training set for resampling and hyper-parameter tuning, and training the model. After the model was trained, we used the test set to evaluate the performance of the model.

3.6 Data Resampling

We mentioned earlier in section 3.1 that only about 6% of the cases actually are fracture cases. To improve the performance of the model we've implemented a resampling technique known as Synthetic Minority Over-sampling (SMOTE).

3.7 Hyperparameter Optimization

Hyperparameter tuning or optimization is the process of choosing a set of optimal hyperparameters for a learning algorithm. In contrast to model parameters, hyperparameter is the configuration that is external to the model. In this thesis, we used cross validation technique for tuning the hyperparameters. We've used k-fold cross validation where we set the value of k as 10. In 10-fold cross-validation, the training dataset is divided into 10 folds, and for each fold, we choose the current fold as a test set and remaining folds a a training set. We used sckit learn's randomized search cross validation method to find the best hyperparameters for different algorithms.

Chapter 4

Results

4.1 Fragility Fracture Prediction

We implemented four different algorithms: logistic regression, random forest, gradient boosting and multi-layer perceptron, for analysis of MrOs dataset. Following are the results obtained for each of models.

4.1.1 Logistic Regression

The AUC-ROC curve for logistic regression model with 307 SNPs is shown in figure 4.1. The area under curve is found to be .816. The confusion matrix for the same model can be seen in figure 4.2. The ROC AUC curve and confusion matrix for logistic regression with 1103 SNPs are shown in figure 4.3 and 4.4, AUC was higher with 0.904. Recall and precision for 307 SNPs were calculated to be 0.533 and 0.220 respectively. For 1103 SNPs recall was calculated to be 0.55 and precision was found to be 0.297.

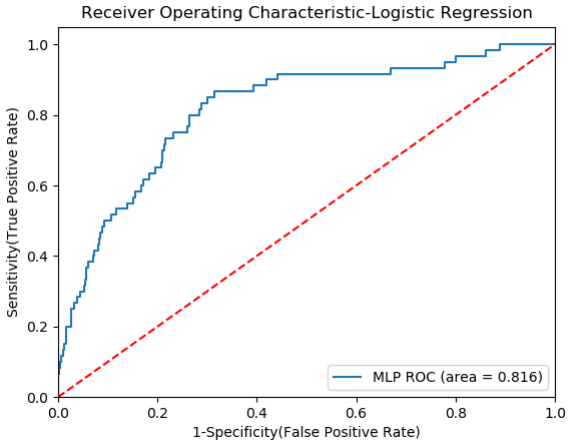

Figure 4.1: AUC-ROC curve for logistic regression (307 SNPs)

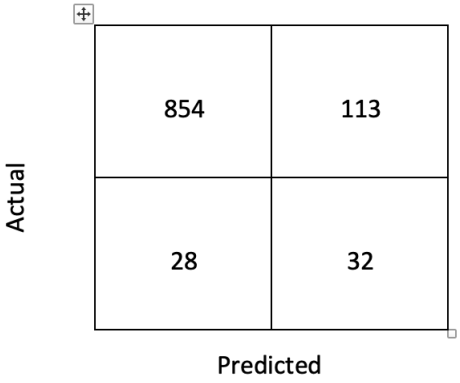

Figure 4.2: Confusion matrix for logistic regression (307 SNPs)

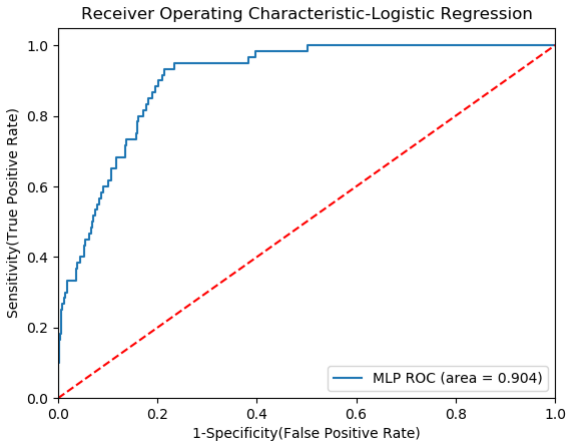

Figure 4.3: AUC-ROC curve for logistic regression (1103 SNPs)

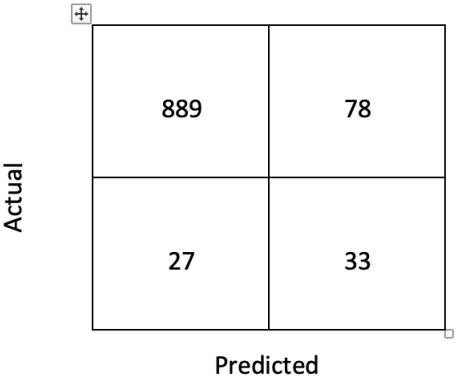

Figure 4.4: Confusion matrix for logistic regression (1103 SNPs)

4.1.2 Random Forest

The random forest model performed better than the logistic model, in both SNPs cases.. The overall evaluation for random forest model can be seen from figures 4.5 to 4.8. The AUC for 307 SNPs was 0.875 and for 1103 SNPs it was 0.916.

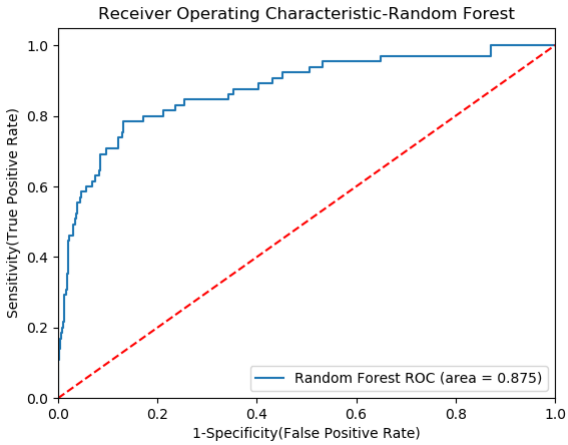

Figure 4.5: AUC-ROC curve for random forest (307 SNPs)

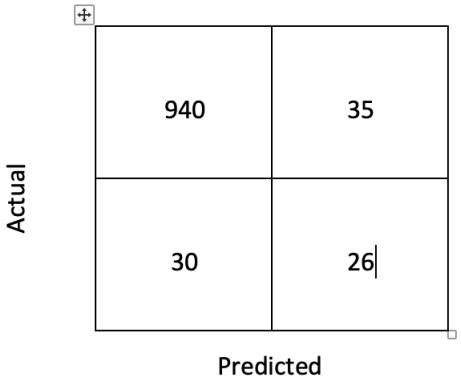

Figure 4.6: Confusion matrix for random forest (307 SNPs)

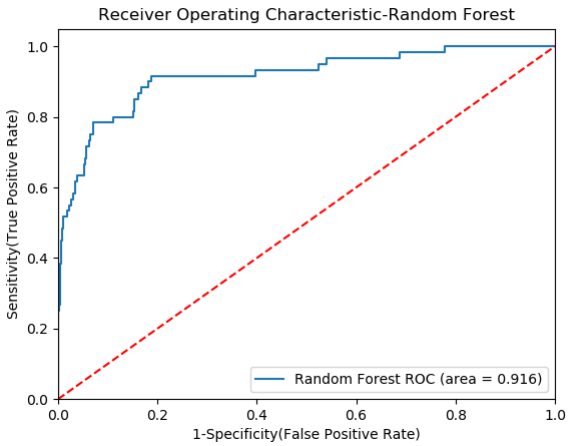

Figure 4.7: AUC-ROC curve for random forest (1103 SNPs)

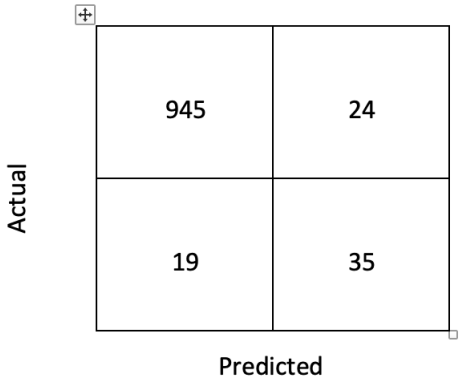

Figure 4.8: Confusion matrix for random forest (1103 SNPs)

4.1.3 Gradient Boosting

The AUC-ROC curve for gradient boosting is summarized in figure 4.9-4.12. The AUC for 307 SNPs was 0.866 and for 1103 SNPs it was 0.933.

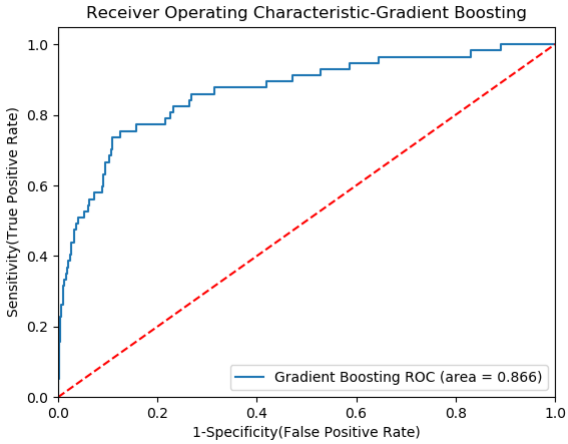

Figure 4.9: AUC-ROC curve for gradient boosting (307 SNPs)

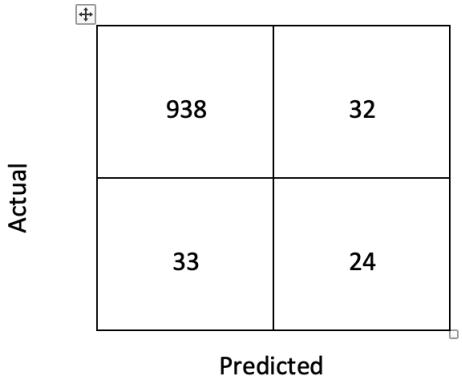

Figure 4.10: Confusion matrix for gradient boosting (307 SNPs)

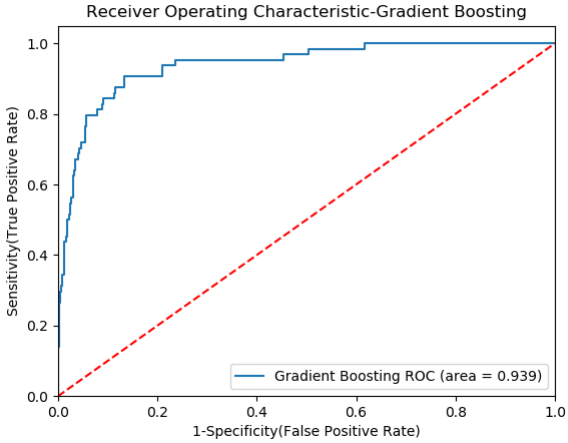

Figure 4.11: AUC-ROC curve for gradient boosting (1103 SNPs)

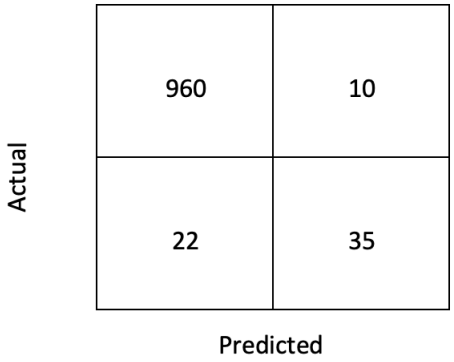

Figure 4.12: Confusion matrix for gradient boosting (1103 SNPs)

4.1.4 Multilayer Perceptron

Among the four algorithms for classifying fracture cases, MLP achieved the best result with ROC AUC curve of 0.97 for 307 SNPs and 0.981 for 1103 SNPs. The recall and precision for 307 SNPs were found to be 0.533 and 0.84 respectively. And, recall and precision for 1103 SNPs were 0.70 and 0.84 respectively. The overall performance for the model is as shown in figures 4.13-4.16

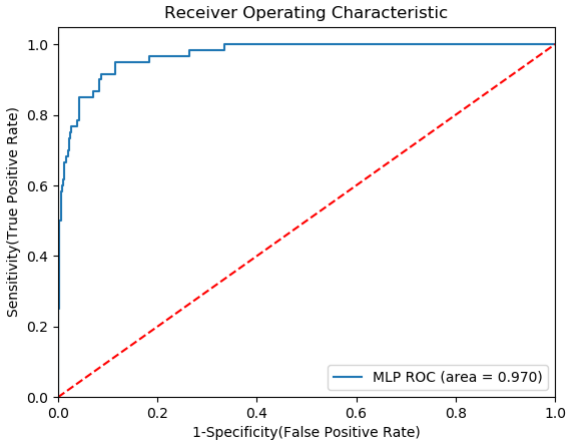

Figure 4.13: ROC AUC curve for MLP

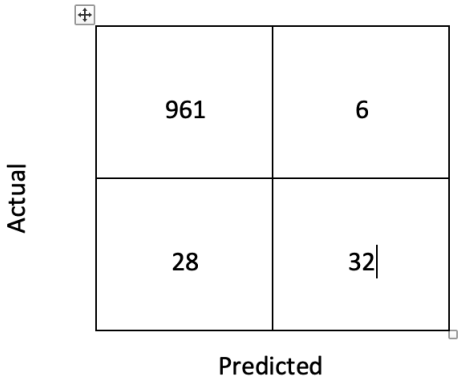

Figure 4.14: Confusion Matrix for MLP

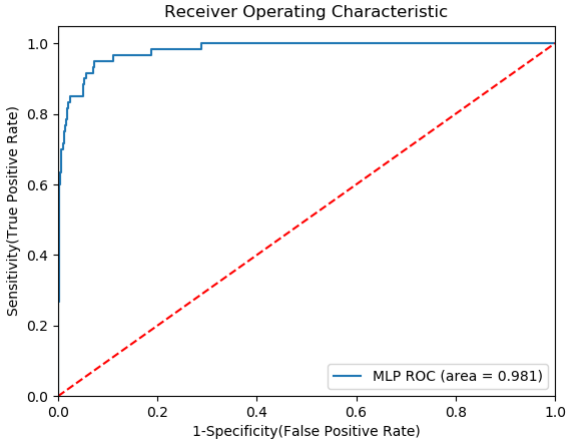

Figure 4.15: ROC AUC curve for MLP (1103 SNPs)

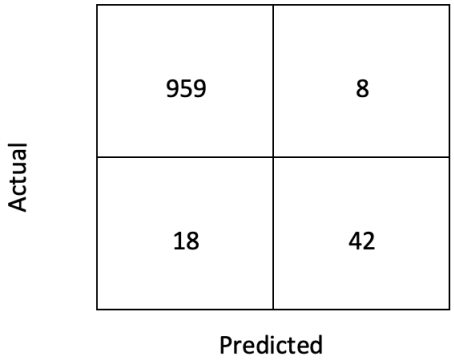

Figure 4.16: Confusion Matrix for MLP (1103 SNPs)

4.1.5 Fracture Prediction Results Summary

Table 4.1 summarizes the results for four different models with AUC, recall, and precision values for 307 SNPs and figure 4.2 summarizes the results for 1103 SNPs.

Model	AUC	Recall	Precision
Logistic Regression	0.816	0.53	0.220
Random Forest	0.875	0.464	0.426
Gradient Boosting	0.946	0.421	0.551
MLP	0.970	0.533	0.84

Table 4.1: Summary of the experiment results for fragility fracture risk prediction (307 SNPs)

Model	AUC	Recall	Precision
Logistic Regression	0.904	0.55	0.297
Random Forest	0.937	0.64	0.593
Gradient Boosting	0.933	0.614	0.77
MLP	0.981	0.70	0.84

Table 4.2: Summary of the experiment results for fragility fracture risk prediction (1103 SNPs)

4.2 Bone Mineral Density Prediction

For BMD prediction we implemented four different algorithms: linear regression, random forest, gradient boosting and multi-layer perceptron. In these experiments, we tried to predict Hologic total hip BMD (B1THD as in dataset) value. The output value is continuous variable and the evaluation metric used for this prediction was MSE. Following are the results obtained from different models for BMD prediction.

4.2.1 Linear Regression

The MSE for training and test datasets for 307 and 1103 SNPs are shown in figures 4.17 and 4.18. The mean squared error for 307 SNPs was found to be 0.1046 and for 1103 SNPs it was found to be 0.1030 on test data.

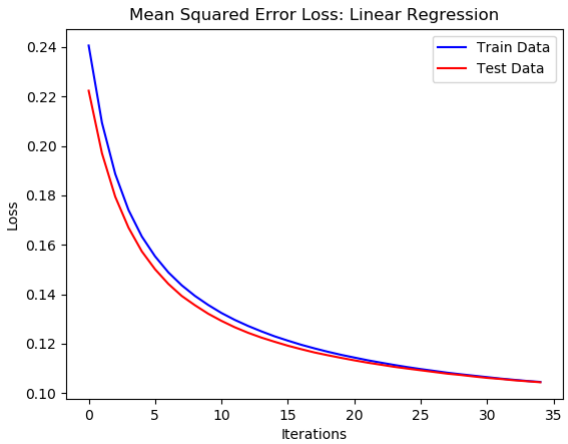

Figure 4.17: MSE vs iterations in linear regression (307 SNPs)

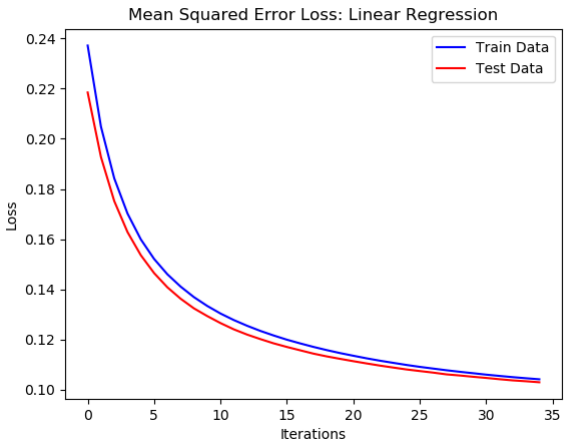

Figure 4.18: MSE vs iterations in linear regression (1103 SNPs)

4.2.2 Random Forest

Random forest better compared to other models. The MSE plot for training and test dataset for both 307 and 1103 SNPs are shown in figures 4.19 and 4.20. The mean squared loss for test data was only 0.00459 for 307 SNPs and 0.00433 for 1103 SNPs.

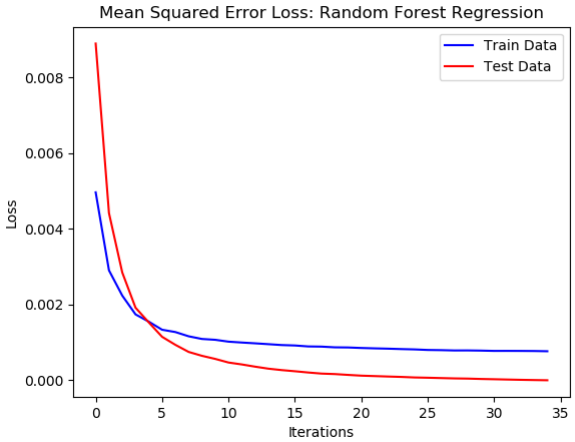

Figure 4.19: MSE vs Iterations in random forest (307 SNPs)

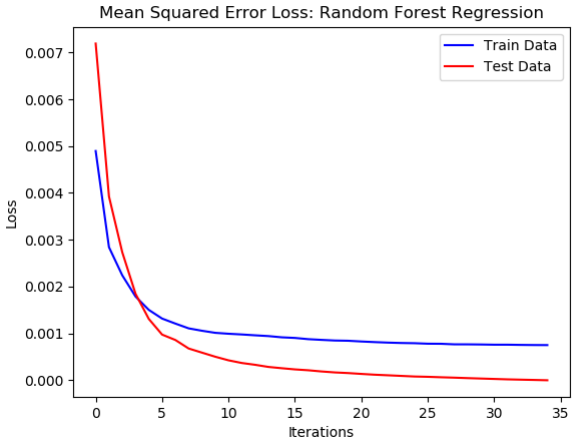

Figure 4.20: MSE vs Iterations in Random Forest (1103 SNPs)

4.2.3 Gradient Boosting

Gradient boosting had the second best performance for the prediction of Hologic BMD. The MSE for training and test data for 307 and 1103 SNPs are shown in figures 4.21 and 4.22. The mean

squared error for test set was 0.01143 for 307 SNPs and for 1103 SNPs the error was same i.e. 0.01143.

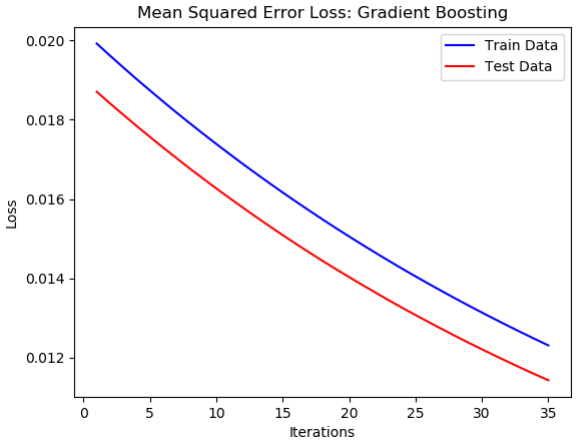

Figure 4.21: MSE vs iterations in gradient boosting (307 SNPs)

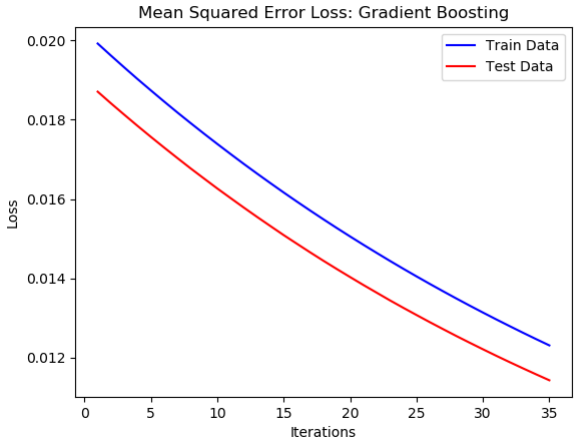

Figure 4.22: MSE vs iterations in gradient boosting (1103 SNPs)

4.2.4 Multilayer Perceptron

The mean squared error for the test set was 0.0972 and 0.0978 for 307 and 1103 SNPs respectively. The MSE for training and test data is shown in figures 4.23 and 4.24.

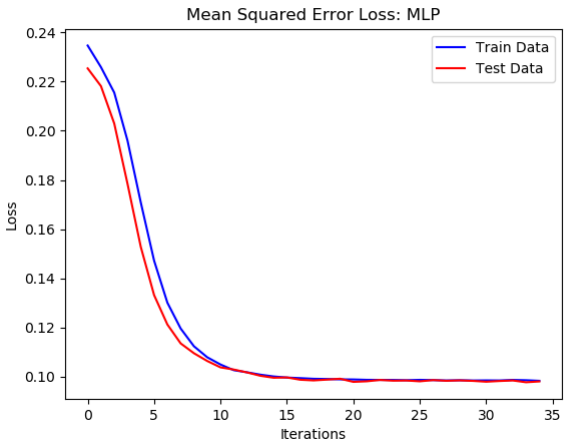

Figure 4.23: MSE vs Iterations in MLP (307 SNPs)

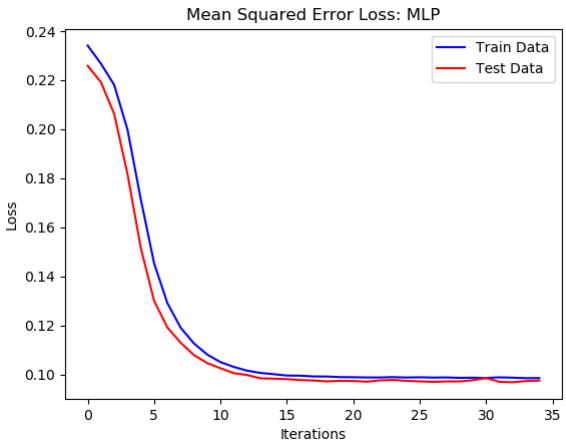

Figure 4.24: MSE vs Iterations in MLP (1103 SNPs)

4.2.5 Bone Mineral Density Prediction Results Summary

Table 4.3 and 4.4 summarize the MSE values for train and test set in both 307 and 1103 SNPs experiments.

Model	MSE for Train Set	MSE For Test Set
Linear Regression	0.1038	0.1046
Random Forest	0.007655	0.00459
Gradient Boosting	0.0123	0.0114
Multilayer Perceptron	0.0981	0.0978

Table 4.3: MSE for different models for BMD prediction (307 SNPs)

Model	MSE for Train Set	MSE For Test Set
Linear Regression	0.1038	0.1030
Random Forest	0.00076	0.00459
Gradient Boosting	0.01143	0.0123
Multilayer Perceptron	0.0979	0.0972

Table 4.4: MSE for different models for BMD prediction (1103 SNPs)

Chapter 5

Conclusions and Future Works

In this thesis, we employed supervised machine learning approach for predictive analysis for the osteoporosis data set. We performed two different predictive analysis using different machine learning models. For the first part of the analysis – fragility prediction – multi-layer perceptron (MLP) performed better than other predictive models. We had class unbalanced distribution of the data set, in which the model tends to be biased towards majority samples class. To tackle this problem, we implemented Synthetic Minority Over-sampling Technique (SMOTE), so that performance of the model can be increased. The best performance for this analysis was seen in MLP with AUC being 0.970 for 307 SNPs and 0.981 for 1103 SNPs. The recall and precision were calculated as 0.533 and 0.84 respectively for 307 SNPs. For 1103 SNPs the recall increased to 0.70 and precision stayed same at 0.84. Among four models logistic regression had poor AUC. The AUC for 307 SNPs was 0.816 and it increased to 0.904 in case of 1103 SNPs.

For the second part of the analysis, we implemented four different algorithms and the performance evaluation was done on the basis of mean squared loss in train and test data. The output variable was the hip bone mineral density values, which is continuous, and was plotted versus number of iterations to train the model. The performance for random forest was better among others and the worst performance was shown by linear regression with the mean squared error of 0.103 in 1103 SNPs and 0.1046 in 307 SNPs for test data.

For both analysis, we used different phenotype risk information along with clinical risk factors and weighted genetic risk scores (GRS) for each chromosomes. We did separate analysis for 307 and 1103 SNPs that were significantly associated with fracture risk. Recent studies have shown that there are more than 1300 SNPs associated with fracture risk for osteoporosis, so these SNPs can be used for further predictive analysis in the future. A good AUC value was obtained for the

classification problem; but the recall and precision weren't as high in both sets of SNPs. Further work can be conducted to improve the recall and precision. Deep learning tools can be implemented so as to get higher precision and higher recall in future. Also, we had 5133 patients data in the dataset for the analysis and further more data can be collected for better or more complicated analysis. The new studied risk factors can be included for better understanding and proposing a good predictive model for minimizing the osteoporosis risk fracture.

Appendix A

Certification of CITI Program

Figure A.1: Biomedical IRB course (basic) Certification

Bibliography

[AEA+17] Koray Acici, Cagatay Berke Erdas, Tunc Asuroglu, Munire Kilinc Toprak, Hamit Erdem, and Hasan Ogul. A random forest method to detect parkinson's disease via gait analysis. *International Conference on Engineering Applications of Neural Networks*, 744:611–614, 2017.

[BEC] https://becominghuman.ai/ensemble-learning-bagging-and-boosting-d20f38be9b1e [Online Accessed March 13, 2019].

[CCK+13] Hsueh-Wei Chang, Yu-Hsien Chiu, Hao-Yun Kao, Cheng-Hong Yang, and Wen-Hsien Ho. Comparison of classification algorithms with wrappe-based feature selection for predicting osteoporosis outcome based on genetic factors in a taiwanese women population. *International Journal of Endocrionolgy*, 2013, 2013.

[CDA+17] Egejuru Ngozi Chidozie, Mhambe Priscilla Dooshima, Balogun Jeremiah Ademola, Femi Komolafe, and Idowu Peter Adebayo. Osteoporosis risk predictive model using supervised machine learning algorithms. *Science Research*, pages 78–87, 2017.

[DAT] https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/[Online Accessed March 17,2019].

[Dey16] Ayun Dey. Machine learning algorithms: A review. *International Journal of Computer Science and Information Technologies*, 7(3), 2016.

[EBI] https://www.ebi.ac.uk/training/online/course/gwas-catalog-exploring-snp-trait-associations/why-do-we-need-gwas-catalog/what-are-genome[Online Accessed March 14,2019].

[ESE+12] K Estrada, U Styrkarsdottir, E Evangelou, Yi-Hsiang Hsu, E L Duncan, Evangelia E Ntzani, Ling Oei, Omar M E Albagha, Najaf Amin, JP Kemp, Daniel L

Koller, Guo Li, Ching-Ti Liu, Ryan L Minster, Alireza Moayyeri, Liesbeth Vandenput, Dana Willner, Su-Mei Xiao, Laura M Yerges-Armstrong, Hou-Feng Zheng, Nerea Alonso, Joel Eriksson, Candace M Kammerer, Stephen K Kaptoge, Paul J Leo, Gudmar Thorleifsson, Scott G Wilson, James F Wilson, Ville Aalto, Aaron K Aragaki Markku Alen, Thor Aspelund, Jacqueline R Center, Zoe Dailiana, David J Duggan, Melissa Garcia, Natalia Garcia-Giralt, Sylvie Giroux, Goran Hallmans, Lise Bjerre Husted Lynne J Hocking, Karen A Jameson, Rita Khusainova, Ghi Su Kim, Charles Kooperberg, Theodora Koromila, Marcin Kruk, and Marika Laaksonen. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature Genetics*, 44, 2012.

[Eth10] Ethem Alpaydin. *Introduction to Machine Learning.* The MIT Press, 2010.

[FKBMF+18] Vincenzo Forgetta, Julyan Keller-Baruch, Audrey Durand Marie Forest, Sahir Bhatnagar, John Kemp, John A Morris, John A Kanis, Douglas P Kiel, Eugene V McCloskey, Fernando Rivadeneira, Helena Johannson, Nicholas Harvey, Cyrus Cooper, David M Evans, Joelle Pineau, William D Leslie, Celia MT Greenwood, and J Brent Richards. Machine learning to predict osteoporotic fracture risk from genotypes. *bioRxiv*, pages 1–3, 2018.

[GD98] M.W. Gardner and S.R. Dorling. Artificial neural networks (the multilayer perceptron) – review of applications in the atmospheric sciences. *Atmospheric Environment*, 1998.

[GEE] https://www.geeksforgeeks.org/decision-tree/[Online Accessed March 12,2019].

[GHR] https://ghr.nlm.nih.gov/primer/genomicresearch/snp [Online Accessed March 21, 2019].

[IAA14] Theodoros Iliou, Christos-Nikolaos Anagnostopoulos, and George Anastassopoulos. Osteoporosis Detection Using Machine Learning Techniques and Feature Selection. *International Journal on Artificial Intelligence Tools*, 23:1–2, 2014.

[IMP] https://imputationserver.sph.umich.edu/index.html!pages/home[Online Accessed October 20, 2018].

[KDN] https://www.kdnuggets.com/2017/10/random-forests-explained.html[Online Accessed March 13,2019].

[Kim18] Stuart K. Kim. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral densit, osteoporosis and fracture. *Plos One*, 13, 2018.

[KMG+17] JP Kemp, JA Morris, CM Gomez, Vincenzo Forgetta, Nicole M Warrington, Scott E Youlten, Jie Zheng, Celia L Gregson, Elin Grundberg, Katerina Trajanoska, Andrea S Pollard John G Logan, Penny C Sparkes, Elena J Ghirardello, Rebecca Allen, Victoria D Leitch, Natalie C Butterfield, Davide Komla-Ebri, Anne-Tounsia Adoum, Katharine F Curry, Jacqueline K White, Fiona Kussy, Keelin M Greenlaw, Changjiang Xu, Nicholas C Harvey, Cyrus Cooper, David J Adams, Celia MT Greenwood, Matthew T Maurano, Stephen Kaptoge, Fernando Rivadenerira, Jonathan H Tobias, Peter I Croucher, Cheryl, L Ackert-Bicknell, JH Duncan Basset, Graham R Williams, J Brent Richards, and David M Evans. Identification of 153 new loci associated with heel bone mineral density and functional involvement of gpc6 in osteoporosis. *Nature Genetics*, 49, 2017.

[MED] https://medium.com/@jayeshbahire/perceptron-and-backpropagation-970d752f4e44 [Online Accessed March 13, 2019].

[MEM] https://medium.freecodecamp.org/machine-learning-mean-squared-error-regression-line-c7dde9a26b93[Online Accessed March12,2019].

[MKY+19] John A. Morris, John P. Kemp, Scott E. Youlten, Laetitia Laurent, John G. Logan, Ryan C.Chai, Nicholas A. Vulpescu, Vicenzo Forgetta, Aaron Kleinman, C. Marcelo Sergio Sindhu T. Mohanty, Julian Quinn, Loan Nguyen-Yamamoto, Amiee-Lee Luco, Jinchu Vijay, Marie-Michelle Simon, Albena Pramatarova, Carolina Medina-Gomez, Katerina Trajanoska, Elena J. Ghirardello, Natalie C. Butterfield, Katharine F. Curry, Voctoria D. Leitch, Penny C. Sparkes, Anne-Tounsia Adoum, Naila S. Mannan, Davide S.K. Komla-Ebri, Andrea S. Pollard, Hannah F. Dewhurst, Thomas A.D. Hassal, Michael-John G. Beltejar, 23andMe Research Team, DJ Adams, SM Vaillancourt, S. Kaptoge, P. Baldock, C. Cooper, J. Reeve, Evangelia E. Ntzani, E. Grundberg, D. Goltzman, DJ Adams, DA Hinds CJ Lelliott,

CL Ackert-Bicknell, Yi-Hsiang Hsu, MT Maurano, PI Croucher, GR Williams, JH Duncan Bassett, DM Evans, and JB Richards. An atlas of genetic influences on osteoprososis in humans and mice. *Nature Genetics*, 51:258–261, 2019.

[MLR] https://ml-cheatsheet.readthedocs.io/en/latest/logistic_regression.html [Online Accessed March 11,2019].

[Mur12] Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[RJA+17] Aditya Rawat, Akshav Jain, Arpit Arora, Bhumika Gupta, , and Naresh Dhami. Analysis of various decision tree algorithms for classification in data mining. *International Journal of Computer Applications*, 163, 2017.

[TOWa] https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a[Online Accessed March 09,2019].

[TOWb] https://towardsdatascience.com/multi-label-text-classification-with-scikit-learn-30714b7819c5 [Online Accessed March 07,2019].

[TOWc] https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd [Online Accessed March 11,2019].

[TOWd] https://towardsdatascience.com/understanding-gradient-boosting-machines-9be756fe76ab [Online Accessed March 13,2019].

[WIL] https://en.wikipedia.org/wiki/Linear_regression[Online Accessed March 09,2019].

[YKK+13] Tae Keun Yoo, Sung Kean Kim, Deok Won Kim, Joon Yul CHoi, Wan Hyung Lee, Ein Oh, and Eun-CHeol Park. Osteoporosis risk prediction for bone mineral density assessment of postmenopausal women using machine learning. *Synapse*, 54(6):1321–1323, 2013.

Curriculum Vitae

Graduate College

University of Nevada, Las Vegas

Bibek Bhattarai

bibekbhattarai.brt@gmail.com

Degrees:

Bachelor Degree in Computer Engineering 2012

Tribhuvan University, Nepal

Thesis Title: Machine Learning Approach for prediction of Bone Mineral Density And Fragility Fracture in Osteoporosis

Thesis Examination Committee:

Chairperson, Dr. Fatma Nasoz, Ph.D.

Committee Member, Dr. Justin Zhan, Ph.D.

Committee Member, Dr. Laxmi Gewali, Ph.D.

Graduate Faculty Representative, Dr. Qing Wu, Ph.D.