PROJECT REPORT

PROJECT: EFFECTIVE AUTOMATED PREDICTION OF VERTEBRAL COLUMN PATHOLOGIES.

MENTOR: MR. ASHWINI THAPLIYAL

Submitted by:

Ankush Kamboj

ACKNOWLEDGEMENT

I have taken efforts in this project. However, it would not have been possible without the kind support and help of my mentor Mr. Ashwini Thapliyal. I would like to extend my sincere thanks to him for guiding me through the field of Medical Diagnosis, for constant supervision, motivation as well as for providing necessary information regarding the project.

I would also like to express my gratitude towards the members of Oil and Natural Gas Corporation for their kind co-operation and encouragement which helped me in completion of this project.

CERTIFICATE

This is to certify that this project titled “Effective Automated Prediction of Vertebral Column Pathologies” submitted by Ankush Kamboj (DTU/2K15/CO/032) in partial fulfilment for the requirements for the award of Bachelor of Technology Degree in Computer Engineering (COE) at Delhi Technological University is an authentic work carried out by the student under my supervision and guidance. To the best of my knowledge, the matter embodied in the report has not been submitted to any other university or institute for the award of any degree or diploma.

Mr. Ashwini Thapliyal
(Programmer)
Oil and Natural Gas Corporation

ABSTRACT

Medical science is characterized by the correct diagnosis of a disease and its accurate classification to avoid complexities at treatment/medication stage. This is often accomplished by a physician based on experience without much signal processing aids. It is envisioned that a sophisticated and intelligent medical diagnostic/classification system may be helpful in making right decisions

Classifying data is one of the most common task in Machine learning. Machine learning provides one of the main features for extracting knowledge from large databases from enterprises operational databases. Machine Learning in Medical Health Care is an emerging field of very high importance for providing prognosis and a deeper understanding of medical data. Most machine learning methods depend on a set of features that define the behavior of the learning algorithm and directly or indirectly influence the performance as well as the complexity of resulting models.

We investigate different algorithms such as ***KNearest Neighbors, Random forest, Logistic Regression*** *along with* ***back-propagation Neural Network (BPNN*** *for the diagnosis of spondylolisthesis and Disk-Hernia. We then use this model to deploy our application to automate priority-based follow-up appointment scheduling and also forward referral application to make appointments with specialty hospitals.*

The system has been implemented in Python platform and trained using proprietary dataset and benchmark dataset from UCI machine learning repository.

CONTENTS

INTRODUCTION

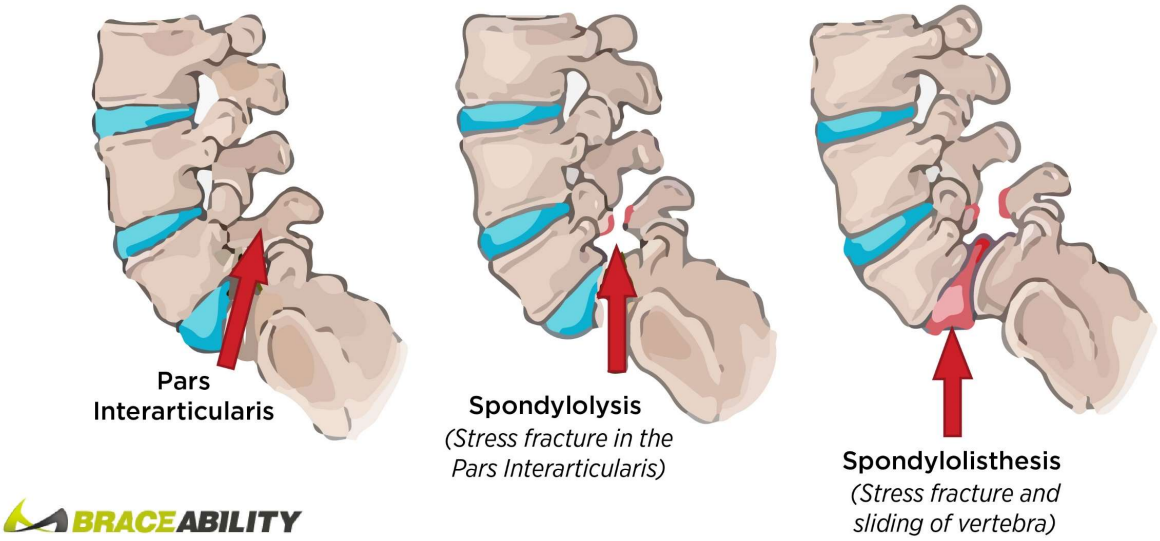
Vertebral Column

Various physical diseases come with age, of which many are inherited and inevitable. The vertebral column plays an important role in body motion, which supports the head and trunk and protects the nerves governing body activities and sensations. In addition, it is the main transfer passage of the nerves

Vertebral column is an integral part of human body. It is a structure which consists of usually thirty-three vertebrates, out of which twenty-four are articulating and nine are fused vertebras. It is found in the dorsal aspect of the torso and is separated by number of intervertebral discs. These thirty-three vertebras are divided in five different groups which include seven vertebrae in cervical curve, twelve vertebrae in thoracic curve, five vertebrae in lumbar curve and nine vertebrae in sacral curve. Intervertebral discs are interposed between the vertebral bodies, and serve not only as shock absorber for the column but also provide the normal mobility between the adjacent vertebrae. Each disc consists of a soft central portion of spongy material. The two vertebral disorders discussed in this paper are disc hernia and spondylolisthesis. Disc hernia is an intervertebral disc protrusion that is produced by the effect of flexion force acting upon the most mobile portions of the spine. A sudden strain with the spine in an unguarded position will rupture the tough annulus, allowing portions of the torn annulus and soft nucleus to escape into the spinal canal. Spondylolisthesis is a condition in which a lower lumbar vertebra, usually the fifth slips forward through the plain of the intervertebral disc below it and so carries with it the whole of the upper portion of the spine.

Experiment Setup

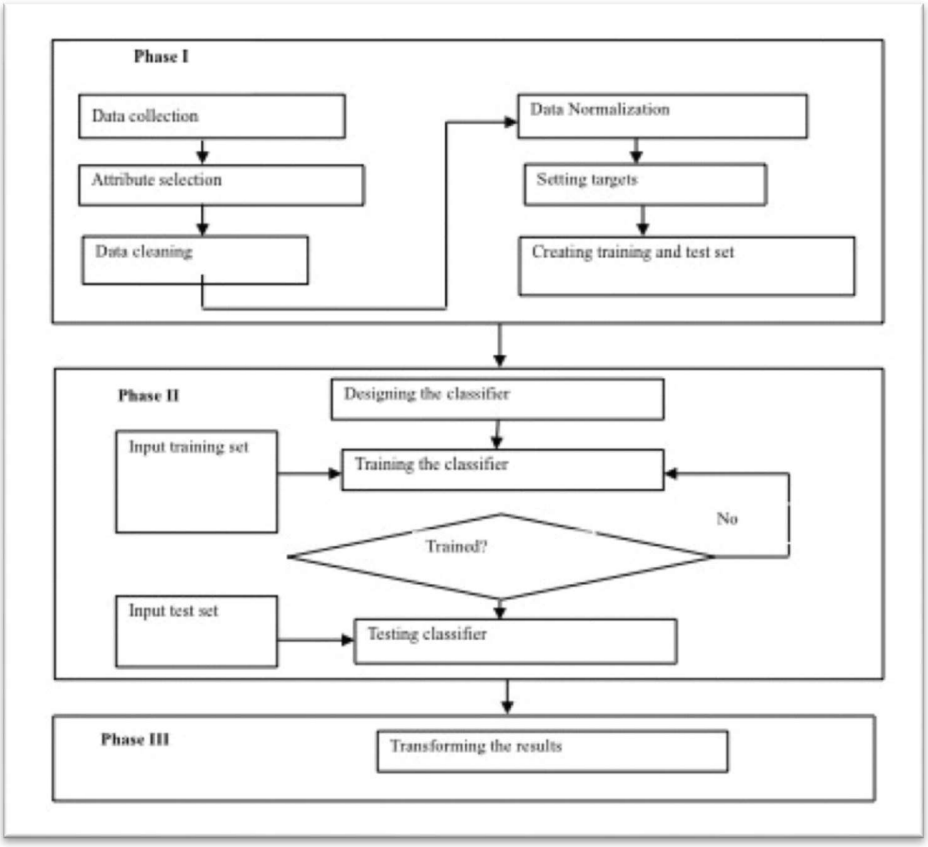
The problem area of vertebral column in humans is spinal cord. Here we have taken data set having values for six biomechanical features used to classify orthopedic patients into 3 classes (normal, disk hernia or spondylolisthesis) or 2 classes (normal or abnormal). The first task involves in categorizing patients as fit in to one out of three categories: Normal (100 patients), Disk Hernia (60 patients) or Spondylolisthesis (150 patients). For the second task, the categories Disk Hernia and Spondylolisthesis were combined into a single category labelled as 'abnormal'. Thus, the second task consists in classifying patients as fitting to one out of two categories: Normal (100 patients) or Abnormal (210 patients). Each patient is characterized in the data set by six biomechanical attributes resultant from the shape and alignment of the pelvis and lumbar spine (in this order): pelvic occurrence, pelvic tilt, lumbar lordosis position, sacral slope, pelvic radius and mark of Spondylolisthesis. The following settlement is used for the class labels: DH (Disk Hernia), Spondylolisthesis (SL), Normal (NO) and Abnormal (AB). A herniated disk and Spondylolisthesis are two possibly painful circumstances that can affect the steadiness and function of the spinal column. While herniation affects the discs between the spinal bones (vertebrae), Spondylolisthesis affect the bones themselves.



PROPOSED METHOD FOR DIAGNOSIS OF VERTEBRAL COLUMN DISORDERS

A: Methodology:

We have used four classifiers to diagnose vertebral column disorders, which includes a kNN classifier, a logistic regression model, a backpropagation neural network and random forest classifier. The disorders are classified as disc hernia spondylolisthesis and normal patients. The methodology used by all the classifiers is divided in to three phases. The detail of these phases is discussed below. Figure 1 presents the block diagram of these phases.



Phase 1: Pre-Processing

In this phase the data is first collected. During the collection the attributes are selected after discussing them with the doctor. We have incorporated all the attributes as they are significant and have a huge influence on the class labels. Next the data is analyzed and cleaned. In the cleaning process the missing and noisy values are handled. Afterwards the data is normalized. Subsequently, each sample in the data set is allotted its targets class. The entire data set is divided into two sets, the training and the test sets. In case of the validation set, the training set is further divided in to two sub groups, the training group and the validation group. After this the dataset is ready to be provided to the classifiers.

Phase 2: Design Classifier

After the pre-processing is complete, the data is presented to the classifiers. All the classifiers belong to the supervised learning category. For the supervised classifier the training set includes both the samples and their targets. These trained classifiers learn the pattern from the samples provided to them and diagnose the correct disorder. Each classifier is trained two times with two different datasets; i.e. the 60% ratio and the 10-fold cross validation. The performance of all four classifiers is evaluated. If the classifier is not trained the training process is repeated with different structure and functions.

Phase 3: Post-Processing

In this phase the results of the trained classifiers are converted into a form that is easily understandable. The results show whether a person is normal or has a vertebral column disorder and the type of disorder either disc hernia or spondylolisthesis.

B: Material

The data set used for classifying vertebral column disorders is taken from the UCI machine learning database. The dataset contains 310 samples out of which 60 samples are of disc hernia, 150 samples of spondylolisthesis and 100 normal samples. Each sample has six attributes. All the attributes are numerical. The second and sixth attribute
contains some samples having negative values as well. All these attributes are used to classify between disc hernia, spondylolisthesis and normal patients. The attribute along with their minimum and maximum values are presented in table I and **figure**.

Attributes	**Values**	
	Minimum	*Maximum*
Pelvic Incidence	26.1479	129.834
Pelvic Tilt	-3.7599	49.4319
Lumbar Lordosis Angle	14.00	125.7424
Sacral Slope	13.3669	121.4296
Pelvic Radius	70.0826	163.071
Degree Spondylolisthesis	-11.0582	418.5431

CLASSIFIERS STRUCTURE

A. Feedforward Backpropagation Neural Network

The NN we used has an architecture 6-5-4-3. The input layer has 6 nodes because each sample has six features and so there are six values in each input record. The hidden and the output layers have 5, 4 and 3 neurons respectively. These neurons perform all the computation. The activation or transfer function we used for both the hidden and output layer neurons is hyperbolic tangent sigmoid function. It is relatively faster than the other activation functions.

$$a = tansig(n) = \frac{2}{1 + e^{(-2*n)}} - 1$$

Where **n** is a value of the hidden and output layer neurons given to the hyperbolic tangent sigmoid activation function. The error is calculated using the mean squared error performance function:

$$Eav = 1/N \sum_{n=1}^{N} E(n)$$

Where $\mathbf{E_{av}}$ is the average error of the network, **N** is the total number of training samples and **E(n)** is the network error. The learning algorithm used is the gradient descent with momentum weight and bias learning function. It
changes the weights in such a way that it reduces the error between the output and the targets. The learning function
is:

$$dW = mc * dWprev + (1 - mc) * lr * gW$$

Where **dW** is the weight change for a particular neuron from its input and error, $\mathbf{dW_{prev}}$ is previous weight change, **lr** is the learning rate in our case it is 0.01**, mc** is momentum constant which is 0.9 in our situation and **W** is either the weight or bias. The weights are changed based on the gradient decent method.
The training algorithm we used is a variant of backpropagation algorithm which is levenberg-marquardt backpropagation. It is based on the error correction rule. The backpropagation algorithm uses the jacobian matrix which contains first derivative of network error with respect to the weights and bias. The Levenberg–Marquardt algorithm is a combination of steepest descent and the gauss–newton algorithm. Its updating rule is:

$$w_{k+1} = w_k - [J^T J + \mu I]^{-1} J^T e$$

Where **w** is the weight, **J** is the jacobian matrix, **T** means transpose, μ is the combination coefficient which is always
positive, I is the identity matrix, and e is the error. The $\mathbf{J^t}\boldsymbol{e}$ is the gradient. The combination coefficient allows the Levenberg–Marquardt algorithm to switch between the steepest decent and netwon algorithms during the training process. The training set provided to FFNN is divided in two parts the training set and the validation set. The validation set monitors the training process and stops training when the error starts to increase instead of decrease. Both the data distribution (80% ratio and 10-fold cross validation) used the same FFNN architecture, training, learning, activation and performance functions.
The weights are converged after 100 epochs in case of 80% ratio data distribution approach whereas in 10-fold approach weights are converged after 50 epochs.

B: Logistic Regression

Logistic regression, despite its name, is a linear model for classification rather than regression. Logistic regression is also known in the literature as logit regression, maximum -entropy classification (MaxEnt) or the log-linear classifier. In this model, the probabilities describing the possible outcomes of a single trial are modeled using a logistic function.

The implementation of logistic regression in scikit-learn can be accessed from class Logistic-Regression. This implementation can fit a multiclass (one-vs-rest) logistic regression with optional L2 or L1 regularization.

As an optimization problem, binary class L2 penalized logistic regression minimizes the following cost function:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

Similarly, L1 regularized logistic regression solves the following optimization problem:

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

Definition of the logistic function:

An explanation of logistic regression can begin with an explanation of the standard logistic function. The logistic function is useful because it can take an input with any value from negative to positive infinity, whereas the output always takes values between zero and one and hence is interpretable as a probability. The logistic function is defined as follows:

$$\sigma(t) = \frac{e^t}{e^t + 1} = \frac{1}{1 + e^{-t}}$$

C: Fuzzy k-Nearest Neighbours

In pattern recognition, the kNearest Neighbors algorithm (or kNN for short) is a nonparametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. kNN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The kNN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for kNN classification) or the object property value (for kNN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A shortcoming of the kNN algorithm is that it is sensitive to the local structure of the data.

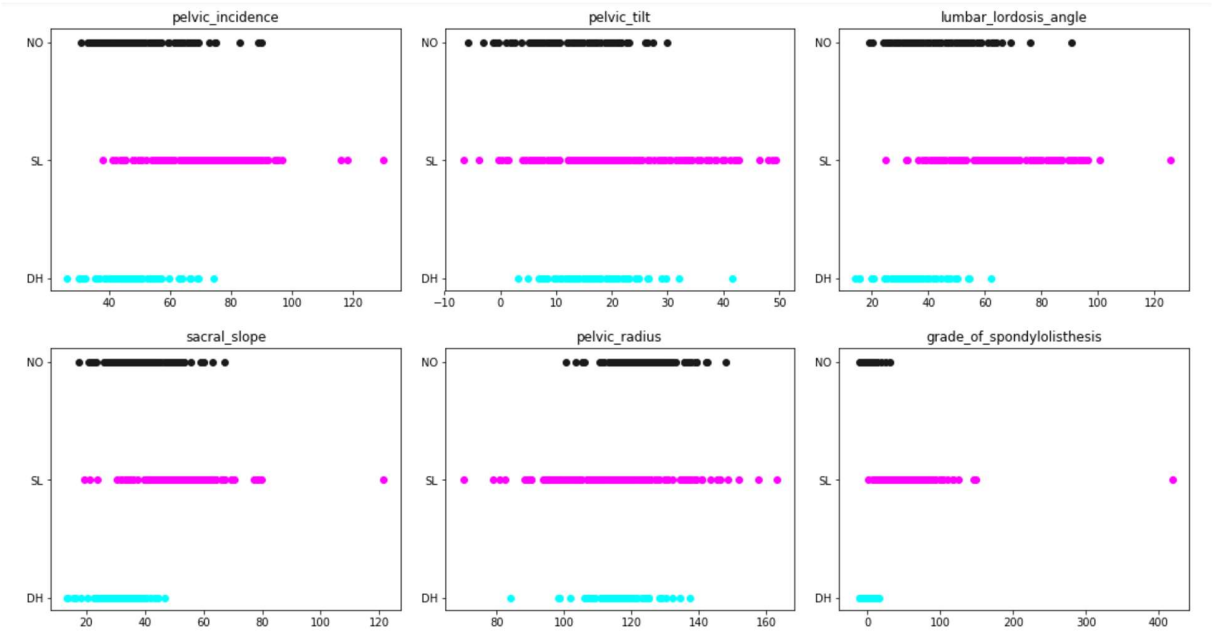
The basis of the algorithm is to assign membership as a function of the vector's distance from its k-nearest neighbors and those neighbors' memberships in the possible classes. The fuzzy algorithm is similar to the crisp version in the sense that it must also search the labeled sample set for the k-nearest neighbors. Beyond obtaining these k samples, the procedures differ considerably. While the fuzzy Kk-nearest neighbor procedure is also a classification algorithm the form of its results differ from the crisp version. The fuzzy k-nearest neighbor algorithm assigns class membership to a sample vector rather than assigning the vector to a particular class. The advantage is that no arbitrary assignments are made by the algorithm. In addition, the vector's membership values should provide a level of assurance to accompany the resultant classification.

D: Random Forests

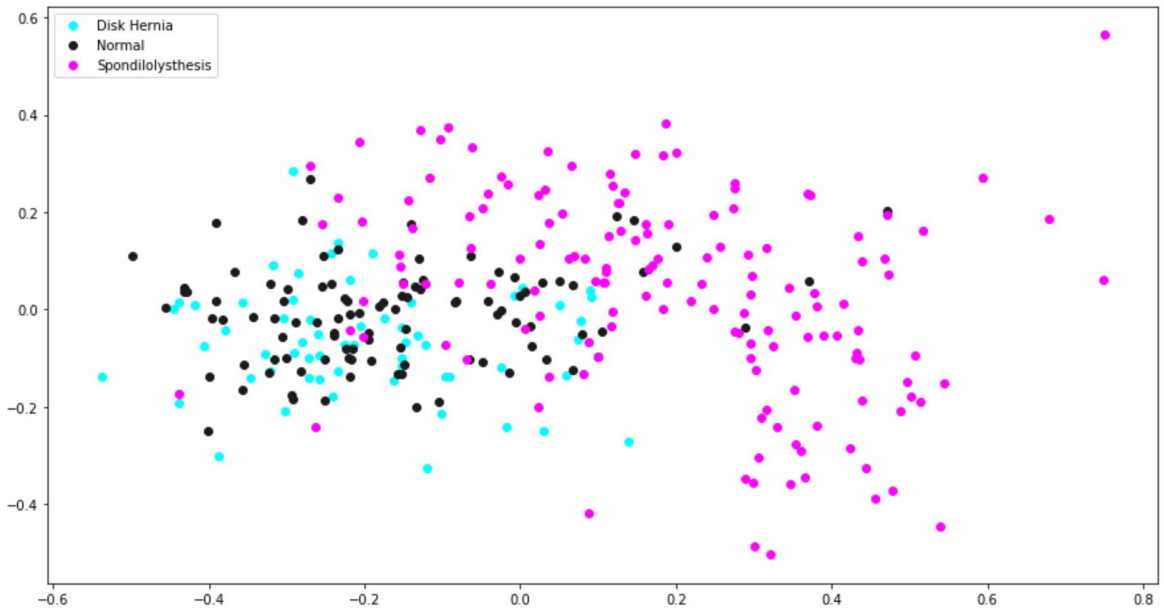
Random forests are built by combining the predictions of several trees, each of which is trained in isolation. Unlike in boosting (Schapire & Freund, 2012) where the base models are trained and combined using a sophisticated weighting scheme, typically the trees are trained independently and the predictions of the trees are combined through averaging. There are three main choices to be made when constructing a random tree. These are the method for splitting the leafs, the type of predictor to use in each leaf, and the method for injecting randomness into the trees. Specifying a method for splitting leafs requires selecting the shapes of candidate splits as well as a method for evaluating the quality of each candidate. Typical choices here are to use axis aligned splits, where data are routed to subtrees depending on whether or not they exceed a threshold value in a chosen dimension; or linear splits, where a linear combination of features are thresholded to make a decision. The threshold value in either case can be chosen randomly or by optimizing a function of the data in the leafs. In order to split a leaf, a collection of candidate splits are generated and a criterion is evaluated to choose between them. A simple strategy is to choose among the candidates uniformly at random, as in the models analyzed in Biau et al. (2008). A more common approach is to choose the candidate split which optimizes a purity function over the leafs that would be created. The most common choice for predictors in each leaf is to use the average response over the training points which fall in that leaf. Criminisi et al. (2011) explore the use of several different leaf predictors for regression and other tasks, but these generalizations are beyond the scope of this paper. We consider only simple averaging predictors here. Injecting randomness into the tree construction can happen in many ways. The choice of which dimensions to use as split candidates at each leaf can be randomized, as well as the choice of coefficients for random combinations of features. In either case, thresholds can be chosen either randomly or by optimization over some or all of the data in the leaf.

DATASET ANALATICS

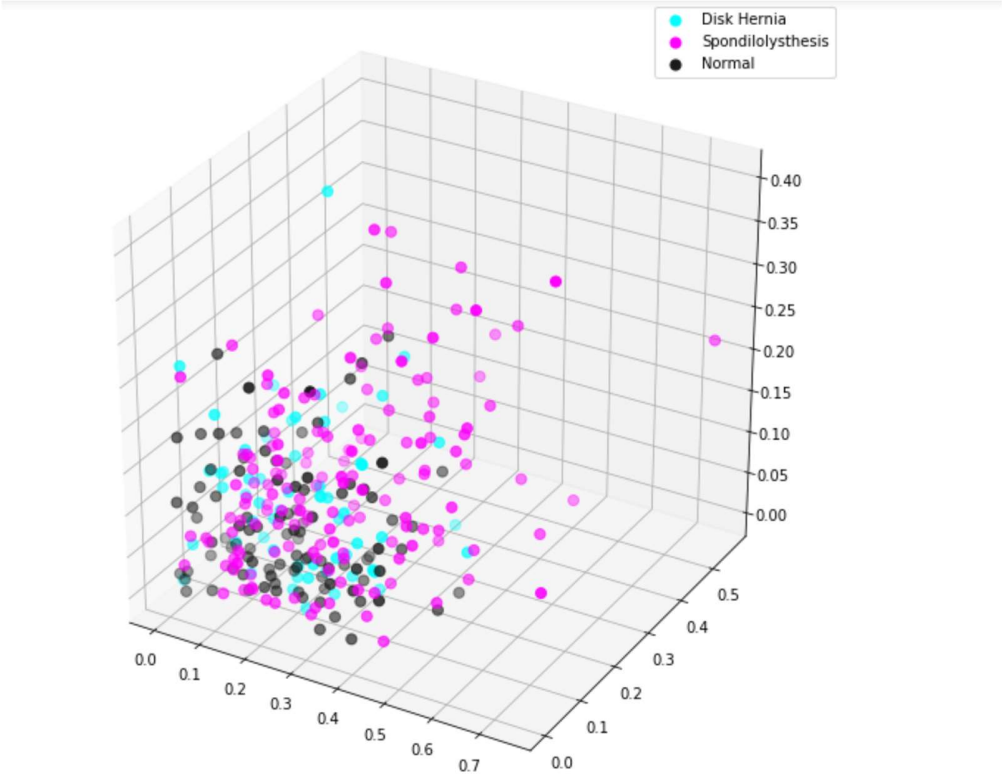
A: Cross-Sections



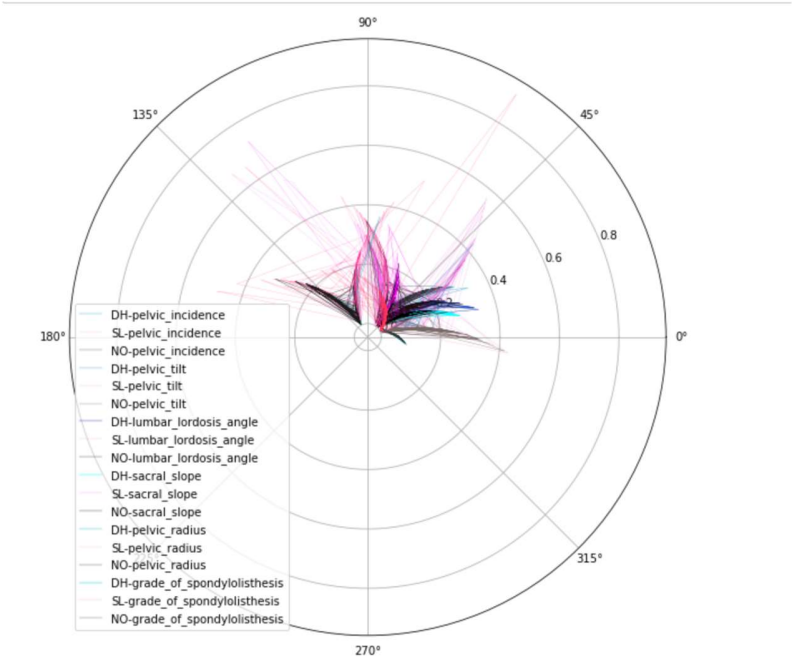
B: PCA (2-components) Plot



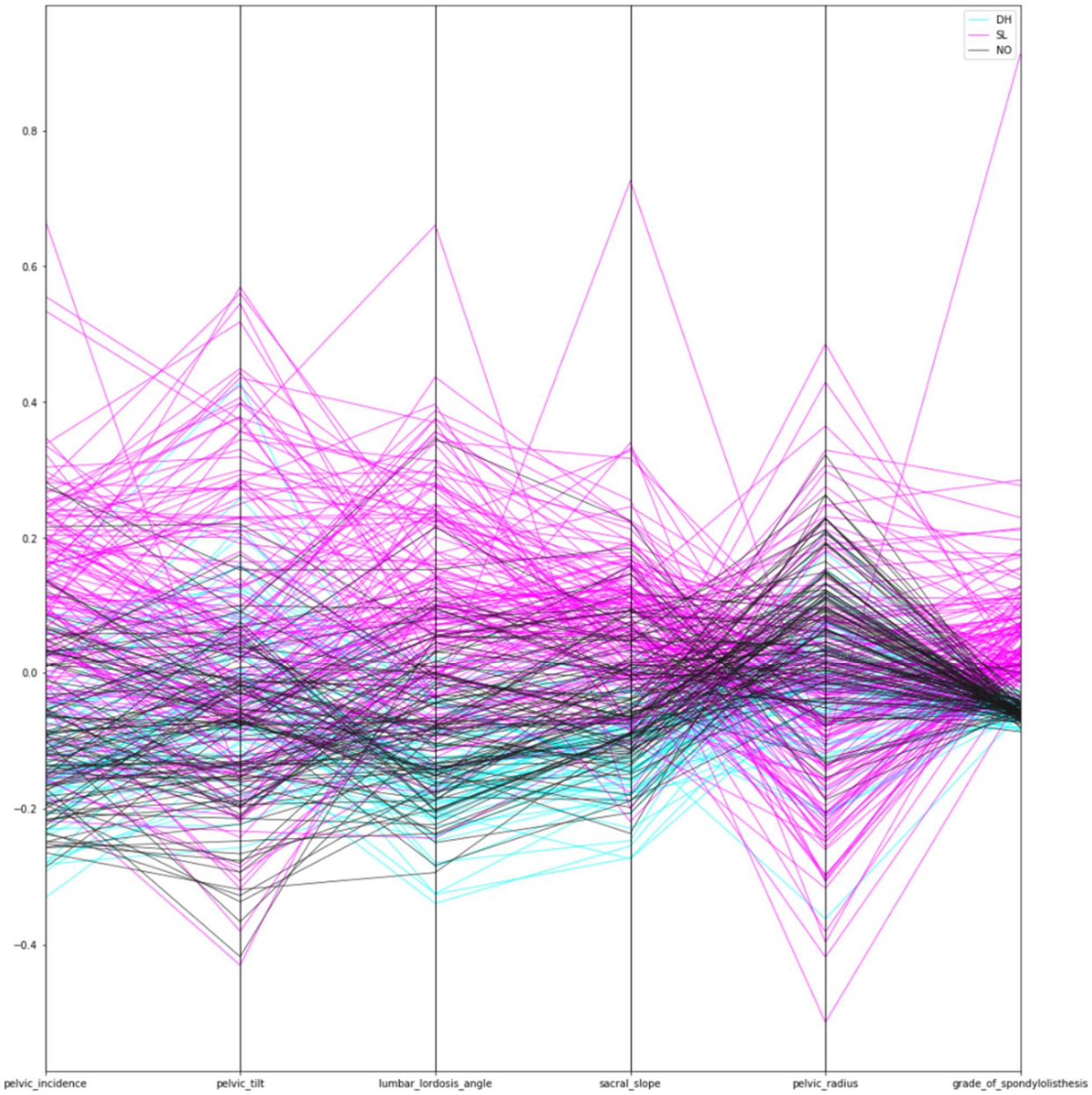
C: PCA (3-components) Plot:



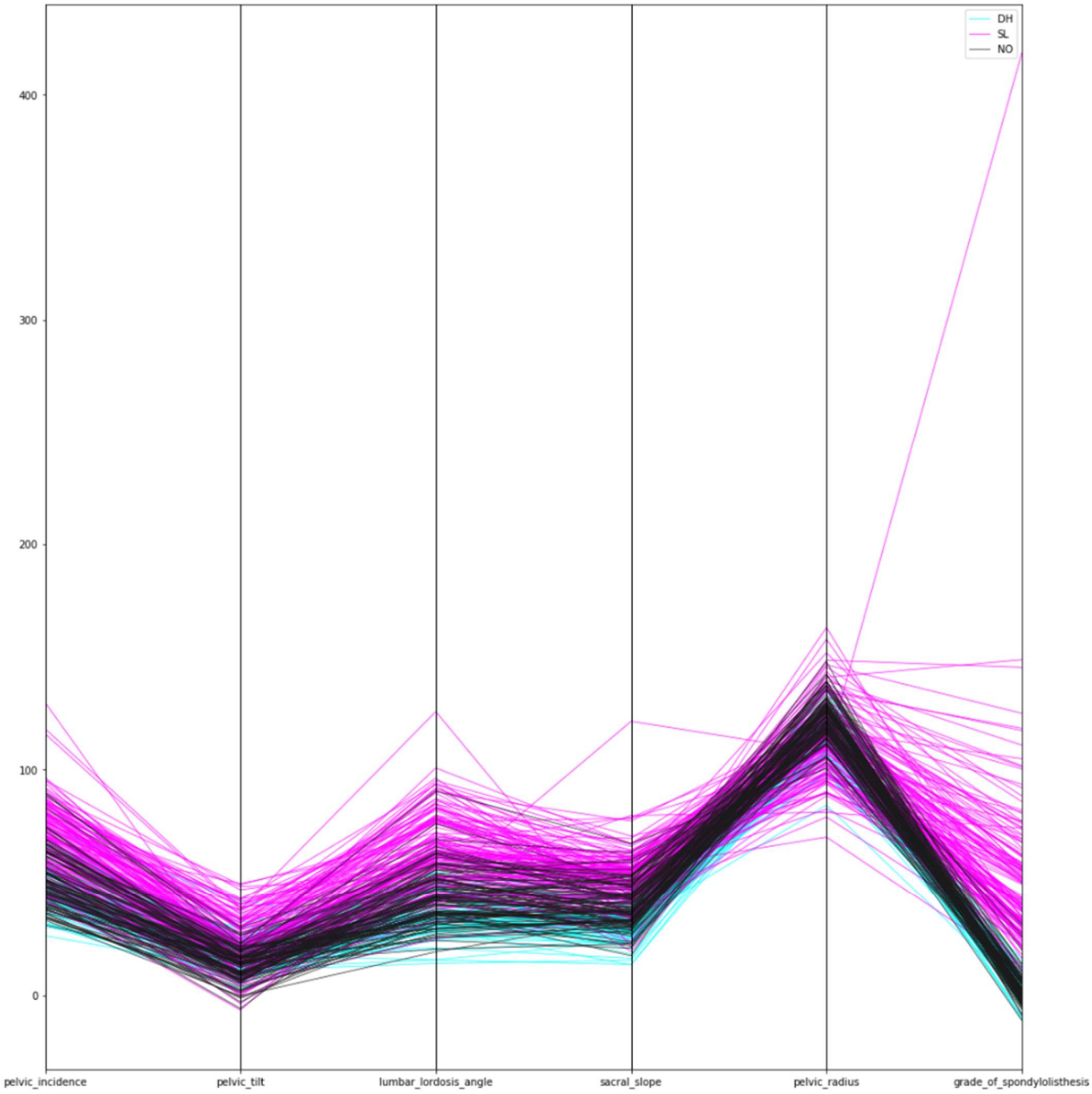
D: Polar Representation:



E: Parallel Coordinates (Normalized):



F: Parallel Coordinates:



ALGORITHM IMPLEMENTATION