Prediction of Spinal Abnormalities using Machine Learning Techniques

Azian Azamimi Abdullah
School of Mechatronic Engineering
Universiti Malaysia Perlis
Arau, Perlis, Malaysia
azamimi@unimap.edu.my

Atieqah Yaakob
School of Mechatronic Engineering
Universiti Malaysia Perlis
Arau, Perlis, Malaysia
atieqahyaacobrk85@gmail.com

Zunaidi Ibrahim
Technopreneur At UniMAP Sdn Bhd (TUSB), Taman Mutiara, Kangar, Perlis, Malaysia
zunaidi@unimap.edu.my

***Abstract*— Lower back pain can be caused by many complications with any parts of the body in the lumbar spine. The compilation of a medical diagnosis is crucial to the medical practitioners in order for them to give a convenient treatment for the low back pain. The machine learning models that applied in the medical field for disease diagnosis assists medical experts in the diseases identification based on the symptoms at an early stage. This research aims to identify the most significant physical parameters that contribute to spinal abnormalities and also predict spinal abnormalities based on collected physical spine data by using unsupervised machine learning approaches such as Principal Component Analysis (PCA), and also using supervised machine learning approaches such as K-Nearest Neighbors (KNN) and Random Forest (RF). As a result, degree spondylolisthesis is the most significant parameter that contributes to spinal abnormalities. As a comparison of results between RF classifier and KNN classifier, KNN classifier performed better than RF classifier since the percentage of accuracy of KNN algorithm (85.32%) are higher compared to RF classifier (79.57%).**

Keywords*— low back pain; machine learning; K-Nearest Neighbors (KNN); Random Forest (RF); spondylolisthesis.

I. INTRODUCTION

The main function of the vertebral column is to protect the spinal cord [1]. One of the vertebral column segments, lumbar vertebrae functions to bear the body weight. The lower the vertebra is in the low back, the greater the weight it must endure. Since the five lumbar vertebrae (L1-L5) are the largest vertebrae in the vertebral column, it makes possible for them to support the weight of the upper part of human body. The low back is known-well as structures that connect the bones, joints, nerves, ligaments, and muscles; all operating together to provide body support, body strength, and body flexibility. However, this complicated structure also makes the lumbar spine easily persuade to injury and pain.

Lower back pain can be caused by many complications with any parts of the body in the lumbar spine [2]. Lower back pain can be caused by an irritation or problem in the spine structures. An uncomplicated muscle strain at the low back might be one of the occasions for an emergency room visit, while a degenerating disc might generate only slight most likely distress to the patient.

In order to examine soft tissues or disc damage on the spinal column, computed tomography (CT) scans or magnetic resonance imaging (MRI) might be required to identify the source of pain and affirm the symptoms [3]–[5]. For instance, if a disc problem is suspected, it may need an imaging test that can give a detailed image by showing the location and size of the herniated disc and influenced nerve roots.

Since there are a lot of the imaging studies on detecting low back problems, this research is aimed to use collected physical spine data to identify a person, whether has spinal abnormalities or not. From the collected physical spine data, there are 12 features that measured which are pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, degree spondylolisthesis, pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle and scoliosis slope. The findings may be used as initial steps towards an automatic discrimination between normal and abnormal spines, which may assist practitioners in the clinical treatment of lower back pain. The compilation of a medical diagnosis is crucial to the clinician in order for them to give a convenient treatment for the pain. Therefore, by using machine learning approach and a set of patient's data, we able to tackle the problems.

The aim of this research is to identify a person whether has spinal abnormalities or not by using collected physical spine data. In order to fulfill the aim, the objectives have been set which are to identify the most significant physical parameters that contribute to spinal abnormalities and second, predict spinal abnormalities based on collected physical spine data by using unsupervised machine learning approaches such as Principal Component Analysis (PCA) and Random Forest (RF), also using supervised machine learning approaches such as K-Nearest Neighbors (KNN) and RF.

II. METHODOLOGY

The flow or process that would be used throughout this research in order to achieve the objectives is shown in Fig. 1.

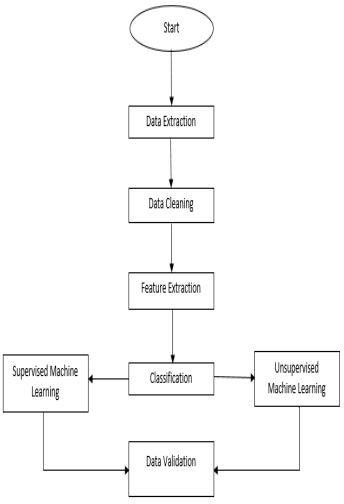

Fig. 1. Research flowchart.

978-1-5386-8369-9/18/$31.00 ©2018 IEEE

A. Data Extraction

The dataset that used in this research was extracted from Kaggle website [6]. This dataset has 310 patients' records which classify two classes. All the attributes are in the numerical attribute. The dataset contains the classification of the patients classified into one of two categories: Normal (100 patients) or Abnormal (210 patients). Every patient is represented as a pattern with 12 biomechanical attributes, according to the following physical parameters: pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, degree spondylolisthesis, pelvic slope, direct tilt, thoracic slope, cervical tilt, sacrum angle and scoliosis slope.

B. Data Cleaning

By using R language programming, the first procedure or technique that been used throughout this research after extracting the dataset is data cleaning [7]. One of the methods of data cleaning is by replacing all the attributes that only filled with numeric indicators, with the name of features. In this section, any information or words, which is not used for feature extraction and classification, have been varnished.

C. Feature Extraction

PCA is a statistical method used to reduce the number of variables in a dataset [8]. PCA was used to observe a variable covariance and variable correlation matrices in datasets. PCA also requires a pre-processing process in order to enter in the variability dimension and avoid scale effects, respectively.

The data that used PCA algorithm would be implemented in a graph or plot and shows the relationship between first and second principal component of the dataset. The first principal component PC1 has the highest variance across data. The second principal component PC2 is uncorrelated with PC1 which also has high variance.

To identify the most important features, the p-value is used to signify that the data is statistically significant and under that statistical data which able to identify the physical parameters or features that can discriminate between normal and abnormal spines. The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected, which means if p-value is small (typically lower than 0.05), the parameters are said to be significant.

After getting results of p-value for each parameter, the parameters that have the p-value > 0.05 would not be used as they are not significant. The remaining parameters will be used with RF classifier in order identify the most important features.

D. Classification using supervised machine learning

Random forest classifier as supervised machine learning is used to predict and classify the data into normal and abnormal spines as an outcome [9].

There are two stages in RF algorithm, one is random forest creation, and the other is to make a prediction from the random forest classifier created in the first stage. The whole process is shown below, and it's easy to understand using the figure. Here shows the Random Forest creation pseudocode:

1. Randomly select "K" features from total "m" features where k << m
2. Among the "K" features, calculate the node "d" using the best split point
3. Split the node into daughter nodes using the best split
4. Repeat 1st to 4th steps until "l" number of nodes has been reached
5. Build forest by repeating steps 1 to 4 for "n" number times to create "n" number of trees

In the next stage, with the random forest classifier created, we will make the prediction. The random forest prediction pseudocode is shown below:

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
2. Calculate the votes for each predicted target
3. Consider the high voted predicted target as the final prediction from the random forest algorithm

Besides that, we also using K nearest neighbor classifier to classify and predict the spinal abnormalities. KNN is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure, for instance, distance functions [10]. The data is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function. If k = 1, then the case is simply assigned to the class of its nearest neighbor. We can implement a KNN model by following the below steps:

1. Load the data
2. Initialise the value of k
3. In order to get the predicted class, iterate from 1 to total number of training data points
    1. Calculate the distance between test data and each row of training data. Here we will use Euclidean distance as our distance metric since it's the most popular method..
    2. Sort the calculated distances in ascending order based on distance values
    3. Get top k rows from the sorted array
    4. Get the most frequent class of these rows
    5. Return the predicted class

Besides that, we also implement the weighted KNN algorithm in order to predict and classify the data into normal and abnormal spines in this research. Choosing the number of nearest neighbors, which means determine the value of *k* plays a crucial role in determining the efficacy of the model. A high *k*-value has an advantage which includes reducing the variance due to the noisy data.

Based on the results of test samples that illustrated using true positive (TP), true negative (TN), false positive (FP) and false negative (FN) from the confusion matrix, the following quality of the classification models can be evaluated and calculated by using the following equations:

1. Accuracy

$$\%\ Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)}\ x\ 100 \quad (1)$$

It determines the percentage of true results of the test samples.

2. Precision or Positive Predictive Value

$$\%\ Precision = \frac{TP}{TP+FP}\ x\ 100 \quad (2)$$

It measures the numbers of predicted abnormal patients among positive results.

3. Sensitivity or True Positive Rate

$$\%\ Sensitivity = \frac{TP}{TP+FN}\ x\ 100 \quad (3)$$

It determines the ability of the classifier to classify abnormal patients correctly.

4. Specificity or True Negative Rate

$$\%\ Specificity = \frac{TN}{TN+FP}\ x\ 100 \quad (4)$$

It measures the predicted normal patients among actual numbers of normal patients.

E. Data Validation

After the comparison of RF and KNN classifier been made, we will use the classifier that has the highest accuracy for data validation. By using different dataset, the model that has the highest accuracy will be used to predict the performance of the other dataset. The different dataset will have some similar steps; data extraction, data cleaning, and classification.

The data set used for classifying vertebral column disorders is taken from the UCI machine learning database [11]. The dataset has 310 patients' records which also classify two classes: normal and abnormal spine.

III. RESULTS AND DISCUSSION

A. Data Cleaning

Figure 2 shows the result after data cleaning as the last column that contained some information but not in numerical have been removed during data cleaning process as shown in Fig. 3.

```
  pelvic_incidence pelvic_tilt lumbar_lordosis_angle sacral_slope
1       63.0278175 22.55258597           39.60911701  40.47523153
2       39.05695098 10.06099147           25.01537822  28.99595951
3       68.83202098 22.21848205           50.09219357  46.61353893
4       69.29700807 24.65287791           44.31123813  44.64413017
5       49.71285934 9.652074879             28.317406  40.06078446
6       40.25019968 13.92190658            25.1249496  26.32829311
  pelvic_radius degree_spondylolisthesis pelvic_slope Direct_tilt
1   98.67291675             -0.254399986  0.744503464     12.5661
2   114.4054254              4.564258645  0.415185678     12.8874
3   105.9851355             -3.530317314  0.474889164     26.8343
4   101.8684951              11.21152344  0.369345264     23.5603
5   108.1687249              7.918500615  0.543360472      35.494
6   130.3278713              2.230651729  0.789992856      29.323
  thoracic_slope cervical_tilt sacrum_angle scoliosis_slope predicted_class
1        14.5386       15.30468    -28.658501         43.5123        Abnormal
2        17.5323       16.78486    -25.530607         16.1102        Abnormal
3        17.4861       16.65897    -29.031888         19.2221        Abnormal
4        12.7074       11.42447    -30.470246         18.8329        Abnormal
5        15.9546        8.87237    -16.378376         24.9171        Abnormal
6        12.0036       10.40462     -1.512209          9.6548        Abnormal
                                                              NA
1
2
3 Prediction is done by using binary classification.
4
5
6        Attribute1  = pelvic_incidence  (numeric)
```

Fig. 2. The dataset before data cleaning process.

```
  pelvic_incidence pelvic_tilt lumbar_lordosis_angle sacral_slope
1         63.02782   22.552586              39.60912     40.47523
2         39.05695   10.060991              25.01538     28.99596
3         68.83202   22.218482              50.09219     46.61354
4         69.29701   24.652878              44.31124     44.64413
5         49.71286    9.652075              28.31741     40.06078
6         40.25020   13.921907              25.12495     26.32829
  pelvic_radius degree_spondylolisthesis pelvic_slope Direct_tilt
1      98.67292               -0.254400    0.7445035     12.5661
2     114.40543                4.564259    0.4151857     12.8874
3     105.98514               -3.530317    0.4748892     26.8343
4     101.86850               11.211523    0.3693453     23.5603
5     108.16872                7.918501    0.5433605     35.4940
6     130.32787                2.230652    0.7899929     29.3230
  thoracic_slope cervical_tilt sacrum_angle scoliosis_slope predicted_class
1        14.5386       15.30468    -28.658501         43.5123        Abnormal
2        17.5323       16.78486    -25.530607         16.1102        Abnormal
3        17.4861       16.65897    -29.031888         19.2221        Abnormal
4        12.7074       11.42447    -30.470246         18.8329        Abnormal
5        15.9546        8.87237    -16.378376         24.9171        Abnormal
6        12.0036       10.40462     -1.512209          9.6548        Abnormal
```

Fig. 3. The dataset after data cleaning process.

B. Feature Extraction

As unsupervised learning which PCA method used on the data, the result has been interpreted into a plot as shown in Fig. 4. From the plot, there are clearly distinguished between normal patients and abnormal patients as normal in black dot and abnormal in red dot.

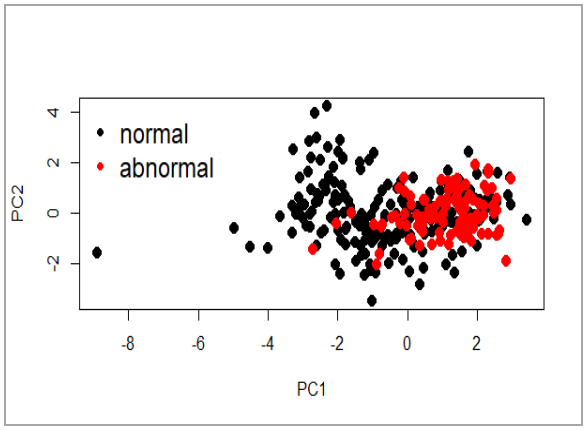

Fig. 4. Principle Component Analysis (PCA) plot.

C. Statistical Analysis

Since t-test has been used to find the p-value, the p-value is used to find the most important parameters or features. Table 1 shows the p-value of each feature after using t-test.

From the table, the degree of spondylolisthesis has the smallest p-value followed by pelvic incidence, pelvic tilt, lumbar lordosis angle, pelvic radius and sacral slope that have *p*-value < 0.05. Meanwhile, pelvic slope, direct tilt, thoracic slope, sacrum angle and scoliosis angle have similar p-value which is 1.0. Features that have small p-value which p-value is less than 0.05 indicates that the features or physical parameters are significant. The parameters that have *p*-value more than 0.05 would not be used in the classifier.

TABLE I. THE P-VALUES OF EACH FEATURES

Features	*p*-value
Degree Spondylolisthesis	3.589907×10^{-26}
Pelvic Incidence	1.115125×10^{-11}
Pelvic Tilt	1.679709×10^{-10}
Lumbar Lordosis Angle	$7.6544428 \times 10^{-10}$
Pelvic Radius	1.233155×10^{-9}
Sacral Slope	1.362633×10^{-4}
Cervical Tilt	4.993720×10^{-1}
Pelvic Slope	1.000000×10^{0}
Direct Tilt	1.000000×10^{0}
Thoracic Slope	1.000000×10^{0}
Sacrum Angle	1.000000×10^{0}
Scoliosis Slope	1.000000×10^{0}

Figure 5 shows box plot of degree spondylolisthesis of abnormal and normal person. From the boxplot, the degrees of spondylolisthesis of abnormal patients are bigger or higher compared to the normal patients.

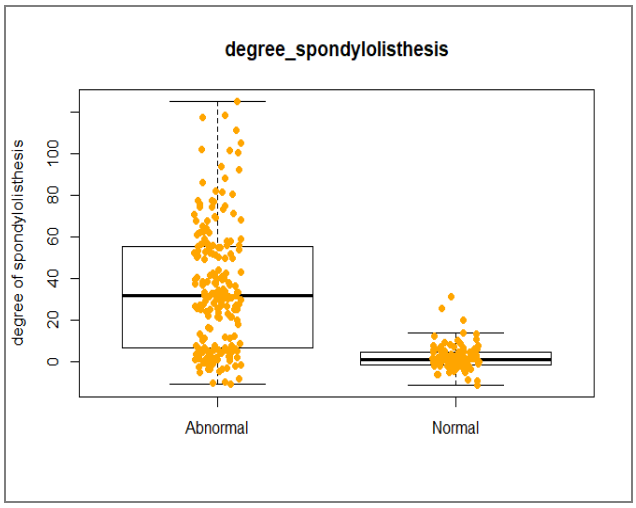

Fig. 5. Boxplot of degree spondylolisthesis of abnormal and normal spine.

D. Random Forest Classifier

Random forest (RF) classifier is used to identify the most significant features that contribute to spinal abnormalities, as shown in Fig. 6. Not surprisingly, degree spondylolisthesis still in the top list of features that could lead to spinal abnormalities to the patients following with pelvic radius and pelvic incidence.

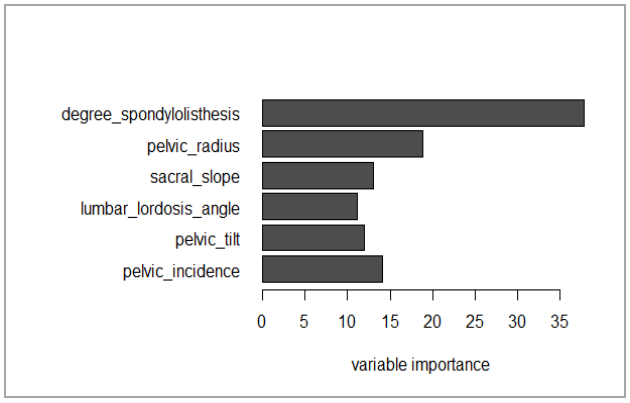

Fig. 6. The most significant features obtained by random forest (RF) classifier.

Next, the dataset have split up into two; 80% train data, 20% test data and 70% train data, 30% test data. RF classifier is used to classify the spine into normal or abnormal spine. After the data is trained, the test data is used to predict the results. The results show the accuracy is higher with 30% test data rather than with 20% test data as shown in Table 2.

TABLE II. THE RESULT OF ACCURACY, SENSITIVITY, SPECIFICITY, AND PRECISION AFTER USING 20% AND 30% OF THE DATA AS TEST DATA IN RANDOM FOREST CLASSIFIER

Test data	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)
20%	79.03	84.62	72.73	82.50
30%	**79.57**	87.50	78.13	80.33

E. K-Nearest Neighbors Classifier

Same with RF classifier, the dataset have split up into two; 80% train data, 20% test data and 70% train data, 30% test data. We implemented three types of KNN algorithm, which are KNN; weighted triangular and weighted rectangular to predict the results. In the end, we will compare the results between KNN and weighted KNN. Table 3 shows the result of using 20 % of data as test data while Table 4 shows the result of using 30 % of data as test data.

TABLE III. KNN RESULTS (20% TEST DATA) WHERE A IS ACCURACY, B IS SENSITIVITY, C IS SPECIFICITY AND D IS PRECISION

Value of *k*	KNN(%)	Triangular weighted KNN (%)	Rectangular weighted KNN (%)
1	A=79.84 B=84.05 C=71.00 D=86.02	A=82.90 B=87.86 C=72.50 D=87.17	A=82.90 B=87.86 C=72.50 D=87.17
2	A=79.68 B=83.10 C=72.50 D=86.65	A=81.78 B=87.15 C=70.50 D=86.19	A=81.78 B=95.27 C=70.50 D=86.19
3	A=84.36 B=86.19 C=80.50 D=90.43	A=81.62 B=86.67 C=71.00 D=86.47	A=82.74 B=89.05 C=69.50 D=86.18
4	A=78.07 B=80.00 C=74.00 D=86.75	A=81.94 B=88.33 C=68.50 D=85.54	A=82.90 B=90.48 C=67.00 D=85.28
5	A=81.13 B=85.00 C=73.00 D=86.96	A=81.94 B=86.43 C=72.50 D=87.02	A=82.91 B=88.81 C=70.50 D=86.45
6	A=84.03 B=87.38 C=77.00 D=89.10	A=82.26 B=88.10 C=70.00 D=86.20	A=83.07 B=89.29 C=70.00 D=86.26
7	**A=85.32** B=84.52 C=95.50 D=93.25	A=83.07 B=88.81 C=71.00 D=86.65	A=83.39 B=88.10 C=73.50 D=87.68
8	A=81.29 B=84.28 C=75.00 D=87.77	A=82.42 B=87.62 C=71.50 D=86.95	A=83.23 B=87.86 C=73.50 D=87.62
9	A=81.45 B=84.05 C=76.00 D=88.25	A=84.52 B=90.55 C=75.00 D=88.41	A=83.07 B=88.57 C=71.50 D=86.85
10	A=80.48 B=84.52 C=72.00 D=86.39	A=83.07 B=88.10 C=72.50 D=87.11	A=80.32 B=84.05 C=72.5 D=86.69

TABLE IV. KNN RESULTS (30% TEST DATA) WHERE A IS ACCURACY, B IS SENSITIVITY, C IS SPECIFICITY AND D IS PRECISION

Value of k	KNN (%)	Triangular kernel KNN (%)	Rectangular kernel KNN (%)
1	A=79.57 B=83.49 C=71.33 D=86.14	A=78.92 B=83.65 C=81.33 D=85.23	A=78.92 B=83.65 C=69.00 D=85.23
2	A=77.42 B=83.17 C=65.33 D=83.49	A=78.28 B=85.87 C=62.33 D=82.85	A=78.28 B=85.87 C=62.33 D=62.85
3	A=82.37 B=86.03 C=74.67 D=87.79	A=77.74 B=83.02 C=67.33 D=84.02	A=79.46 B=84.60 C=68.67 D=85.18
4	A=80.43 B=83.17 C=74.67 D=87.44	A=79.46 B=86.03 C=65.67 D=84.12	A=81.62 B=88.57 C=67.00 D=85.11
5	A=80.97 B=86.51 C=69.33 D=85.65	A=80.46 B=86.51 C=68.00 D=85.07	A=81.29 B=86.19 C=71.00 D=86.28
6	A=83.08 B=86.03 C=76.60 D=88.65	A=79.03 B=85.40 C=65.67 D=84.02	A=80.65 B=86.51 C=68.33 D=85.28
7	A=81.08 B=82.70 C=77.67 D=88.67	A=78.92 B=85.24 C=65.67 D=83.95	A=79.79 B=86.03 C=66.67 D=84.47
8	**A=83.67** B=86.98 C=77.00 D=88.87	A=80.00 B=85.56 C=68.33 D=85.20	A=80.22 B=85.56 C=69.00 D=85.51
9	A=83.12 B=87.14 C=74.67 D=88.05	A=79.46 B=85.24 C=67.33 D=93.19	A=80.54 B=86.51 C=68.67 D=85.04
10	A=80.75 B=83.18 C=75.67 D=88.03	A=80.65 B=85.72 C=70.00 D=85.80	A=81.51 B=84.60 C=75.00 D=87.75

The result of accuracy, sensitivity, specificity and precision are an average of ten times experiment. From Table 3 and Table 4, we can see that in any value of k, KNN have better results than in weighted KNN. Value of accuracy in 20% test data is higher than in 30% test data which in the value of k = 7 (85.32%).

F. Data Validation

For data validation, we use the performance model from KNN classifier since the KNN accuracy value is higher than RF accuracy value.

By using different dataset, we will evaluate the performance of KNN model in value of k = 7 of 20% test data. Table 5 shows the actual numbers of abnormal and normal patients by using 20% of test samples and the predicted numbers of abnormal and normal patients when

using k = 7 on the test samples after using the KNN classifier.

Table 6 indicates the percentage of accuracy; sensitivity, specificity and precision after the results have been repeated about 10 times.

TABLE V. ACTUAL NUMBER OF PATIENTS AND PREDICTED NUMBER OF PATIENTS AFTER USING KNN CLASSIFIER

Patient	Actual No.	Predicted No.
Abnormal	42	36
Normal	20	17

TABLE VI. THE PERCENTAGE OF ACCURACY, SENSITIVITY, SPECIFICITY AND PRECISION FOR DATA VALIDATION

Accuracy	Sensitivity	Specificity	Precision
86.13 %	90.24%	77.50%	89.48%

IV. CONCLUSIONS

As a conclusion, the most significant parameter that contributes to spinal abnormalities is degree spondylolisthesis. Spondylilosthesis is a common disease even though the symptoms is undetectable in early stage. Since the cause could happen on normal people in daily basis life while doing daily works, spondylolisthesis could happen to most people, regardless of their age. While it may not be possible to reverse the degenerative changes that occur with aging, it is possible to strengthen the muscles that surround the spine. With helps of physical therapist to stabilize the lumbar spine, it will often result in a decrease in symptoms of low back and leg pain to the point where surgery becomes unnecessary. This type of therapy must emphasize active rehabilitation, which means that the patient must work actively to strengthen the muscles of the abdomen, low back, and core. The type of therapy that we employ emphasizes core conditioning and strengthening and our therapists will instruct you on how to do these exercises properly. If the symptoms are relatively mild and are still able to exercise, hike, and play some sports, then often a Pilates or a Yoga program may be very beneficial, less costly, and more convenient than going to a physical therapist.

As a comparison of results between RF classifier and KNN classifier, we can see that KNN classifier performed better than RF classifier since the percentage of accuracy, sensitivity, specificity and precision of KNN algorithm (85.32%) are higher compared to RF classifier (79.57%). Last but not least, since KNN classifier has better accuracy than RF classifier, by using the developed model, data validation is done by using different dataset from UCI database. In data validation, the accuracy is quite high. It is proven that the model is good.

ACKNOWLEDGMENT

The authors are grateful for financial support provided by Universiti Malaysia Perlis (UniMAP) via research grant 9009-00053.

REFERENCES

[1] P. Brinckmann, "Pathology of the vertebral column," *Ergonomics*, vol. 28, no. 1, pp. 77–80, 1985.

[2] K. H. Allen R. Last, "Chronic lower back pain," *Am. Fam. Physician*, vol. 79, no. 12, pp. 1067–1074, 2009.

[3] Z. Ahmad, R. Mobasheri, T. Das, S. Vaidya, S. Mallik, M. El-Hussainy, and A. Casey, "How to interpret computed tomography of the lumbar spine," *Annals of the Royal College of Surgeons of England*, vol. 96, no. 7. pp. 502–507, 2014.

[4] B. Tins, "Technical aspects of CT imaging of the spine," *Insights Imaging*, vol. 1, no. 5–6, pp. 349–359, 2010.

[5] A. Liguori, F. Galli, M. Gurgitano, A. Borelli, M. Pandolfi, F. Caranci, A. M. Magenta Biasina, G. G. M. Pompili, C. L. Piccolo, V. Miele, C. Masciocchi, and G. Carrafiello, "Clinical and instrumental assessment of herniated discs after nucleoplasty: A preliminary study," *Acta Biomed.*, vol. 89, pp. 220–229, 2018.

[6] Kaggle Inc., "The Home of Data Science," *Kaggle is the world's largest community of data scientists*, 2014. [Online]. Available: http://www.kaggle.com.

[7] J. Monogan, "An Introduction to R," *User Man.*, no. October, pp. 1–12, 2009.

[8] H. Abdi and L. J. Williams, "Principle component analysis," *Wiley Interdiscip. Rev. Comput. Stat.*, vol. 2, no. 4, pp. 433–459, 2010.

[9] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[10] M. Steinbach and P.-N. Tan, "kNN: k-Nearest Neighbors," in *The Top Ten Algorithms in Data Mining*, 2009, p. 208.

[11] K. Bache and M. Lichman, "UCI Machine Learning Repository," *University of California Irvine School of Information*, vol. 2008, no. 14/8. p. 0, 2013.