

EFFECTIVE PREDICTION OF VERTEBRAL COLUMN PATHOLOGIES USING RANDOM FOREST

K.N. Nithya

Assistant Professor, Department of Computer Science,
Shri Sakthi Kailassh Women's College, Salem, Tamil Nadu, India
knnithya.10@gmail.com

P. Suresh

Head, Department of Computer Science,
Salem Sowdeswari College, Salem, Tamil Nadu, India
sur_bhoo71@rediffmail.com

Abstract - The advent of spine research has gradually improvised using Artificial Intelligence and Machine learning techniques. Data mining had a greater influence on minimizing the complexities of medical diagnosis. The most concentrated aspects of medical diagnosis are feature selection and classifications. Accurate classification minimizes human intervention especially in handling most sensitive cases like vertebral column disorders. In the medical domain, the major controversy is distinguished from healthy and unhealthy spines. There were several existing methods are available for classifying the vertebral column disorders. But still, the need for advancement has grabbed attention among the research communities. In this paper, we proposed a random forest method for achieving enhanced classification results. The proposed mechanism consists of two phases and the first phases consist of random forest creation using N decision trees. The second phase is accurate prediction using the obtained tree structure from the first phase. The proposed method of efficiency is calculated based on accuracy, specificity, sensitivity, and F-measure. The experimental work is done using UCI medical dataset on MATLAB and WEKA tools. The observation is conducted between SPRINT [12], Ensemble Classifiers [11] with our proposed RF model. From the observation, it is proved the classification accuracy is quite improved with the implementation of our proposed random forest mechanism. The prediction achieved by random forest is 98 % which is more efficient in comparison to others especially on the terms of high accuracy and classifying speed.

Keywords: Data Mining, classification, Medical diagnosis, Vertebral column pathologies, and ethical implications.

1. Introduction

Smart healthcare system, is the major need for most of the growing nations. Once a country's health and wealth are reflected by its computing technology with major fields like business, healthcare, etc. On considering medical diagnosis, in recent years there is a vast contribution towards its research areas. Especially in dealing with critical and major diseases. The evolution of technology in medical fields is significant this is because of accuracy and time consumption. Even an expert physician required more time to confirm the diagnosis but that also not finalize, this lack can be overcome using technology. In this way, data mining plays a vital role in the fields of the medical domain. In the human race there day-by-day several diseases are evolving among which low back pain (LBP). Nowadays LBP becomes a common disease among all ages from younger to aged [1]. A survey report states about 85% of peoples have suffered from LBP in most of the major countries [2, 3]. The low back pain caused because of the vertebral column pathologies. The vertebral column is the major part of the human body for movement, support, and protections. It consists of a series of 33 bones enclosed with the thoracic curve, thoracic curve, sacral curve, cervical curve and lumbar curve separated by intervertebral discs [4]. Determining the pathologies from a normal human, a physician needs to examine the pattern signs and symptoms [5]. In the industry, several computerized learning techniques are discussing spinal challenges and abnormalities. Among which the known technologies are Bayesian classifier with Gibbs distribution, Support Vector Machine (SVM), Artificial Neural Network (ANN), k-means for valuing the degree of disc space, biomechanical measures effects, and orientation for identifying normal and pathologic conditions [6, 7, and 8]. An automatic spine analysis can be possible by effective vertebrae segmentation and identification [9]. This segmentation and identification are done with several spine images that require a proper vertebra segmentation algorithm. The disc hernia and spondylolisthesis are the popularly known vertebral disorders. The lumbar

region is the major part where vertebral column pathologies are present. For detecting those pathologies MRI or radiology images are the best sources [10].

This paper is organized as follows, section 1 consists of an introduction about the verbal column and other aspects of spinal disorders and technology evolved for the identification. In section 2 several related works representing the classifications on the aspects of separating normal, hernia and spondylolisthesis cases. In section problem statements and our contributions were also discussed. Section 3 consists of the proposed works and their working mechanism. In section 4 experimental work is discussed and finally, section 5 consists of the conclusion part.

2. Related work

K.N. Nithya et al [11] proposed the Prediction of Vertebral Column Pathologies Using Ensemble Classifiers. This work is the extension of her previous work with the SPRINT algorithm a single classifier. The addressed the possibility of poor selection using the single classifier and develops multiple-classifier techniques with multiple voting model known as Ensemble classifier. This work is evaluated based on the similarity features with MATLAB and WEKA tools using UCI medical dataset. The outcomes of the results prove its accurate efficiency with multiple-classifier techniques of overall 85.5%.

K.N. Nithya et al [12] developed a Novel Sprint Algorithm for diagnosing Vertebral Column Disorders. The author distinguishes a healthy and unhealthy spine using the Sprint algorithm. SPRINT algorithm works in the principle of Decision Tree which considers lumbar and sacral parameters for achieving better classification results. The experimental work is carried out on the WEKA tool with three sets of datasets. The results obtained by the SPRINT algorithm are much better in comparison to KNN and rep-tree methods.

Shaiful Islam et al [13] describe the feature selections and classification of spinal abnormalities using machine learning methods. The main contribution is towards detecting the low back pain disorders from the abnormality data. The experiment is carried out in two ways such as with or without feature filtering. The main motto of this work identifying of most influential features that overcome the misleading effect from the given input data. The proposed mechanism with traditional random forest achieves 94% accuracy comparing to others.

Vijayalakshmi et al [14] investigated the Vertebral Column Pathologies diagnosis with kNN classifiers. The main reason for this implementation is improvising the pattern recognition. By which the accurate identification of pathologies effectively declares either the data has disc hernia or Spondylolisthesis, etc. The author used a dataset from the UCI machine learning repository and achieves about 88.31% of successful rates compared to other classifiers.

Fabio Galbusera et al [15] discussed various research contributions of Artificial intelligence and machine learning in spine research. It encloses several tools and novel approaches evolved in identifying the image segmentation, prediction of clinical outcomes and radiological images. In this review, the author describes several decision support systems, computer-aided diagnosis, and complications. Besides, the author submitted an accountable work about data privacy and security with all related issues.

Sana Ansari et al., [16] developed diagnosis and vertebral column pathologies detection with machine learning classifiers. It is a combined approach of generalized regression neural network, support vector machine and feeds forward neural network. The experimental work is conducted with 10 fold cross-validation and holdout methods applying various kernel and activation functions. The experimental work dataset is taken from the UCI machine learning repository and accrued the classification accuracy of 88.56% compared to others.

Yavuzunal and ErdincKocer et al [17] proposed the identification of vertebral column disorders using the combined approach of backpropagation neural network and naive Bayesian classifiers. The author conducted the observation over the input dataset taken from the UCI machine learning repository and achieves the average classification accuracy of about 84.35% on both classifiers.

2.1. Problem Statements

- The need for improvement in the range of accuracy
- Improvisation on handling a multi-range of the dataset with complexities
- Need to have enhanced classification method for handling of noisy data
- Clear segmentation on the important parameters such as selection features
- Exact determination on dealing with spine curvature
- Leverage due to complexities and increase of time consumptions

2.2. Our Contributions

Our Contributions The previous contribution consists of the evolution of two mechanisms in the state of improving classification accuracy. They are the SPRINT algorithm and the Ensemble classifier. The SPRINT algorithm is a single way classifier which has the possibility of poor selection. The Ensemble classifier is a multiple-classifier technique that achieves better results than the SPRINT algorithm. As the Ensemble classifier

is a multiple-classifier technique there is an increase of time in prediction. Even though it has good accuracy, on the research basis we found the implementation of an enhanced random forest algorithm will give better accuracy results than the others.

2.3. Reason for implementing Ensemble methods

- It is advanced from tradition random forest algorithm [13]
- Two-stage performance such as N decision tree and prediction from the decision tree
- Applying bootstrap random sampling method for retrieving the K training sets
- Applicable for handling huge datasets
- Consist of a simple voting system
- Forest decision tree structure consists of maximum prediction
- Majority with maximum prediction achieves higher range accuracy

2.4 Proposed Methodology

The proposed Random Forest model is derived from the traditional random forest (RF) algorithm. Initially, in 2001 it was proposed by Breiman [18] as a combination algorithm. By which the prediction is based on its random forest regression, continuous value, and random forest classification. Most of the work proved random forest algorithm results with higher accuracy prediction avoiding the noise and abnormal values. The working mechanism of the RF classification algorithm consists of two phases, under which in the first phase the subsamples are extracted from original samples employing the bootstrap resampling method. Based on that for every sample, a decision tree is created. Secondly, the classified decision tree enables the voting system and the decision tree which gets a higher vote taken as the final prediction result.

The working of RF algorithm is stated by three steps such as;

(1) Select the training set: The bootstrap random sampling method is employed for retrieving the K training sets from its original dataset (M properties). As per the original training set, the training set is taken along with its size.

(2) Build the RF model: For all the bootstrap training sets a classification regression tree is created resulting in K decision trees forming as a “forest” and the trees in the forest are not altered. Each tree structure does not resemble with best features due to internal nodes. In this instance, branching is done by the random selection of $m \leq M$ of all features.

(3) Create simple voting: The decision tree’s training process is independent of each other. The training of random forests works parallel with enhancing its efficiency. The trained K decision trees are combined in the same way for building the RF. Once the classifications are done on the input samples then the simple voting system is enabled on each decision tree for deriving the output. The independent sample determination, distributed decision trees, and each decision tree’s final categorization are determined by the RF algorithm.

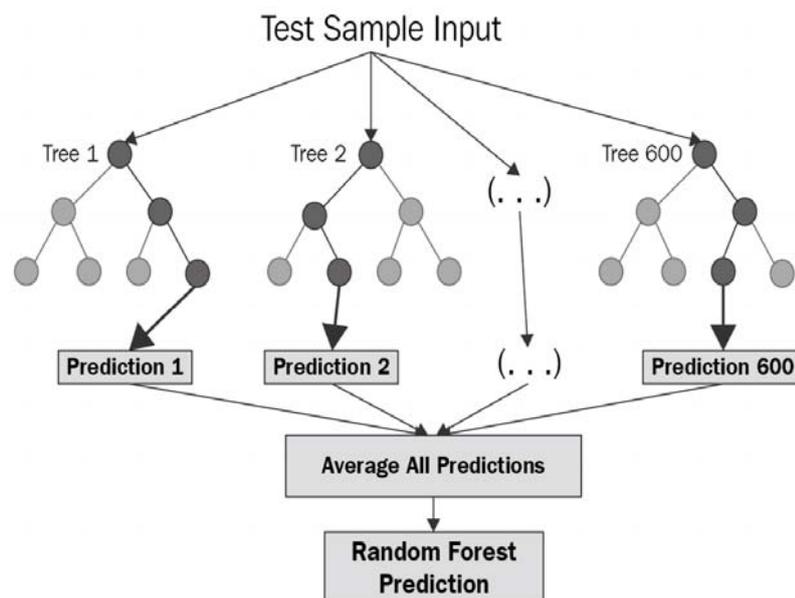


Fig 1 random forest structure

The proposed Random Forest model working mechanism consists of two-phase. In the first phase, the random forest is created by combining the N decision tree. Then the second phase deals with predictions at each created tree from the first phase.

The Random forest structure as on the above figure and its working process are explained as follows: Step-1: Random K data points are selected from the training set.

Step-2: The decision trees are built using the associated selected data points known as subsets.

Step-3: The number N for decision trees needs to build is chosen.

Step-4: Repeat Step 1 & 2.

Step-5: Next the predictions on each decision tree.

The new data points are categorized using a simple voting system the data point with multiple votes is taken as final. The below example describes the working of the RF algorithm simply; Example: Consider that dataset consists of the number of fruit images and this dataset is taken to the Random forest classifier. Initially, the dataset is segmented into subsets and several decision trees are created. Under the training phase, every decision tree has a prediction result at the stage of a new data point occurs. Based on which the Random Forest classifier predicts the final decision with data point having maximum votes.

2.5 Experimental results

2.5.1 Dataset

for conducting the observation datasets are taken from UCI. UCI is one of the known resources among the research community which consists of an ample number of medical datasets and data generators. Most of the existing work discussed in the related work section has declared the same source for conducting their experiments. A statement is proved UCI is peculiar for determining the medical research which contributes to the needed diagnostic parameters among the students, educators, and research scholars. For our research, a dataset is taken with a combination of 150 spondylolisthesis disorders, 100 healthy subjects, and 60 disk hernia disorders. Each dataset consists of six attributes known as pelvic incidence, lumbar lordosis angle, pelvic radius, sacral slope, pelvic tilt and grade of spondylolisthesis. These attributes are get classified under biomechanical feature vectors which are the major factor for diagnosing vertebral column pathologies.

2.5.2 Tools

In this research work, MATLAB and WEKA tools are taken for announcing the classification outputs. Both the tools are open-sourced and commonly used desktop environment for conducting an iterative analysis. MATLAB is a programming language consist of directly exploit matrix and array mathematics. It also facilitates with Live Editor under new scripts that are created with code combinations, output, and formatted text in an executable notebook. WEKA is one of the well-known popular data mining tools for conducting research analysis. It enables modification according to the research need and having the facility of re-implementing several data mining algorithms.

2.6 Result and Discussions

The research works result declaration is based on several evaluation matrices. The matrices taken for the observation and its obtained results are discussed below; 5.3.1 Classifier Evaluation Index: The common terms for declaring the prediction model's performance are accuracy (ACC), recall, precision (PPV), and the area under the curve (AUC). The confusion matrix is applied for evaluating these matrices. This matrix representation consists of columns and rows. In which the column determines the prediction categories and sum of the values states its category. The row determines the actual categories and sum of the values representing the categories of data observations. Below table 1 declares the training set and its obtained results.

Table-1 Training set results

Correctly Classified Instances 304	98.0645
Incorrectly Classified Instances 6	1.9355 %
Kappa statistic	0.969
Mean absolute error	0.0512
Root mean squared error	0.1157
Relative absolute error	12.2887 %
Root relative squared error	25.3725 %
Total Number of Instances	310

2.6.1 Confusion Matrix

A confusion matrix is an immediate statement generated based on the prediction results of a classification problem. The predictions can be classified into two kind's correct predictions and incorrect predictions. The count values and broken down determines whether the prediction comes under correct predictions or incorrect predictions. The best thing it does depicts the classifier error but also portrays what kind of error is done by that respective classifier.

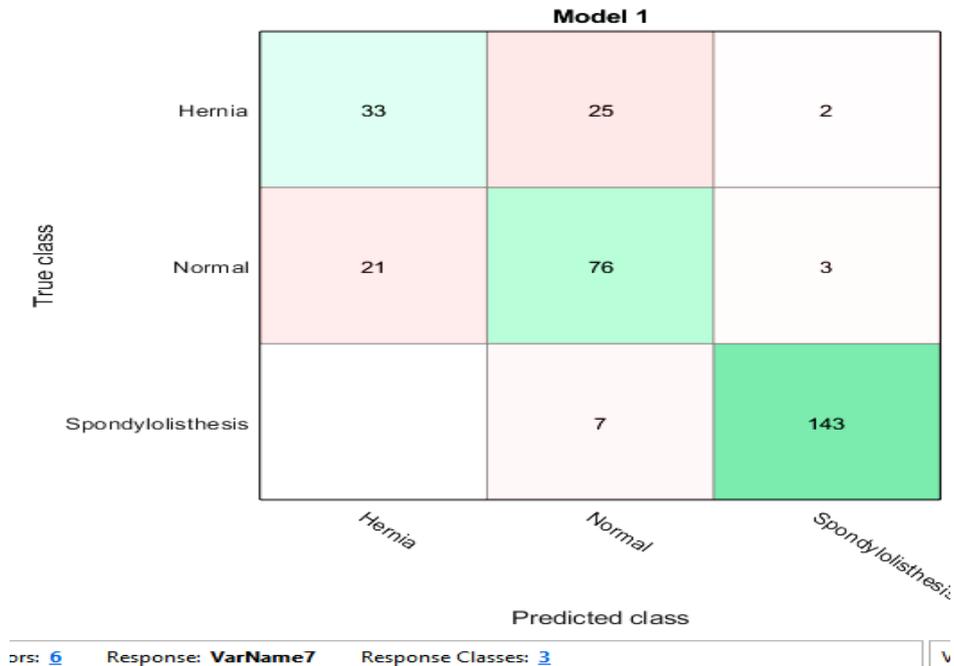


Fig-2 Number of observation

The above fig2 is the obtained result under the implementation of the confusion matrix. The output consists of true class in the X-axis and predicted class on the Y-axis. The observation is categorized in form three categorized as per the obtained values such as Hernia, Normal and spondylolisthesis.

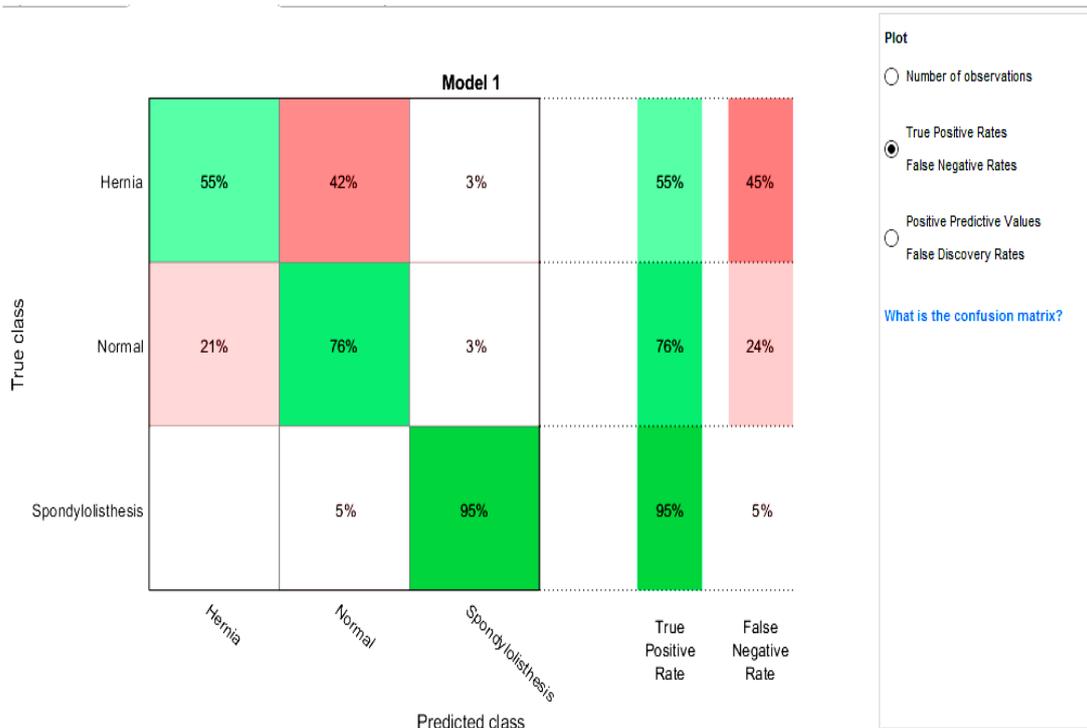


Fig-3 True positive and false-positive rates

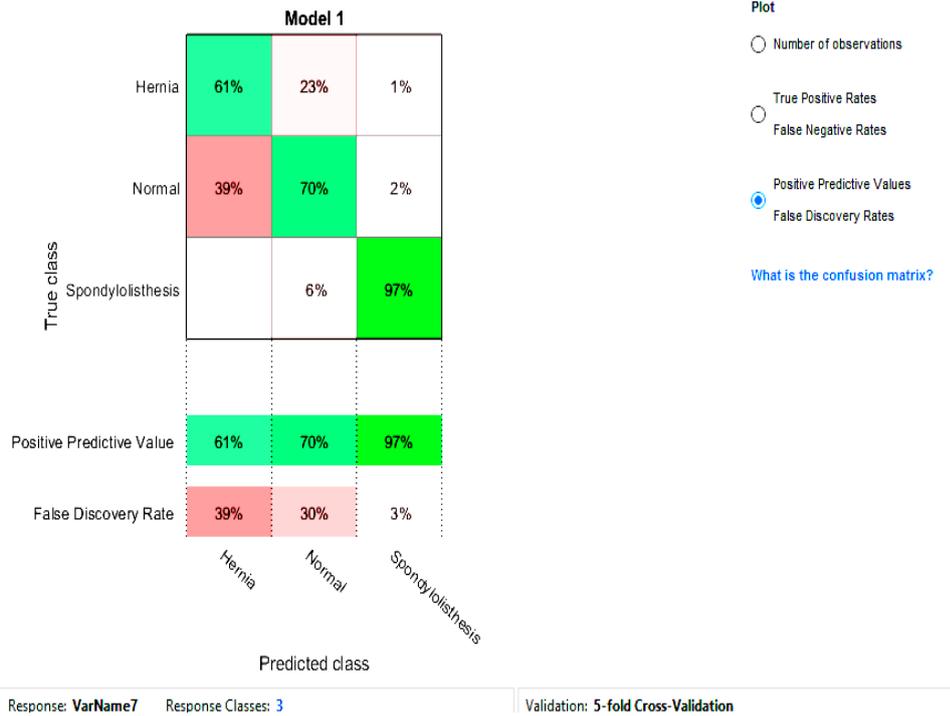


Fig-3 True positive and false-positive rates

Definition of the Terms •

Positive (P): Observation is positive

- Negative (N): Observation is not positive
- True Positive (TP): Observation is positive, and is predicted to be positive.
- False Negative (FN): Observation is positive, but is predicted negative.
- True Negative (TN): Observation is negative, and is predicted to be negative.
- False Positive (FP): Observation is negative, but is predicted positive.

Image 3 shows the true positive rate and false-negative rate respectively in the aspect of Hernia, Normal and spondylolisthesis. By which the true positive rate for hernia is 58% and 42% of the false-negative rate. Normal cases have a true positive rate of 76% and 24 % of the false-negative rate. For the spondylolisthesis, the true positive rate is 95% and the false-negative rate is 5%. Fig 4 shows the positive predictive and false discovery rates. The positive predictive for Hernia, Normal and spondylolisthesis are 67%, 70%, and 97% respectively. Whereas the false discovery rates for Hernia, Normal and spondylolisthesis are 33%, 30%, and 3% respectively.

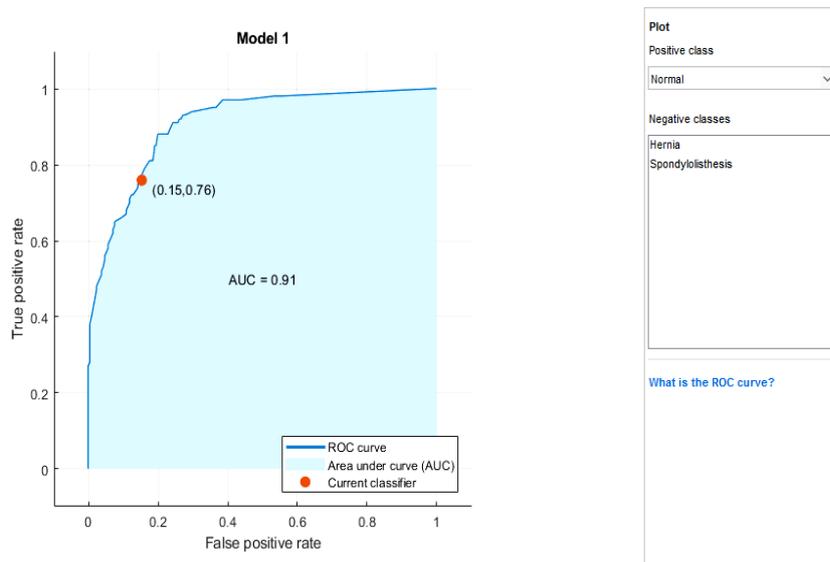


Fig-5 ROC plot for Normal

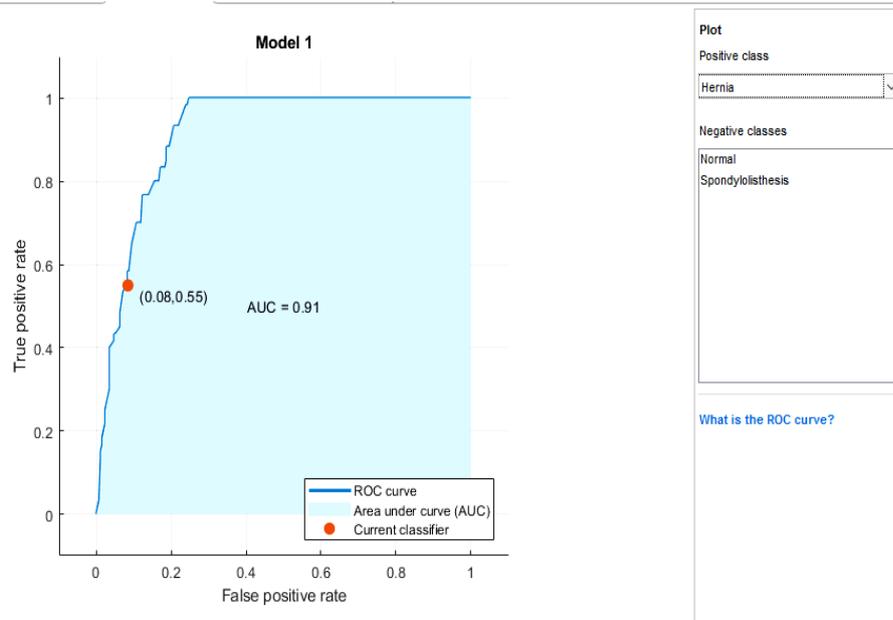


Fig-6 ROC plot for Hernia

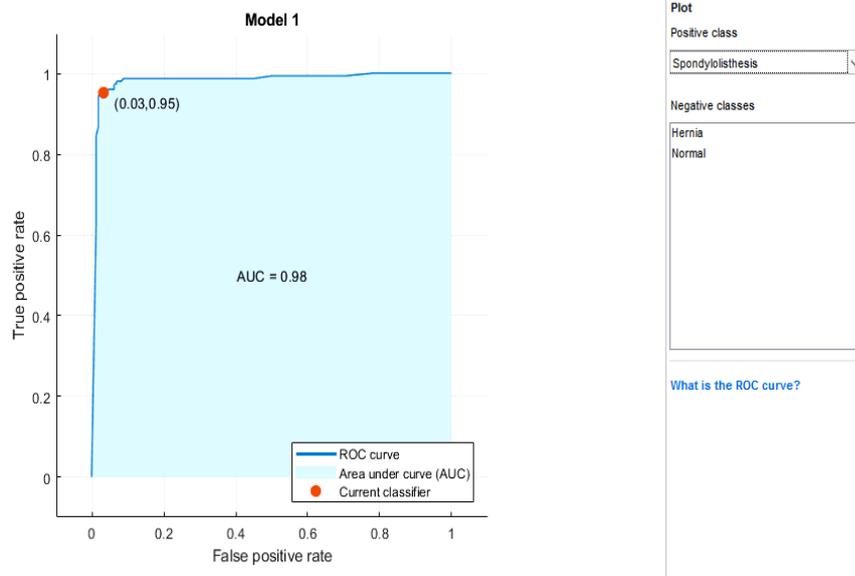


Fig-7 ROC plot for Spondylolisthesis

The ROC is a receiver operating characteristic curve and it consists of a graphical plot determining the diagnosis ability of the classifiers. The above figures 5, 6 & 7 have the redpoint which indicates the recall observation under which ROC for normal cases with 0.91, ROC for Hernia with 0.91 and ROC for Spondylolisthesis is 0.98 respectively.

2.6.2 Recall

The recall represents the total number of correctly classified positive examples divided by the total number of positive examples. The high recall values (a small number of FN) states the identification of correctly recognized examples. It can be stated as below;

$$Recall = \frac{TP}{TP + FN}$$

2.6.3 Precision

Precision is the value calculated by dividing the total number of correctly classified positive examples with the total number of predicted positive examples. The value with high precision is noted as positive is indeed positive (a small number of FP). It is stated as below;

$$\text{Precision} = \frac{TP}{TP + FP}$$

2.6.4 F-measure

F-Measure is obtained by measuring Precision and Recall. Instead of Arithmetic Mean, it uses Harmonic Mean for calculating F-measure because it deals with extreme values widely. On comparing the three, F-Measure is always smaller than the Precision or Recall. F-measure can be stated as below;

$$F - \text{measure} = \frac{2 * \text{recall} + \text{precision}}{\text{recall} + \text{precision}}$$

Table- 2 Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.983	0.012	0.952	0.983	0.967	1	Hernia
1	0.987	1	0.993	1	1	Spondylolisthesis
0.95	0.005	0.99	0.95	0.969	1	Normal
0.981	0.01	0.981	0.981	0.981	1	

shows the obtained results and its calculation based on TP-rate, FP-rate, Precision, Recall, F-measure and ROC for Hernia, Spondylolisthesis and normal classes respectively.

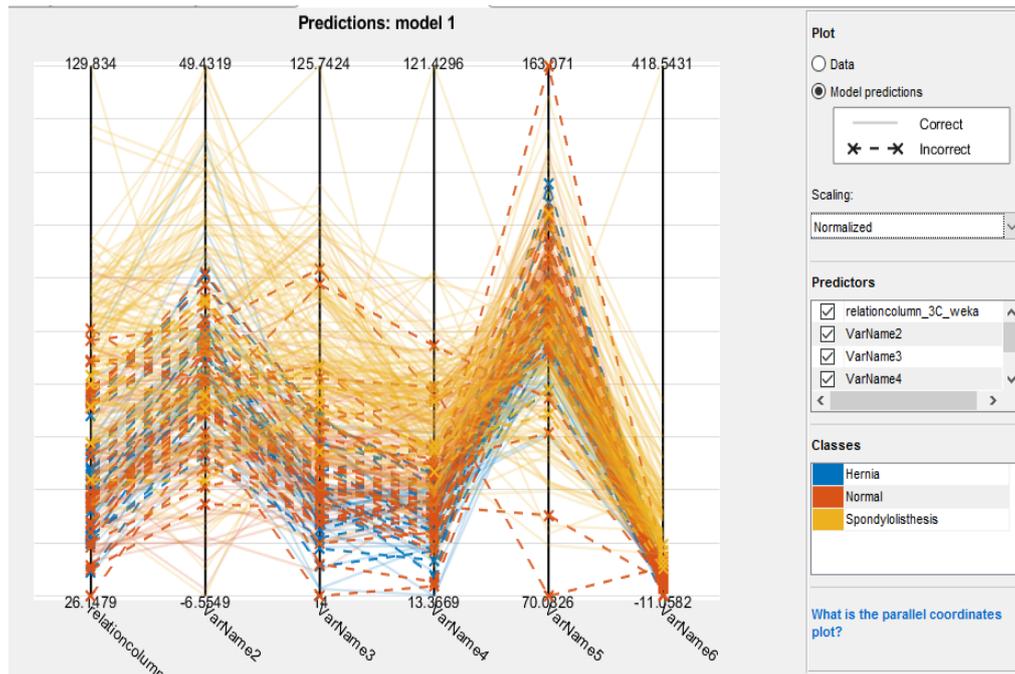


Fig-8 Parallel coordinates plot Normalized

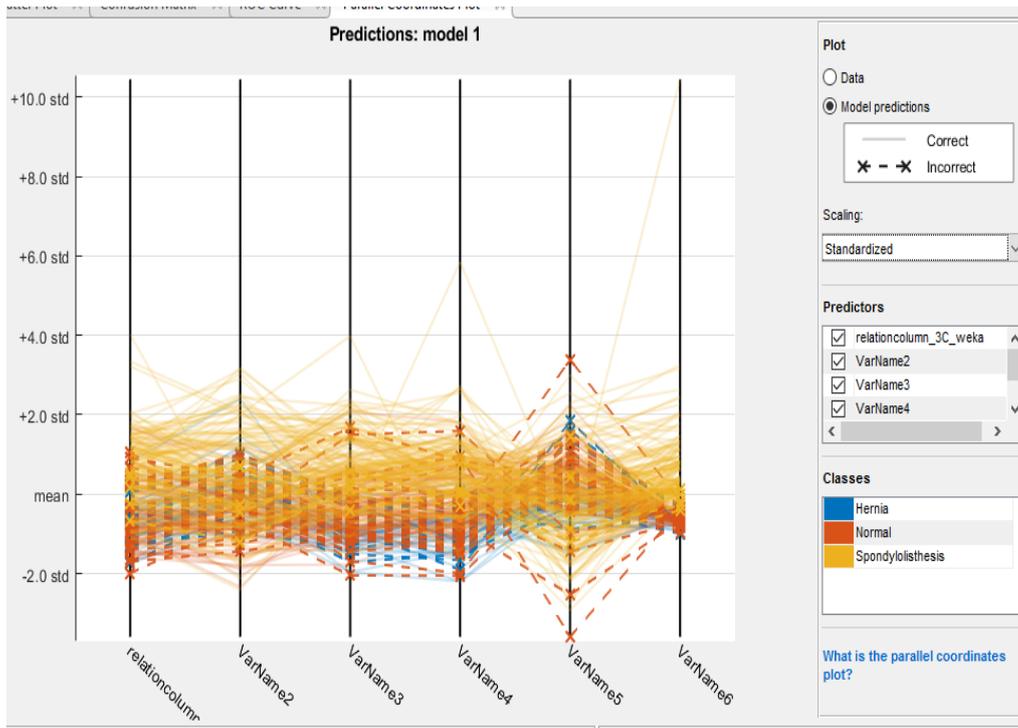


Fig-9 Parallel coordinates plot Standardized

The above fig 8 & 9 are the obtained parallel coordinates under normalized and standardized states of the prediction model. The straight line shows the correct prediction and the dotted line shows the incorrect prediction. On both the images, the maximum straight line is illustrated which determines the maximum vote for predictions.

Table 3: Overall performance comparison some exiting algorithms

S.No	Classification algorithm	Accuracy performance
1	SPRINT [12]	91%
2	Ensemble[11]	95 %
3	Random forest	98%

The experimental work is conducted to prove the efficiency of the proposed Random forest model with SPRINT [12] and Ensemble classifiers [11]. The evaluation matrices for all the three models are the same with the data source such UCI and tools taken which are discussed in [12 & 11] respectively. Based on the observation on the aspect of classification accuracy, all three algorithms' performance is tabulated. Table 3 shows the obtained accuracy value achieved by each classifier respectively. That is SPRINT with 91%, Ensemble with 95% and RF with 98%. From this table, its clear among the three RF achieves a higher accuracy value of about 98% proving its performance far better than the other two.

3. Conclusion

In this paper, the need for enhanced identification and classification in determining the diagnosis accuracy is discussed. There are several existing works are done based on improving accuracy such as SPRINT [12], Ensemble Classifiers [11], etc. The detailed works of this research area are discussed in the related work section. To enhance the accuracy, the proposed work has the implementation of the RF model which has two phases. In the first phase, the random forest is created by combining the N decision tree. Then the second phase deals with predictions at each created tree from the first phase. To show the performance of the proposed model a comparison work is carried out with the dataset taken from UCI under the MATLAB and WEKA tools. The algorithm taken for comparison is SPRINT, Ensemble Classifiers and RF. The results and discussion section describes the various factors that deal with the comparison. It is proved the proposed RF achieves the accuracy of 98% along with other evaluation metrics comparing to the other two models. Thus the RF has enhanced performance on the state of accuracy and classification speed in the effective prediction of vertebral column pathologies.

References

- [1] Devon I. Rubin, Epidemiology and Risk Factors for Spine Pain, *Neurologic Clinics*, Vol. 25, No. 2, 2007, Pages 353-371.
- [2] Low Back Pain Fact Sheet, National Institute of Neurological Disorders and Stroke, USA. Available: <https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/FactSheets/Low-Back-Pain-Fact-Sheet>. Last accessed: 1 October 2018.
- [3] Janet K. Freburger et. al., The Rising Prevalence of Chronic Low Back Pain, US National Library of Medicine, National Institutes of Health, *Arch Intern Med*. 2009 Feb; 169(3): 251258.
- [4] Dana Andersen, Timothy Billiar, David Dunn, John Hunter, Jeffrey Mathews and Rapheal E. Pollock. *Brunicardi Schwartz's Principles of Surgery*, 9th edition, Ed: McGraw Hill, 2009.
- [5] Tom Petersen, Mark Laslett and Carsten Juhl, Clinical classification in low back pain: best-evidence diagnostic rules based on systematic reviews. *BMC Musculoskeletal Disorders*; 2017. 18:188.
- [6] Alomari, R. S., Corso, J. J., Chaudhary, V., and Dhillon, G., Toward a clinical lumbar CAD: herniation diagnosis. *International journal of computer-assisted radiology and surgery* 6(1):119–126, 2011.
- [7] Chamrathy, P., Stanley, R. J., Cizek, G., Long, R., Antani, S., and Thoma, G., Image analysis techniques for characterizing disc space narrowing in cervical vertebrae interfaces. *Comput. Med. Imaging Graph.* 28:39–50, 2004.
- [8] Berthonnaud, E., Dimnet, J., Roussouly, P., and Labelle, H., Analysis of the sagittal balance of the spine and pelvis using shape and orientation parameters. *Journal of spinal disorders & techniques* 18(1):40–47, 2005.
- [9] Yao, J., Burns, J.E., Munoz, H., Summers, R.M., 2012. Detection of vertebral body fractures based on cortical shell unwrapping, in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. volume 7512 of LNCS, pp. 509–516.
- [10] Unal Y, Kocer HE. Diagnosis Of Pathology On The Vertebral Column With Back Propagation And Naïve Bayes Classifier, In *Technological Advances In Electrical, Electronics And Computer Engineering (TAECE)*, International Conference On IEEE 276: 2792013.
- [11] K.N. Nithya, Dr.P. Suresh, “Effective Prediction of Vertebral Column Pathologies Using Ensemble Classifiers”, The Mattingly Publishing Co., Inc. January - February 2020 ISSN: 0193 - 4120 Page No. 6362 - 6369
- [12] K.N. Nithya, P. Suresh, “Diagnosis of Vertebral Column Disorders using A Novel Sprint Algorithm”, *International Journal of Recent Technology and Engineering (IJRTE)* ISSN: 2277-3878, Volume-8 Issue-3, September 2019
- [13] Md. Shaiful Islam, Md. Asaduzzaman, Mohammad Masudur Rahman, “Feature Selection and Classification of Spinal Abnormalities to Detect Low Back Pain Disorder using Machine Learning Approaches”, 1st International Conference on Advances in Science, Engineering and Robotics Technology 2019 (ICASERT 2019)
- [14] Vijayalakshmi G V, Mohan Kumar M, “Diagnosis of Vertebral Column Pathologies using kNN Classifier”, *International Journal of Engineering Research & Technology (IJERT)* ISSN: 2278-0181 Published by, www.ijert.org RTESIT – 2019
- [15] Fabio Galbusera, Gloria Casaroli, Tito Bassani, “Artificial intelligence and machine learning in spine research”, *JOR Spine*. 2019;2:e1044, <https://doi.org/10.1002/jsp2.1044>
- [16] Ansari, S., Sajjad, F., Naveed, N., & Shafi, I. (2013, June). Diagnosis of vertebral column disorders using machine learning classifiers. In *2013 International Conference on Information Science and Applications (ICISA)* (pp. 1-6). IEEE.
- [17] Unal, Y., & Kocer, H. E. (2013, May). Diagnosis of pathology on the vertebral column with backpropagation and Naïve Bayes classifier. In *2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAECE)* (pp. 276-279). IEEE.