A Comparative Analysis of Machine Learning Algorithms in Vertebral Column Disorders Classification

Mohammad Z. Chowdhury and Dr. Charles C. Tappert
Seidenberg School of CSIS Pace University, Pleasantville NY 10570

***Abstract*— Spine disorder detection is a critical step for correct diagnosis of diseases such as Spondylolisthesis and Disk Hernia. Only experienced physician can identify abnormalities in the spine without any signal processing aids. However, capacity of human diagnostic suffers under adverse conditions, as stress, fatigue. With the recent advent of cost-effective hardware, computer aided classification system may assist physician in making right decision. In this paper we have applied four machine learning algorithms; Logistic Regression, K-Nearest Neighbor, Random Forest and XGBoost in vertebral column dataset to classify spine disorder and comparison results are presented.**

***Index Terms*—Disk hernia, Machine learning, Spondylolisthesis**

I. INTRODUCTION

Vertebral column, which consists of thirty-three vertebrates, invertebrate discs, nerves, muscles, medulla is an integral part of human body. Because of vertebral column, three level of body movements; frontal, sagittal and transversal is possible. These vertebrates are divided in five different groups and each vertebrae is separated by intervertebral disc, which serves as shock absorber and provides normal mobility between adjacent vertebrae. The result of small or several traumas in the column gradually injures the structure of the intervertebral disc which led to spine disorder such as Disk hernia and Spondylolisthesis. Thus, the arrangement and the features of those vertebrae represent important biomarkers for predicting spine disorder.

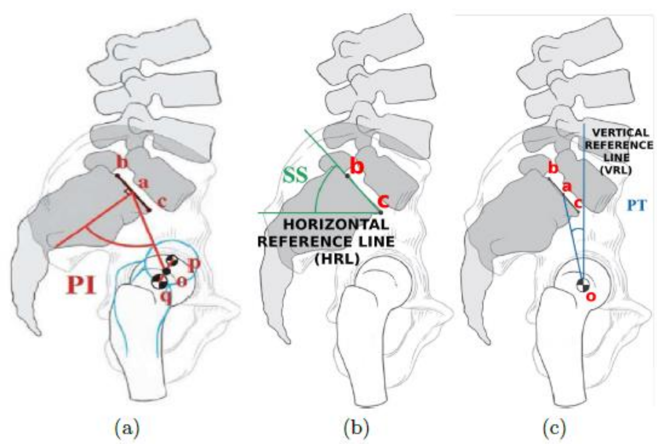

Figure 1 Spino-pelvic system. Pelvic incidence (PI) is defined as an angle subtended by line oa in fig. 1a. Sacral Slope (SS) is defined as the angle between the sacral endplate (bc) and the horizontal reference line (HRL), in Fig. 1b. Pelvic Tilt (PT) is defined as the angle between the vertical reference line (VRL) and the line joining the middle of the sacralend plate and the axis of the femoral heads in Fig. 1c.

However, parameters analysis of spino-pelvic system (figure 1) is labor intensive and error prone. Hence, there is a need to develop automated disease classification of spine disorder. Additionally, at remote area where specialist physicians are not available, sophisticated automated classification may assist in making right diagnosis.

The rest of the paper is organized as follows: A review of ML application in medical domain is presented in Section II. Dataset description and methodology are described in Section III. The experimental setup and description of ML algorithms are explained in Section IV. The experiment results and discussion are reported in Section V. Concluding remarks are presented in Section VI.

II. RELATED WORK

Spine disorders such as Disk hernia and Spondylolisthesis, are degenerative changes in spine over time [1]. Detecting the progression of these diseases creates a challenge to radiologist due subtle changes of biomechanically characteristics of disks. Radiologist must go through historical images to detect the changes in the early stages of diseases which is very expensive and time consuming. Incorporating machine learning technique in medical diagnosis aiding process has got tremendous progress in last decade.

However, credibility of these kind of ML system hinders widespread adoption due to contradiction to physician's proposed diagnosis and due to unreliable rendition of complex cases with very similar features [2]. Based on this observation, [3] proposed a framework enabling the classifier to add additional class, a rejection class (in addition to normal and abnormal classes), which is the critical case that would require manual expert analysis. [3] authors extended the works of [4] by exploiting overlapping classes region in higher dimensional space in ordinal dataset. Two binary classifiers were trained by heavily penalized the error weight and if output from these classifiers contradict each other, it is labeled as rejection class.

Ansari et al [5] experimented with three different algorithms using the same dataset for classification purpose. Among ANN (artificial neural networks), GRNN (generalized regression neural network) and SVM (support vector machine), GRNN outperformed other two by adopting 50% ration approach in training and test dataset. Important features selection using genetic algorithm and bagging technique showed improvement in classification of spine disorders [7]. Instead of disorder classifications, [9] emphasis on single classification disorder and further investigation revealed grade of spondylolisthesis (feature) is very important parameter for spondylolisthesis disorder detection.

III. METHODOLOGY

A. Dataset

We use the Vertebral Column dataset released by Dr. Henrique da Mota which contains 310 patient's data extracted from sagittal panoramic radiographies of spine. From this, 100 patients don't have any spinal disorders, hereafter referred to as normal. The remaining data are collected from the patients operated due to disc hernia (60 patients) or spondylolisthesis (150 patients). Therefore, this dataset is composed of 100 normal and 210 abnormal cases. Each of 310 patients spino-pelvic system is described by six biomechanical attributes: pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and degree of spondylolisthesis. The relationship between these attributes and common spinal pathologies (e.g. disc hernia and spondylolisthesis) was originally proposed in this paper [7].

B. Data pre-process

Although abnormal class to normal class ratio is 2:1, we considered this dataset is balanced. Initial data analysis revealed that degree of spondylolisthesis is important feature due its scattered separation from rest of the features. All of six features are in numerical form and there is no missing value. For classification purpose, we transformed class attribute to its correspondent numeric form using label encoder.

C. Classifier design

Four types of supervised classifiers were chosen including Logistic regression, K-Nearest Neighbor, Random Forest and XGBoost. Baseline models of all four classifiers were trained and evaluated on same dataset and their correspondence parameters were chosen as default in scikit-learn package.

IV. EXPERIMENTS

Workflow of this study as follows: First, a baseline model was trained and evaluated with same distribution of incidents to determine the baseline performance. Second, forward sequential features selection algorithm was implemented to choose most important features. Third, based on selected features, optimum hyperparameters search was perform using grid search with reasonable search space. Fourth, Model was trained with selected features along with fine-tuned parameters and its performance is observed. Fifth, a 10-fold cross validation was performed to evaluate the model performance (figure 2). Dataset was divided into two sets; training set and test set with correspondence 8:2 ratio.

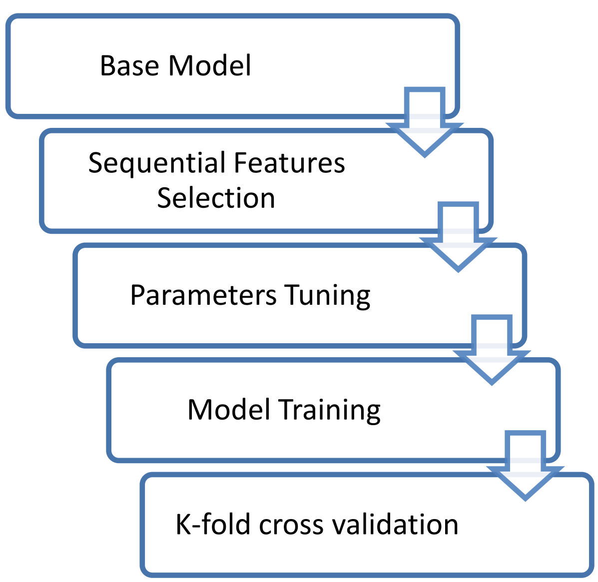

Figure 3 Five-stages workflow

A. Logistic Regression

Baseline model was trained along with penalty parameter l2, and inverse regularization parameter value 0.01. Model accuracy was 90.32 %. Four important features, pelvic incident, sacral slope, pelvic radius, and degree of spondylolisthesis were chosen based on performance value. A sperate model was trained with selected feature along with tuned hyperparameters. Model accuracy achieved same performance with fewer features. Since dataset is relatively small, we performed 10-fold cross validation to estimate the model accuracy.

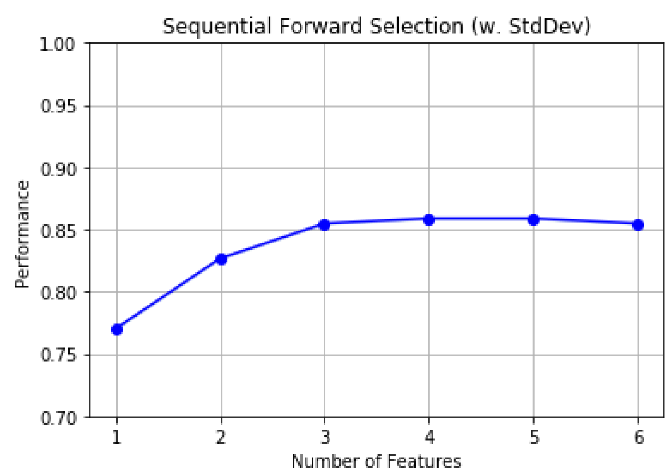

Figure 2 Sequential forward features selection in logistic regression

B. Random Forest Classifier

Baseline model was trained along with number of trees 10, and random state 4. Model accuracy was 85.48 %. Five important features, pelvic tilt numeric, lumber lordosis angle, sacral slope, pelvic radius, and degree of spondylolisthesis were chosen based on performance value. A sperate model was trained with selected feature along with tuned hyperparameters; number of trees 20, max depth 3. Model accuracy achieved 80.64%. Since dataset is relatively small, we performed 10-fold cross validation to estimate the model accuracy.

C. K-Nearest Neighbor Classifier

Baseline model was trained along with k 10 and model accuracy was 88.70%. Four important features, pelvic tilt numeric, sacral slope, pelvic radius, and degree of spondylolisthesis were chosen based on performance value. Model complexity was evaluated using WSS (with sum of square) and found that k=23 performed better. A sperate model was trained with selected feature along with tuned hyperparameters; k=3. Model accuracy achieved 90.32%. Since dataset is relatively small, we performed 10-fold cross validation to estimate the model accuracy.

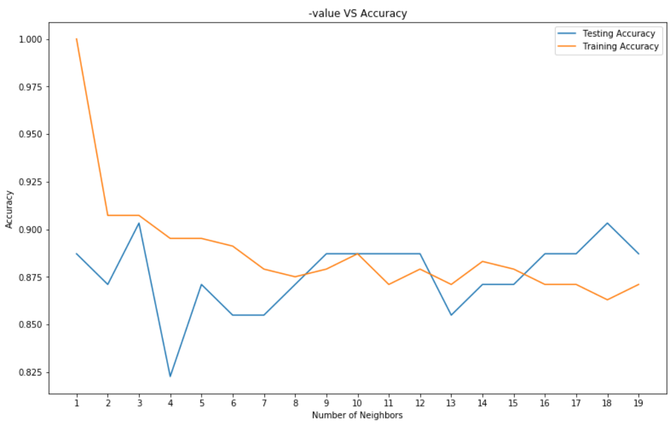

Figure 4 KNN model complexity. Accuracy and K is measure using With Sum of Square

D. XGBoost Classifier

Baseline model was trained along with number of tress 10, and max depth 3. Model accuracy was 85.48%. Features selection process revealed no difference in performance between 5 and 6 features. A sperate model was trained with all features along with tuned hyperparameters; number of trees 50 and max depth 3. Model accuracy achieved 88.70%. Since dataset is relatively small, we performed 10-fold cross validation to estimate the model accuracy.

V. RESULT

We performed five-stages operation on four selected algorithms on same distribution of dataset. Baseline model's performance evaluation are presented in table 1. Among four supervised model classifiers, logistic regression performed better than rest of three algorithms on same distribution of dataset with the accuracy of 90.32%. Reason being, baseline logistic regression model's false positive rate is smaller than rest of three models.

Table 1 Base model accuracy

Model	Test Accuracy
Logistic regression	**90.32%**
Random Forest Classifier	85.48%
K-Nearest Neighbor Classifier	88.70 %
XGBoost Classifier	85.48%

Next, important features selection was performed through sequential forward approach for each classifier algorithms and grid search approach was taken to find optimum hyperparameters for each algorithm. Four different models were trained correspondence to four different classifier algorithms and result is presented in table 2. Logistic regression model has no significant improvement whereas k-nearest neighbor outperformed rest of the three algorithms. Gradient boost tree (XGBoost) got little improvement due to parameter tuning.

Table 2 Models performance on selected features with tuned hyperparameters

Model (tuned)	Test Accuracy
Logistic regression	90.32%
Random Forest Classifier	85.48%
K-Nearest Neighbor Classifier (k-23)	**93.54 %**
XGBoost Classifier	87.09%

In final steps of this study, we performed 10-fold cross validation to see how four different classifiers perform in different variance of dataset and their result are presented in table 3. Due to very small dataset, XGBoost performed very poorly. Logistic regression performance is at same range of Random Forest classifier. KNN performed better since it is small dataset with only 4 features. Fine-tuned KNN model has higher k value (k =23) due to noise in dataset.

Table 3 Performance comparison through 10-fold cross validation

Model	10-fold cv
Logistic regression	85.08% (6.20% dv)
Random Forest Classifier	85.43% (7.13% dv)
K-Nearest Neighbor Classifier (k-23)	**86.29 %** (5.40%)
XGBoost Classifier	80.66% (7.70% dv)

VI. CONCLUSION

Providing efficient model on automatic classification of spinal disorders would assist physicians in decision making process. This paper presents critical analytics on disorder detection on four popular machine learning algorithms. It is concluded that K-Nearest Neighbor performed better on numerical features. In further works, it would be interesting to evaluate degree of spinal disorder and defining a context to detect particular disk's pathological condition.

Acknowledgment

Authors like to thank you Dr. Ron Frank and Prof. Barabasi for their valuable discussion and insights.

REFERENCES

[1] Kushchayev, Sergiy V et al. “ABCs of the degenerative spine.” Insights into imaging vol. 9,2 (2018): 253-274. doi:10.1007/s13244-017-0584-z

[2] Dreiseitl, Stephan, and Michael Binder. "Do physicians value decision support? A look at the effect of decision support systems on physician opinion." Artificial intelligence in medicine 33.1 (2005): 25-30.

[3] da Rocha Neto, Ajalmar R., et al. "Diagnostic of pathology on the vertebral column with embedded reject option." Iberian Conference on Pattern Recognition and Image Analysis. Springer, Berlin, Heidelberg, 2011.

[4] Sousa, Ricardo, Beatriz Mora, and Jaime S. Cardoso. "An ordinal data method for the classification with reject option." 2009 International Conference on Machine Learning and Applications. IEEE, 2009.

[5] Ansari, Sana, et al. "Diagnosis of vertebral column disorders using machine learning classifiers." 2013 International Conference on Information Science and Applications (ICISA). IEEE, 2013.

[6] Ebrahimzadeh, Elias, et al. "A machine learning-based method in order to diagnose lumbar disc herniation disease by MR image processing." MedLife Open Access 1.1 (2018): 1.

[7] Berthonnaud, E., Dimnet, J., Roussouly, P., 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)

[8] Prasetio, Rizki Tri and Dwiza Riana. “A comparison of classification methods in vertebral column disorder with the application of genetic algorithm and bagging.” 2015 4th International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering (ICICI-BME) (2015): 163-168.

[9] Akben, Selahaddin Batuhan. “Importance of the Shape and Orientation of the Spine and Pelvis for the Vertebral Column Pathologies Diagnosis with Using Machine Learning Methods.” Biomedres.Info, 30 Dec. 2016,