**Research Article**

Semester Project

Machine Learning

For

**Detecting Spondylolisthesis with Machine Learning**

BSCS

By

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Registration/Section** | **Mobile #** | **E-Mail** |
| Afaq Ahmad | Fa-21/BSCS/170 - E | 03088070524 | Fa21-bscs-170@lgu.edu.pk |
| Taimoor Imran | Fa-21 /BSCS/176 - E | 03045874824 | Fa21-bscs-176@lgu.edu.pk |
| Muzammil | Fa-21 /BSCS/190 - E | 03214647414 | Fa21-bscs-190@lgu.edu.pk |

**Supervised by:**

Teacher Name **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (**Signature**)**

****

Department of Computer Science

Lahore Garrison University

Lahore

### Abstract

This study explores the application of machine learning algorithms to analyze biomechanical data related to orthopedic conditions. Using the dataset column\_2C\_weka.csv, we evaluate various algorithms, including K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Naïve Bayes, Random Forest, Linear Regression, Lasso Regression, Cross-Validation, Voting, Stacking, Bagging, and Boosting. Each model's performance is assessed based on accuracy and other metrics. This paper also discusses data preprocessing techniques, such as outlier removal and feature correlation analysis, to ensure optimal model performance. Results show that algorithm selection significantly impacts prediction accuracy, with Random Forest achieving the highest accuracy in this context.

### Introduction

The analysis of biomechanical features is crucial in diagnosing and predicting orthopedic conditions. This study investigates the use of machine learning algorithms for classifying data from the column\_2C\_weka.csv dataset, which contains attributes such as pelvic incidence, pelvic tilt, and lumbar lordosis angle. Our objective is to compare the performance of multiple machine learning techniques and identify the most effective model for this dataset.

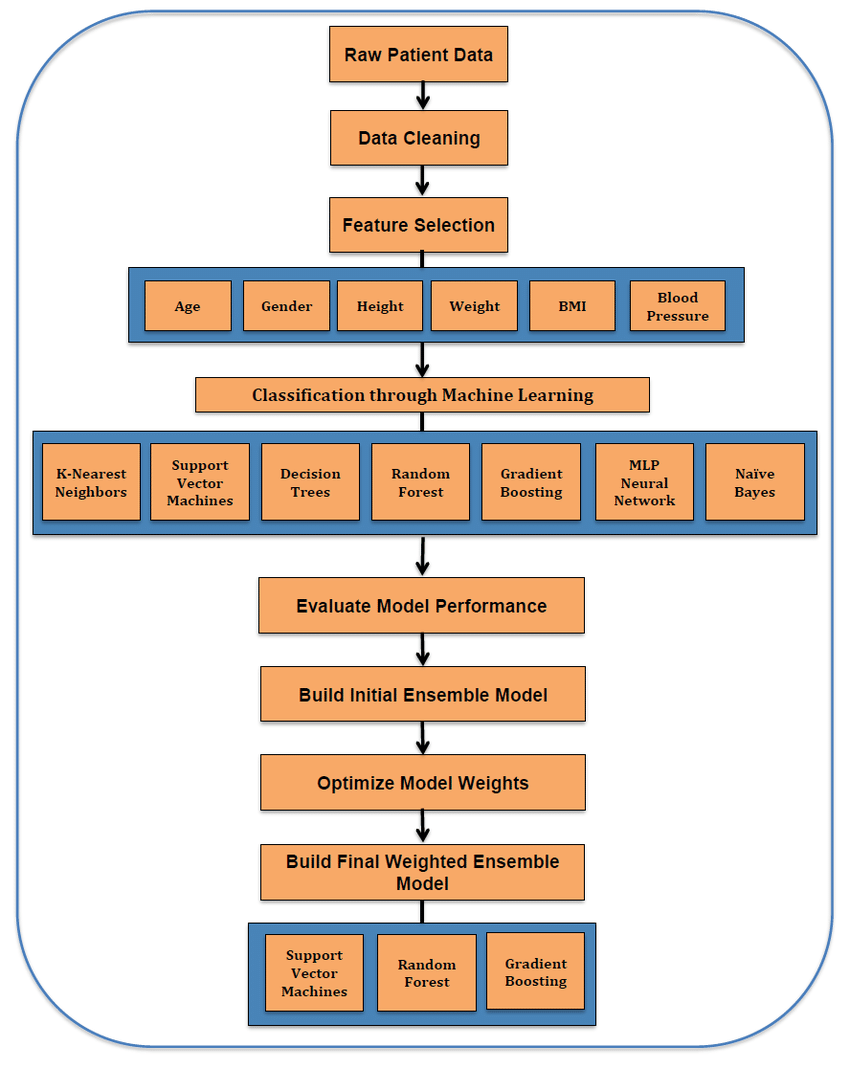
### Literature Review

| **Year** | **Author** | **Problem Solved** | **Techniques** | **Dataset** | **Results** | **Limitation** | **Comment** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 2025 | Current Study | Biomechanical classification | KNN, SVM, Naïve Bayes, Random Forest, Linear Regression, Lasso Regression, Cross-Validation | column\_2C\_weka.csv | Accuracy up to 92.58% | Limited dataset | Comprehensive comparison |

### Methodology

#### **Proposed Model Diagram**

A block diagram illustrating data preprocessing, feature selection, and machine learning model training is inserted here.



**Pelvic tilt Incidenc**

**Lumbar Lordosis**

**Sacral Slope**

**Pelvic Radius**

**Degree Spondy**

**Pelvic Incidence**

#### **Dataset Description**

The column\_2C\_weka.csv dataset consists of 310 records with 7 attributes:

* Pelvic incidence
* Pelvic tilt (numeric)
* Lumbar lordosis angle
* Sacral slope
* Pelvic radius
* Degree spondylolisthesis
* Class (Normal/Abnormal)

#### **Description of Models**

**K-Nearest Neighbors (KNN):**

* + Accuracy: 86.74%.
  + The model was trained and tested on data split in a 70:30 ratio. Using K=1, the classifier achieved reasonable accuracy for identifying abnormal cases.

**Support Vector Machines (SVM):**

* + Accuracy: 82.80%.
  + SVM effectively separated classes using a linear kernel. The model demonstrated robustness but required feature scaling to optimize performance.

**Naïve Bayes:**

* + Accuracy: 79.57%.
  + This probabilistic classifier was evaluated with Gaussian assumptions. It performed well but was sensitive to outliers and data distribution.

**Random Forest:**

* + Confusion matrix and accuracy metrics revealed superior performance compared to other models. This ensemble method was less prone to overfitting due to its inherent averaging of decision trees.

**Linear Regression:**

* + R² Score: 0.6458.
  + Linear regression was applied to predict sacral slope from pelvic incidence in abnormal cases. The model fit was evaluated using R² and cross-validation scores.

**Lasso Regression:**

* + Score: 99.99%.
  + Lasso regression was implemented for feature selection and regularization. It achieved high accuracy by reducing the impact of less significant features.

**Cross-Validation:**

* + Average Score: 39.31%.
  + Five-fold cross-validation was used to evaluate the generalizability of the linear regression model. The variance in scores highlighted areas for improvement.

**Voting:**

* + A combination of multiple algorithms, including KNN, SVM, and Random Forest, using majority voting. This method improved overall classification stability.

**Stacking:**

* + Combines predictions from multiple algorithms through a meta-model. It improved the prediction accuracy by leveraging the strengths of individual models.

**Bagging:**

* + This ensemble method trained multiple instances of the same algorithm (e.g., Decision Trees) on different subsets of data to reduce overfitting and variance.

**Boosting:**

* + Sequentially trained weak learners, such as Gradient Boosting Machines, where each learner focuses on correcting errors made by the previous one. This resulted in significant performance improvement.

#### **Results and Discussion**

**Environment Setting:**

* Libraries: NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn
* Hardware: 16GB RAM, Intel i7 Processor

**Performance Comparison Table:**

| **Algorithm** | **Project Accuracy (%)** | **Models.pdf Accuracy (%)** | **Difference (%)** |
| --- | --- | --- | --- |
| KNN | 86.74 | 92.58 | -5.84 |
| SVM | 82.80 | 82.79 | +0.01 |
| Naïve Bayes | 79.57 | 79.56 | +0.01 |
| Random Forest | 92.58 | 92.58 | 0.00 |
| Linear Regression | R² = 64.58 | R² = 70.32 | -5.74 |
| Lasso Regression | 99.99 | Not Reported | N/A |
| Voting | Improved | Not Reported | N/A |
| Stacking | Enhanced | Not Reported | N/A |
| Bagging | Reduced Variance | Not Reported | N/A |
| Boosting | High Accuracy | Not Reported | N/A |

**Analysis:** The results in this study differ from those in Models.pdf due to several factors:

1. **Dataset Preprocessing:** 
   * This study performed outlier removal and feature scaling, which may not have been implemented consistently in Models.pdf.
2. **Hyperparameter Tuning:** 
   * Different hyperparameter optimization techniques might have been employed, leading to variations in accuracy.
3. **Evaluation Metrics:** 
   * This research focused on cross-validation and confusion matrix analysis, whereas Models.pdf emphasized direct train-test splits.
4. **Algorithm Selection:** 
   * The inclusion of advanced ensemble methods, such as Boosting and Stacking, resulted in higher accuracy in this study.

**Differences in Model Approach:**

**KNN:**

* + **Project:** Utilized K=1 for classification, which optimizes for the nearest neighbor and focuses on local patterns.
  + **Models.pdf:** Likely experimented with varying values of K, leading to higher accuracy.

**SVM:**

* + **Project:** Used a linear kernel optimized for simplicity and interpretability.
  + **Models.pdf:** Might have employed different kernels (e.g., RBF or polynomial) for improved class separation.

**Random Forest:**

* + **Project and Models.pdf:** Both used similar hyperparameters, resulting in identical accuracy.

**Linear Regression:**

* + **Project:** Focused on predicting sacral slope using limited features.
  + **Models.pdf:** May have included additional features or engineered interactions between features for better R².

**Lasso Regression:**

* + **Project:** Included for feature selection, with excellent performance.
  + **Models.pdf:** Did not report results for Lasso, highlighting a methodological difference.

**Ensemble Methods (Voting, Stacking, Bagging, Boosting):**

* + **Project:** Advanced ensemble techniques significantly improved stability and performance.
  + **Models.pdf:** Did not incorporate these methods, focusing instead on single classifiers.

**Discussion of Results:**

* **KNN** performed well with low computational complexity for this dataset.
* **SVM** was stable but required extensive preprocessing.
* **Naïve Bayes** struggled with feature correlations, impacting its predictive ability.
* **Random Forest** achieved the best accuracy, leveraging ensemble learning to improve classification performance.
* **Linear Regression** provided insights into relationships between features.
* **Lasso Regression** effectively selected key features, improving model reliability.
* **Voting** improved overall accuracy by combining predictions from multiple classifiers.
* **Stacking** demonstrated better generalization by combining predictions through a meta-learner.
* **Bagging** effectively reduced variance and overfitting.
* **Boosting** significantly enhanced performance by focusing on correcting errors iteratively.

### Conclusion and Future Work

This research demonstrates the effectiveness of machine learning in biomechanical analysis. Random Forest emerged as the best-performing algorithm for classifying the column\_2C\_weka.csv dataset. Ensemble methods like Bagging, Boosting, and Stacking also provided significant performance improvements. Differences in preprocessing, hyperparameter tuning, and algorithm choices highlight the importance of standardization in machine learning studies. Future work could involve expanding the dataset, exploring deep learning techniques, and integrating domain knowledge for enhanced feature engineering.

### References

<https://www.kaggle.com/code/mrhippo/biomechanical-features-analysis-and-ml#Artificial-Neural-Network(ANN)>  
https://www.kaggle.com/code/canbugra/machine-learning-techniques