BioMedInformatics

Article

Machine-Learning-Based Biomechanical Feature Analysis for Orthopedic Patient Classification with Disc Hernia and Spondylolisthesis

Daniel Nasef [1], Demarcus Nasef [1], Viola Sawiris [1], Peter Girgis [2] and Milan Toma [1,*]

[1] Department of Osteopathic Manipulative Medicine, College of Osteopathic Medicine, New York Institute of Technology, Old Westbury, NY 11568, USA; dnasef02@nyit.edu (D.N.); dnasef01@nyit.edu (D.N.); vsawiris@nyit.edu (V.S.)

[2] Downstate Health Sciences University, State University of New York, 445 Lenox Rd., Brooklyn, NY 11203, USA; peter.girgis@downstate.edu

* Correspondence: tomamil@tomamil.com

Abstract: (1) **Background**: The exploration of various machine learning (ML) algorithms for classifying the state of Lumbar Intervertebral Discs (IVD) in orthopedic patients is the focus of this study. The classification is based on six key biomechanical features of the pelvis and lumbar spine. Although previous research has demonstrated the effectiveness of ML models in diagnosing IVD pathology using imaging modalities, there is a scarcity of studies using biomechanical features. (2) **Methods**: The study utilizes a dataset that encompasses two classification tasks. The first task classifies patients into Normal and Abnormal based on their IVDs (2C). The second task further classifies patients into three groups: Normal, Disc Hernia, and Spondylolisthesis (3C). The performance of various ML models, including decision trees, support vector machines, and neural networks, is evaluated using metrics such as accuracy, AUC, recall, precision, F1, Kappa, and MCC. These models are trained on two open-source datasets, using the PyCaret library in Python. (3) **Results**: The findings suggest that an ensemble of Random Forest and Logistic Regression models performs best for the 2C classification, while the Extra Trees classifier performs best for the 3C classification. The models demonstrate an accuracy of up to 90.83% and a precision of up to 91.86%, highlighting the effectiveness of ML models in diagnosing IVD pathology. The analysis of the weight of different biomechanical features in the decision-making processes of the models provides insights into the biomechanical changes involved in the pathogenesis of Lumbar IVD abnormalities. (4) **Conclusions**: This research contributes to the ongoing efforts to leverage data-driven ML models in improving patient outcomes in orthopedic care. The effectiveness of the models for both diagnosis and furthering understanding of Lumbar IVD herniations and spondylolisthesis is outlined. The limitations of AI use in clinical settings are discussed, and areas for future improvement to create more accurate and informative models are suggested.

Keywords: machine learning; orthopedic patient classification; biomechanical features; Disc Hernia; Spondylolisthesis; Lumbar Intervertebral Disc abnormalities

check for updates

Academic Editor: Moulay A. Akhloufi

Received: 17 September 2024
Revised: 15 October 2024
Accepted: 2 January 2025
Published: 7 January 2025

Citation: Nasef, D.; Nasef, D.; Sawiris, V.; Girgis, P.; Toma, M. Machine-Learning-Based Biomechanical Feature Analysis for Orthopedic Patient Classification with Disc Hernia and Spondylolisthesis. *BioMedInformatics* **2025**, *5*, 3. https://doi.org/10.3390/biomedinformatics5010003

1. Introduction

Lumbar Intervertebral Disc (IVD) abnormalities are a leading cause of chronic lower back pain. Among these abnormalities, lumbar disc herniations and spondylolisthesis are particularly common. IVD herniation occurs when the nucleus pulposus, the gel-like core, breaches the annulus fibrosus, often leading to nerve compression and resultant pain [1].

When the nucleus pulposus bulges it displaces disc material into the spinal canal. This displacement can exert pressure on adjacent nerve roots, resulting in pain and neurological symptoms [2,3]. Approximately 80% of individuals in the United States experience back pain, most of which lumbar disc herniation [3]. Spondylolisthesis, characterized by the slippage of one vertebral body over another, is another significant contributor to lower back pain. It occurs when a fracture to the pars interarticularis causes destabilization of the vertebrae and anterior slippage [3,4]. The resulting instability can lead to nerve compression and various neurological symptoms, exacerbating the patient's pain and functional limitations [2,5]. This condition can arise from various causes, including congenital defects, degenerative changes, or traumatic injuries. The most prevalent form, degenerative spondylolisthesis, occurs because of age-related degeneration of the intervertebral discs and facet joints, leading to instability and displacement of the vertebrae [5]. Isthmic spondylolisthesis, another common type, results from a defect in the pars interarticularis, often due to stress fractures, particularly in athletes engaged in activities that involve hyperextension of the spine [6]. The clinical manifestations of spondylolisthesis can vary widely, ranging from localized back pain to radicular symptoms caused by nerve root compression. When left untreated, disc herniations can progress to cause more nerve damage leading to an array of symptoms including urinary and fecal incontinence, cauda equina syndrome, and irreversible paralysis [7,8]. Similarly, untreated Spondylolisthesis may lead to Chronic back pain, spinal arthritis, nerve damage, paralysis of the legs, and urinary and bowel incontinence [7–9]. Thus, it is imperative to accurately diagnose and treat these conditions before progression. Currently, both lumbar IVD herniations and Spondylolisthesis are diagnosed with imaging studies—most commonly X-rays and CT scans [3,7,10].

Machine learning for diagnostics in orthopedics is a large focus of current research. Previous studies have developed models to diagnose osteoarthritis, osteoporosis, and fracture types (cervical spine and rib fractures) using data from X-rays and other imaging modalities [11,12]. It has also been used to access Acetabular Inclination and Version After Total Hip Arthroplasty (THP) and to assess risk of hip dislocation after THP [13,14]. While the current literature extensively covers the use of deep learning models to make orthopedic diagnoses based on radiological images, there is a gap in literature on use of machine learning models to diagnose off patient biomechanical features.

In terms of Lumbar Disc pathologies, previous studies reviewed the use of ML for diagnosing chronic low back pain (LBP), including intervertebral disc degeneration and spondylolisthesis, reporting high accuracy rates, often above 80% for classification tasks using deep learning models [15]. Notably, neural networks such as convolutional neural networks (CNNs) have been employed to segment and classify spinal structures like discs, vertebrae, and spinal canals [15,16]. Other studies have also extended beyond imaging to assess the role of patient-specific factors like body weight, psychological health, and the duration of symptoms in predicting outcomes such as neuropathic pain after lumbar disc herniation surgery [17]. Models utilizing deep learning approaches have achieved accuracy rates above 90% for identifying herniations, spinal stenosis, and spondylolisthesis in MRI datasets [15,18].

However, challenges remain in integrating these AI systems into clinical workflows [19]. One key limitation is their reliance on high-quality, standardized imaging data, which may not always be available in routine clinical settings [16]. In addition, while AI can process vast amounts of data efficiently, it still struggles to interpret more nuanced patient-specific factors such as variations in biomechanical markers, pain perception, and psychosocial influences [17]. The inclusion of such data, and specifically biomechanical markers, may allow models to more accurately predict patient diagnoses and outcomes [17].

The current study analyzed machine learning models to predict abnormalities in the lumbar spine by utilizing key biomechanical features derived from the shape and orientation of the pelvis and lumbar spine. The dataset for this analysis was sourced from the Kaggle repository [20], and the models were configured using PyCaret version 3.0.3. Two datasets, referred to as '2C' and '3C', were used in the analysis, with '2C' classifying patients into 'Normal' or 'Abnormal' and '3C' further dividing the 'Abnormal' group into 'Disc Hernia' and 'Spondylolisthesis'.

2. Materials and Methods

We used 2 open-source datasets from Kaggle [20] that detailed 6 key biomechanical markers and the lumbar IVD status of orthopedic patients. The biomechanical perameters included pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius, and grade of spondylolisthesis. These makers were all derived from the shape and orientation of the patients' pelvis and lumbar spine. The datasets reported these measurable biomechanical features for each patient in addition to the state of their lumbar disc. One dataset classified patients into three groups based on the state of their lumbar intervertebral discs: "Normal" (n = 100), "Disc Hernia" (n = 60), or "spondylolisthesis" (n = 150). This dataset and models based on this dataset are labeled "3C". The other dataset classified patients into 2 groups based on the state of their lumbar intervertebral discs: "Normal" (n = 100) or "Abnormal" (n = 210). The abnormal group combines patients who suffer from a disc herniation or spondylolisthesis. This dataset and models based on this dataset are labeled "2C".

To analyze and build models for these datasets, we used the PyCaret library, an open-source, low-code machine learning library in Python that automates machine learning work-flows. We used Jupyter Notebook as our integrated development environment (IDE) for writing and executing Python code. The setup in PyCaret was initialized with the appropriate parameters, including the target variable and the specific preprocessing tasks to be performed on the data. Additionally, we applied data transformations and removed multicollinearity among features which prevents potential issues with correlated features that can skew results. Various machine learning models from PyCaret were compared and accessed on their accuracy, Area Under the Curve (AUC), recall, precision, F1, Kappa, MCC, and TT. After comparing the models, the best model for each dataset was selected based on the evaluation metrics. For the 2C dataset, we created an ensemble model by blending the two top performing machine learning models: Random Forest and Logistic Regression models. This blending process combines the strengths of both models to improve overall predictive performance. After creating the ensemble model, it is retrained to ensure that it is optimized for the dataset. For the 3C dataset, the Extra Trees classifier model was selected. Although this model was not the highest performing model, it was still among the best performing models and was selected to prevent overfitting. The selected models were tuned to adjust the hyperparameters of the model to enhance its performance. During this process, we utilized various visualizations to improve our understanding of the models including: Learning Curves, Confusion Matrices, AUC graphs, and Features Importance graphs. These visualizations are useful for accessing model validity and identifying potential areas for improvement to ensure the reliability of lumbar IVD status predictions.

3. Results

The comparison tables show the average score of each model across all folds and are seen in Tables 1 and 2. The learning curves for the model trained in the 2C dataset (Figure 1a) provide a visual representation of the logistic regression's performance as it learns from more data. The training curve, which starts above the cross-validation curve, represents the model's performance on the training data. The cross-validation curve, which rises gradually and eventually plateaus, represents the model's performance on unseen data. At the start of training, the model performs well on the training data but not as well on the unseen data, hence the training curve is above the cross-validation curve. As the model is exposed to more training instances, it begins to generalize better to unseen data, which is reflected in the rising cross-validation curve. The point where the training curve meets the cross-validation curve is significant. This point of convergence indicates that the model has reached a state where it performs similarly on both the training and validation data. This is typically the point where the model has learned to generalize well and is neither overfitting (performing significantly better on the training data than on the validation data) nor underfitting (performing poorly on both the training and validation data).

Table 1. A cross-validation for metric evaluation of various machine learning models on the 2C dataset.

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	TT (Sec)
Logistic Regression	0.8576	0.9294	0.8576	0.8683	0.8590	0.6840	0.6910	0.2380
Ridge Classifier	0.8437	0.8922	0.8437	0.8524	0.8359	0.6188	0.6385	0.0050
Linear Discriminant Analysis	0.8437	0.8922	0.8437	0.8534	0.8377	0.6255	0.6430	0.0050
Gradient Boosting Classifier	0.8338	0.9106	0.8338	0.8373	0.8329	0.6180	0.6230	0.0190
Extra Trees Classifier	0.8299	0.9167	0.8299	0.8302	0.8237	0.5914	0.6009	0.0240
Random Forest Classifier	0.8251	0.9078	0.8251	0.8261	0.8232	0.5934	0.5976	0.0270
Light Gradient Boosting Machine	0.8208	0.8935	0.8208	0.8233	0.8181	0.5838	0.5894	0.0680
K Neighbors Classifier	0.8121	0.8901	0.8121	0.8221	0.8128	0.5771	0.5842	0.0120
Quadratic Discriminant Analysis	0.7931	0.9185	0.7931	0.8615	0.7985	0.5947	0.6362	0.0050
Naive Bayes	0.7792	0.8823	0.7792	0.8286	0.7847	0.5494	0.5772	0.0050
Decision Tree Classifier	0.7792	0.7474	0.7792	0.7828	0.7785	0.4962	0.4994	0.0060
Ada Boost Classifier	0.7606	0.8293	0.7606	0.7634	0.7563	0.4421	0.4501	0.0140
SVM-Linear Kernel	0.7002	0.8865	0.7002	0.7062	0.6489	0.2789	0.3289	0.0060
Dummy Classifier	0.6773	0.5000	0.6773	0.4587	0.5470	0.0000	0.0000	0.0040

Table 2. A cross-validation for metric evaluation of various machine learning models on the 3C dataset.

Model	Accuracy	AUC	Recall	Precision	F1	Kappa	MCC	TT (Sec)
Random Forest Classifier	0.9083	0.9783	0.9083	0.9186	0.9066	0.8623	0.8689	0.0310
Gradient Boosting Classifier	0.8989	0.0000	0.8989	0.9067	0.8979	0.8482	0.8528	0.0520
Extra Trees Classifier	0.8956	0.9833	0.8956	0.9030	0.8942	0.8432	0.8478	0.0310
Light Gradient Boosting Machine	0.8891	0.9681	0.8891	0.8997	0.8875	0.8335	0.8401	0.2950
Decision Tree Classifier	0.8634	0.8978	0.8634	0.8758	0.8608	0.7949	0.8034	0.0080
Logistic Regression	0.8446	0.0000	0.8446	0.8568	0.8451	0.7667	0.7720	0.0330
Ridge Classifier	0.8225	0.0000	0.8225	0.8354	0.8148	0.7336	0.7439	0.0080
Linear Discriminant Analysis	0.8225	0.0000	0.8225	0.8286	0.8201	0.7335	0.7385	0.0090
Naive Bayes	0.7875	0.9115	0.7875	0.7946	0.7806	0.6813	0.6895	0.0100
Quadratic Discriminant Analysis	0.7872	0.0000	0.7872	0.7967	0.7846	0.6804	0.6871	0.0090
K Neighbors Classifier	0.5943	0.7767	0.5943	0.6310	0.5959	0.3910	0.4013	0.0120
Ada Boost Classifier	0.5106	0.0000	0.5106	0.5668	0.4622	0.2668	0.3048	0.0190
SVM-Linear Kernel	0.4200	0.0000	0.4200	0.3752	0.2916	0.1230	0.1951	0.0100
Dummy Classifier	0.3175	0.5000	0.3175	0.1009	0.1531	0.0000	0.0000	0.0070

Similarly, the learning curves for the model trained on the 3C data follow a common pattern seen in machine learning (see Figure 1b). The training curve starting above the cross-validation curve indicates that the model initially has a better performance on the training data than on the validation data. This is a common occurrence as models are typically better at predicting outcomes for data they have already seen (training data) compared to new, unseen data (validation data). The fact that the curves meet at similar values and rise together until they plateau suggests that the model is learning effectively. The meeting point of the curves indicates that the model has learned to generalize from the training data to unseen data, which is a crucial aspect of machine learning models. The subsequent parallel rise of the curves until they plateau indicates that the model continues to improve its performance on both the training and validation data as it is exposed to more training instances. The plateau suggests that the model has reached a point where additional training instances do not significantly improve its performance. This is typically the point where the model has achieved a balance between bias (underfitting) and variance (overfitting).

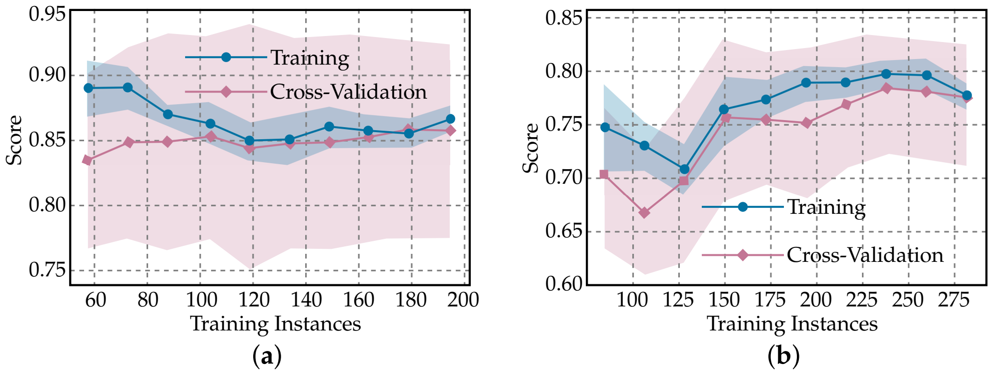


Figure 1. The learning curves from the model trained on the (**a**) 2C dataset, and (**b**) 3C dataset.

The confusion matrix (Figure 2) provided a breakdown of the model's predictions against actual outcomes to identify true positives, true negatives, false positives, and false negatives. For the 2C dataset, Figure 2a, the number 57 represents the True Positives, indicating the cases where the model correctly predicted that the patients have the condition. The number 6 represents the False Positives, which are the cases where the model incorrectly predicted that the patients have the condition, also known as a "Type I error". The number 7 represents the False Negatives, which are the cases where the model incorrectly predicted that the patients do not have the condition, but they actually do, also known as a "Type II error". Lastly, the number 23 represents the True Negatives, which means these are the cases where the model correctly predicted that the patients do not have the condition. Hence, the model correctly predicted 86.02% of the cases as either normal or abnormal.

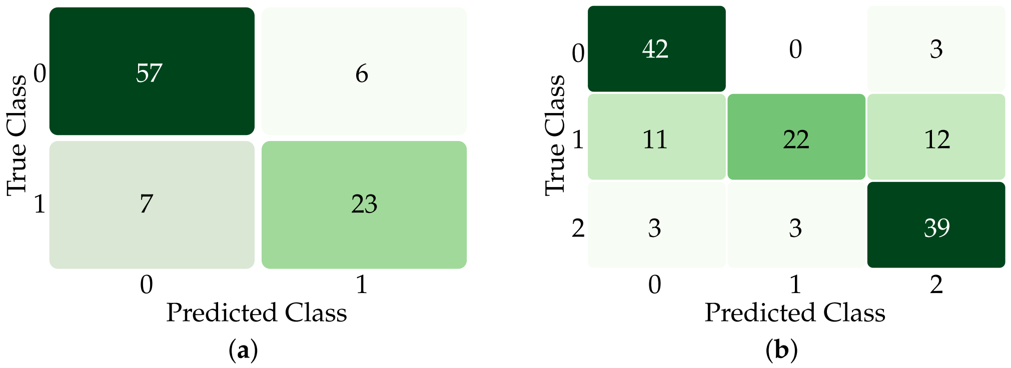
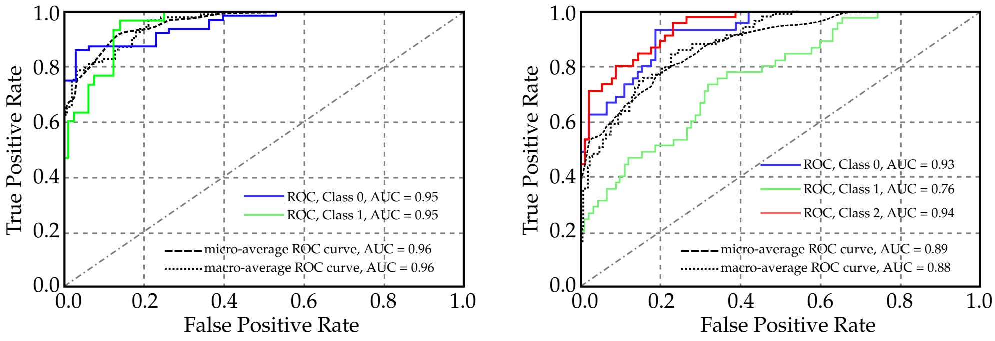


Figure 2. The confusion matrix for the model for the (**a**) 2C dataset (where 0: normal, 1: abnormal), and (**b**) 3C dataset (where 0: normal, 1: disc hernia, 2: spondylolisthesis).

For the 3C dataset, Figure 2b, the matrix represents the performance of a classification model (random forest classifier) on the 3C dataset. The rows represent the actual categories, while the columns represent the predicted categories. Both are ordered as Normal (0), Disc Hernia (1), and Spondylolisthesis (2). The first row represents the cases that are actually Normal. Out of these, 42 were correctly predicted as Normal (True Positives for Normal), none were incorrectly predicted as Disc Hernia (False Positives for Disc Hernia), and

3 were incorrectly predicted as Spondylolisthesis (False Positives for Spondylolisthesis). The second row represents the cases that are actually Disc Hernia. Out of these, 11 were incorrectly predicted as Normal (False Negatives for Disc Hernia), 22 were correctly predicted as Disc Hernia (True Positives for Disc Hernia), and 12 were incorrectly predicted as Spondylolisthesis (False Positives for Spondylolisthesis). The third row represents the cases that are actually Spondylolisthesis. Out of these, 3 were incorrectly predicted as Normal (False Negatives for Spondylolisthesis), 3 were incorrectly predicted as Disc Hernia (False Negatives for Spondylolisthesis), and 39 were correctly predicted as Spondylolisthesis (True Positives for Spondylolisthesis). Hence, the model correctly predicted 76.3% of the cases.

The Area Under the Curve (AUC) (Figure 3) graph displayes ROC curves and indicate the model's ability to distinguish between classes, with higher values representing better performance. The ROC curves are tools used in clinical medicine to evaluate the performance of diagnostic tests or predictive models. The ROC curve is a graphical representation of the sensitivity (true positive rate) versus 1-specificity (false positive rate) of a model across different threshold values. In a ROC curve, a perfect model would yield a point in the upper left corner of the ROC space, representing 100% sensitivity (no false negatives) and 100% specificity (no false positives). This would correspond to an AUC of 1. Conversely, a model no better than random guessing would yield a diagonal line from the bottom left to the top right corners, with an AUC of 0.5. The ROC curve can help in deciding the optimal cut-off point, which is the point closest to the upper left corner of the graph. This point represents the highest sensitivity and specificity.



(**a**) (**b**)

Figure 3. The AUC curve for the model for the (**a**) 2C dataset, and (**b**) 3C dataset.

The feature importance graphs (Figure 4) illustrates the weight of of different features in the model's decision-making process, and demonstrating which biomechanical markers were most influential in predicting lumbar IVD status. A feature importance plot is a graphical representation used in ML to illustrate the significance of different features in a predictive model. The importance of a feature is typically gauged by the increase in the model's prediction error after permuting the feature, a measure known as permutation importance. A higher permutation importance value signifies that the feature is more crucial for the model's prediction.

In the context of the 2C dataset, Figure 4a, the feature importance plot reveals that the feature 'degree spondylolisthesis', which is referred to as 'degree s-listhesis' in the graph, has the highest permutation importance value, approximately 0.22. This suggests that this feature is the most significant for the model's ability to differentiate between the classes in the 2C dataset. The features 'sacral slope' and 'pelvic radius' have the second lowest values, around 0.03, indicating their lesser importance. The features 'pelvic tilt' and 'pelvic incidence' have almost zero importance, implying they do not significantly contribute to the model's predictions. Interestingly, the feature 'lumbar lordosis angle' has a negative importance value of around −0.02.

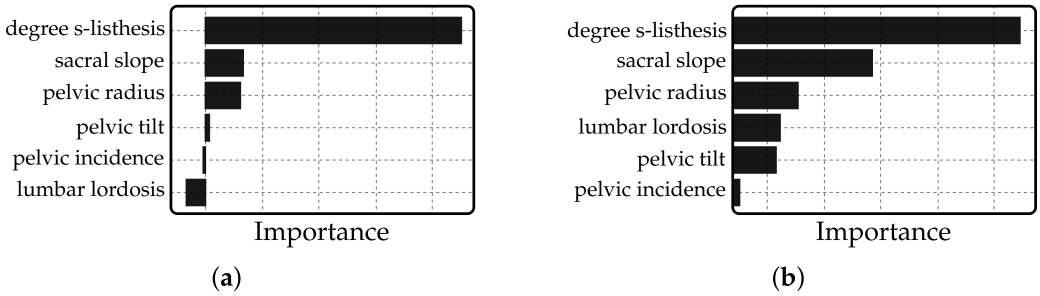


Figure 4. The feature importance graph for the model for the (**a**) 2C dataset, and (**b**) 3C dataset.

From a technical perspective, a negative permutation importance value implies that permuting the feature actually enhances the model's performance. This could occur if the feature is noisy or misleading, leading the model to make more errors when it relies on this feature. From a clinical standpoint, this could suggest that the feature 'lumbar lordosis angle' is not a reliable indicator for the condition being predicted in the 2C dataset.

For the 3C dataset, Figure 4b, the feature importance plot shows a different pattern. The feature 'degree spondylolisthesis' still has the highest importance value of around 0.41, but the importance of the other features has increased compared to the 2C dataset. This indicates that these features are more useful for distinguishing between the three classes in the 3C dataset.

The difference between the 2C and 3C datasets likely reflects the difference in the complexity of the prediction task. The 2C dataset involves a binary classification task, while the 3C dataset involves a multi-class classification task. Technically, this means that the model trained on the 3C dataset needs to rely on more features to distinguish between the three classes. Clinically, this could suggest that a broader range of indicators is needed to accurately diagnose the three conditions represented in the 3C dataset.

4. Discussion

Important insights into the performance of the models utilized were provided by the learning curves for both the 2C and 3C datasets (Figure 1). The evolution of the models' performance with exposure to more training instances was illustrated by the curves. A better performance on the training data was initially exhibited by the models, as indicated by the higher starting point of the training curve. However, a quick generalization to unseen data was learned by the models as they were trained on more data, leading to the convergence of the training and cross-validation curves. The convergence, followed by a parallel rise and eventual plateau of both curves, suggested that a balance between bias and variance was achieved by the models, with performance not significantly improved by additional training instances. Clinically, an effective classification of patients as normal or abnormal (for the 2C data) and into the three categories of Normal, Disc Hernia, and Spondylolisthesis (for the 3C data) could be suggested by these learning curves. However, it is important to note that a good performance on the training and validation data was shown by the models, but their performance in a real-world clinical setting may be influenced by factors such as the quality and representativeness of the data they were trained on.

In the context of clinical implications, an abnormal result in the model with 2C dataset indicates the presence of either Disc Hernia or Spondylolisthesis. This means that the model's predictions can be instrumental in diagnosing these conditions, thereby facilitating early intervention and treatment. However, the presence of False Positives and False Negatives, as depicted in Figure 2a, also highlights the need for further refinement of the model to minimize these errors, as they could potentially lead to misdiagnosis or missed diagnosis. On the other hand, the confusion matrix for the 3C dataset, as depicted in Figure 2b, provides insights into the performance of the random forest classifier model in predicting the conditions of patients-Normal, Disc Hernia, and Spondylolisthesis. The model correctly identified Normal cases 42 times, which means it was successful in distinguishing patients without any conditions from those with Disc Hernia or Spondylolisthesis. However, it incorrectly predicted 3 Normal cases as Spondylolisthesis, which could lead to unnecessary treatment for those patients. For Disc Hernia, the model correctly identified 22 cases but also incorrectly predicted 11 cases as Normal and 12 cases as Spondylolisthesis. This could potentially lead to missed or incorrect treatment, which could have serious health implications for the patients. In the case of Spondylolisthesis, the model correctly identified 39 cases, but it also incorrectly predicted 3 cases as Normal and another 3 cases as Disc Hernia. This could lead to missed diagnosis or incorrect treatment. Overall, the model correctly predicted 76.3% of the cases. While this is a relatively high accuracy, the misclassifications, particularly the False Negatives, could have serious implications in a clinical setting, as they represent missed diagnoses. Therefore, while the model provides a good starting point, it is important to continue refining it to reduce these errors and improve patient care.

The results highlight the role of feature importance in machine learning models, particularly in predicting lumbar IVD status. Our findings underscore the significance of the 'degree spondylolisthesis' feature, which consistently showed the highest permutation importance value in both the 2C and 3C datasets. This suggests that 'degree spondylolisthesis' is a key biomechanical marker in differentiating between the classes in these datasets. Interestingly, the 'lumbar lordosis angle' feature demonstrated a negative importance value in the 2C dataset, indicating that it may not be a reliable indicator for the condition being predicted. This could have significant implications for clinical practice, as it suggests that reliance on this feature could potentially lead to inaccurate predictions. The proposed approach in this study relies heavily on the 'degree spondylolisthesis' feature, based on the data available in the public dataset. It is recognized that the degree of spondylolisthesis may not be a reliable indicator for all types of lumbar IVD abnormalities. For instance, in cases of degenerative spondylolisthesis, other factors such as facet joint degeneration and disc collapse may play a more significant role. Similarly, in cases of isthmic spondylolisthesis, the pars defect may be the primary contributor to the development of the condition. However, due to the constraints of the dataset and the scope of this study, a comprehensive evaluation of these factors was not possible. This study acknowledges its limitations and suggests that future research could benefit from a more nuanced approach that considers these additional factors. This would allow for a more comprehensive understanding of lumbar IVD abnormalities and could potentially lead to more accurate diagnostic tools and treatments.

In comparing the 2C and 3C datasets, this study revealed that the complexity of the prediction task-binary classification in the 2C dataset and multi-class classification in the 3C dataset-likely influences the importance of different features. The increased importance of features in the 3C dataset suggests that a broader range of indicators is needed to accurately diagnose the three conditions represented in this dataset. This finding could guide future research and clinical practice in the field of orthopedics, emphasizing the need to consider a wide range of biomechanical markers in diagnosing various conditions.

While our findings underscore the significant potential of machine learning models for predicting lumbar IVD status, limitations exist and there are opportunities for further refinement. The study relied on six biomechanical features from the Kaggle dataset, yet critical markers for diagnosing lumbar IVD abnormalities were not comprehensively evaluated. This is shown by the questionable reliability of the 'lumbar lordosis angle', indicating a need for further investigation into other biomechanical markers that may be more predictive. In this study, we have focused on the available biomechanical markers provided in the public dataset used for our research. We acknowledge that there are other significant markers such as facet joint orientation, vertebral endplate morphology, and disc degeneration that play a crucial role in the development of lumbar IVD abnormalities. However, these markers were not included in the dataset, limiting our ability to incorporate them into our analysis. We recognize that this limitation may lead to incomplete or inaccurate predictions, particularly in cases where these unconsidered markers are the primary contributors to the development of the condition. Future research could benefit from a more comprehensive dataset that includes these additional markers. The use of data from an online repository, such as this, raises concerns about data quality and representativeness, as machine learning models are only as reliable as the data on which they are trained [21,22]. This study primarily investigates the impact of specific biomechanical markers on lumbar IVD health, based on the available dataset. It is recognized that factors such as age, sex, body mass index, and physical activity level significantly influence the development and progression of lumbar IVD abnormalities. However, due to the scope of this study and the constraints of the dataset, these complex interactions were not accounted for. In cases of patients with complex spinal deformities or multiple levels of spinal degeneration, a more comprehensive evaluation of biomechanical markers and other factors may be necessary for accurate prediction. This study acknowledges its limitations and suggests that future research could benefit from a more nuanced approach that considers the complex interactions between multiple biomechanical markers and other factors.

Furthermore, the proposed approach in this study does not account for the dynamic nature of lumbar IVD health, which can change over time due to various factors such as degeneration, trauma, or surgical intervention. The models produced in this study were designed to predict the state of the lumbar IVDs at the time of data collection. It is recognized that any further degeneration or progression of the disease since then cannot be accurately reported, especially for patients who've suffered significant degeneration since then, which is not uncommon for IVD pathologies. This limitation is due to the constraints of the dataset and the scope of this study. Future research could benefit from a more dynamic model that can account for changes in lumbar IVD health over time.

Another key area for improvement is the reduction of false positives and false negatives, which directly effects diagnostic accuracy [23]. In a clinical setting, missed diagnoses can lead to suboptimal patient outcomes, while misdiagnoses may result in unnecessary follow-up visits, costly tests, and unnecessary invasive procedures that expose the pt to unnecessary risk [24,25]. Even with true positives, the application of models risks identifying IVD abnormalities that lack clinical relevance. This issue is particularly pertinent for IVD conditions, as many abnormalities are asymptomatic and are often discovered incidentally on MRI [26,27]. Such incidental findings can lead to unwarranted treatments and unnecessary healthcare expenditures [28]. While machine learning models offer promising avenues for advancing orthopedic diagnostics, ongoing refinement, validation, and the use of higher-quality data are essential to ensure their efficacy, accuracy, and clinical applicability.

5. Conclusions

The study provides an examination of the role of ML models in the categorization of orthopedic patients. These models, demonstrating an accuracy reaching 90.83% and a precision peaking at 91.86%, have shown their effectiveness in the diagnosis of IVD pathology. The research also offers useful insights into the biomechanical alterations associated with the pathogenesis of lumbar IVD abnormalities. The results indicate that an ensemble of Random Forest and Logistic Regression models is most effective for the 2C classification, while the Extra Trees classifier is superior for the 3C classification. Despite the encouraging outcomes, the study recognizes the limitations of AI implementation in clinical environments and identifies potential areas for future enhancement to develop more precise and informative models. This study makes a contribution to the ongoing efforts to employ data-driven machine learning models to enhance patient outcomes in orthopedic care.

Author Contributions: All authors (D.N. (Daniel Nasef), D.N. (Demarcus Nasef), V.S., P.G. and M.T.) contributed to all aspects of this work. This includes but is not limited to conceptualization, methodology, software management, validation, formal analysis, investigation, resources management, data curation, writing—original draft preparation, writing—review and editing, and visualization. The supervision, and project administration were handled by M.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
IVD	Lumbar Intervertebral Disc
CT	Computed Tomography
THP	Total Hip Arthroplasty
LBP	Low Back Pain
CNN	Convolutional Neural Network
MRI	Magnetic Resonance Imaging
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic

References

1. Adams, M.A.; Roughley, P.J. What is Intervertebral Disc Degeneration, and What Causes It? *Spine* **2006**, *31*, 2151–2161. [CrossRef] [PubMed]
2. Jordon, J.; Konstantinou, K.; O'Dowd, J. Herniated lumbar disc. *BMJ Clin. Evid.* **2009**, *2009*, 1118.
3. Qaraghli, M.I.A.; Jesus, O.D. Lumbar Disc Herniation. In *StatPearls [Internet]*; StatPearls Publishing: Treasure Island, FL, USA, 2024.
4. Wiltse, L.L. The Etiology of Spondylolisthesis. *J. Bone Jt. Surg.* **1962**, *44*, 539–560. [CrossRef]
5. Bydon, M.; Alvi, M.A.; Goyal, A. Degenerative Lumbar Spondylolisthesis. *Neurosurg. Clin. N. Am.* **2019**, *30*, 299–304. [CrossRef] [PubMed]
6. Ganju, A. Isthmic spondylolisthesis. *Neurosurg. Focus* **2002**, *13*, 1–6. [CrossRef]
7. Dydyk, A.M.; Massa, R.N.; Mesfin, F.B. Disc Herniation. In *StatPearls [Internet]*; StatPearls Publishing: Treasure Island, FL, USA, 2023.
8. Bednar, D.A. Cauda equina syndrome from lumbar disc herniation. *Can. Med. Assoc. J.* **2015**, *188*, 284. [CrossRef]
9. KD, W. Spondylolisthesis. In *Campbell's Operative Orthopaedics*, 14th ed.; Elsevier: Philadelphia, PA, USA, 2021; Chapter 40.
10. Tenny, S.; Hanna, A.; Gillis, C.C. Spondylolisthesis. In *StatPearls [Internet]*; StatPearls Publishing: Treasure Island, FL, USA, 2024.
11. Padash, S.; Mickley, J.P.; Vera Garcia, D.V.; Nugen, F.; Khosravi, B.; Erickson, B.J.; Wyles, C.C.; Taunton, M.J. An Overview of Machine Learning in Orthopedic Surgery: An Educational Paper. *J. Arthroplast.* **2023**, *38*, 1938–1942. [CrossRef]
12. Toma, M.; Wei, O.C. Predictive Modeling in Medicine. *Encyclopedia* **2023**, *3*, 590–601. [CrossRef]
13. Rouzrokh, P.; Wyles, C.C.; Philbrick, K.A.; Ramazanian, T.; Weston, A.D.; Cai, J.C.; Taunton, M.J.; Lewallen, D.G.; Berry, D.J.; Erickson, B.J.; et al. A Deep Learning Tool for Automated Radiographic Measurement of Acetabular Component Inclination and Version After Total Hip Arthroplasty. *J. Arthroplast.* **2021**, *36*, 2510–2517.e6. [CrossRef]

14. Rouzrokh, P.; Ramazanian, T.; Wyles, C.C.; Philbrick, K.A.; Cai, J.C.; Taunton, M.J.; Maradit Kremers, H.; Lewallen, D.G.; Erickson, B.J. Deep Learning Artificial Intelligence Model for Assessment of Hip Dislocation Risk Following Primary Total Hip Arthroplasty From Postoperative Radiographs. *J. Arthroplast.* **2021**, *36*, 2197–2203.e3. [CrossRef]
15. D'Antoni, F.; Russo, F.; Ambrosio, L.; Bacco, L.; Vollero, L.; Vadalà, G.; Merone, M.; Papalia, R.; Denaro, V. Artificial Intelligence and Computer Aided Diagnosis in Chronic Low Back Pain: A Systematic Review. *Int. J. Environ. Res. Public Health* **2022**, *19*, 5971. [CrossRef] [PubMed]
16. Wang, P.; Zhang, Z.; Xie, Z.; Liu, L.; Ren, G.; Guo, Z.; Xu, L.; Yin, X.; Hu, Y.; Wang, Y.; et al. Natural Language Processing-Driven Artificial Intelligence Models for the Diagnosis of Lumbar Disc Herniation with L5 and S1 Radiculopathy: A Preliminary Evaluation. *World Neurosurg.* **2024**, *189*, e300–e309. [CrossRef] [PubMed]
17. Wirries, A.; Geiger, F.; Hammad, A.; Bäumlein, M.; Schmeller, J.N.; Blümcke, I.; Jabari, S. AI Prediction of Neuropathic Pain after Lumbar Disc Herniation-Machine Learning Reveals Influencing Factors. *Biomedicines* **2022**, *10*, 1319. [CrossRef] [PubMed]
18. Fan, X.; Qiao, X.; Wang, Z.; Jiang, L.; Liu, Y.; Sun, Q. Artificial Intelligence-Based CT Imaging on Diagnosis of Patients with Lumbar Disc Herniation by Scalpel Treatment. *Comput. Intell. Neurosci.* **2022**, *2022*, 3688630. [CrossRef]
19. Bekbolatova, M.; Mayer, J.; Ong, C.W.; Toma, M. Transformative Potential of AI in Healthcare: Definitions, Applications, and Navigating the Ethical Landscape and Public Perspectives. *Healthcare* **2024**, *12*, 125. [CrossRef]
20. Crawford, C. Biomechanical Features of Orthopedic Patients. Available online: https://www.kaggle.com/datasets/uciml/biomechanical-features-of-orthopedic-patients (accessed on 9 September 2024).
21. Chen, H.; Chen, J.; Ding, J. Data Evaluation and Enhancement for Quality Improvement of Machine Learning. *IEEE Trans. Reliab.* **2021**, *70*, 831–847. [CrossRef]
22. Siebert, J.; Joeckel, L.; Heidrich, J.; Trendowicz, A.; Nakamichi, K.; Ohashi, K.; Namba, I.; Yamamoto, R.; Aoyama, M. Construction of a quality model for machine learning systems. *Softw. Qual. J.* **2022**, *30*, 307–335. [CrossRef]
23. Murphy, P.; Knight, S. Misdiagnosis in Sports Medicine. *Curr. Sports Med. Rep.* **2002**, *1*, 333–337. [CrossRef]
24. Jawad, B.N.; Pedersen, K.Z.; Andersen, O.; Meier, N. Minimizing the Risk of Diagnostic Errors in Acute Care for Older Adults: An Interdisciplinary Patient Safety Challenge. *Healthcare* **2024**, *12*, 1842. [CrossRef]
25. Ahn, Y.; Hong, G.S.; Park, K.J.; Lee, C.W.; Lee, J.H.; Kim, S.O. Impact of diagnostic errors on adverse outcomes: Learning from emergency department revisits with repeat CT or MRI. *Insights Into Imaging* **2021**, *12*, 160. [CrossRef]
26. Dora, C.; Wälchli, B.; Elfering, A.; Gal, I.; Weishaupt, D.; Boos, N. The significance of spinal canal dimensions in discriminating symptomatic from asymptomatic disc herniations. *Eur. Spine J.* **2002**, *11*, 575–581. [CrossRef]
27. Donnally, C.J., III; Hanna, A.; Varacallo, M. *Lumbar Degenerative Disk Disease*; StatPearls Publishing: Treasure Island, FL, USA, 2024.
28. Epstein, N.E.; Hood, D.C. Unnecessary spinal surgery: A prospective 1-year study of one surgeon's experience. *Surg. Neurol. Int.* **2011**, *2*, 83. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.