




Predicting the Risk of Lumbar Prolapsed Disc: A Gene Signature-Based Machine Learning Analysis

Fengfeng Wang · Fei Meng · Stanley Sau Ching Wong 

Received: February 21, 2025 / Accepted: April 24, 2025 / Published online: May 4, 2025
© The Author(s) 2025

ABSTRACT

Introduction: Lumbar prolapsed disc (LPD) is a leading cause of low back pain, contributing significantly to global disability and healthcare burden. This study aimed to develop machine learning models to predict the risk of LPD by analysing gene expression profiles for early detection.

Methods: Transcriptomic data from peripheral blood samples were obtained from the Gene Expression Omnibus (GEO) database, with dataset GSE150408 used for training and GSE124272 for testing. The training dataset included 17 patients with sciatica resulting from LPD, all of whom had magnetic resonance imaging confirmation of single-level LPD at either the L4/5 or L5/S1 levels. Data from 17 healthy volunteers were used as controls. Recursive feature elimination (RFE) was employed to identify the most

relevant gene signatures among 23 pain-related genes. Machine learning models, including support vector machine (SVM), random forest, *k*-nearest neighbours (KNN), logistic regression, and Extreme Gradient Boosting (XGBoost), were trained and evaluated. Model performance was assessed using accuracy, area under the curve (AUC), F1 score, and Matthews correlation coefficient (MCC).

Results: Eight key gene signatures were identified as significant predictors of LPD, with *MMP9* exhibiting the highest importance score. Most of these genes were differentially expressed between patients with LPD and healthy controls ($p < 0.05$). Among the models, random forest demonstrated the highest accuracy (0.80, 95% CI 0.73–0.85) and MCC (0.64, 95% CI 0.53–0.76), followed by KNN, XGBoost, and SVM. Overall, the random forest model exhibited the most robust performance in predicting the risk of LPD.

Conclusion: The results of our study suggest that machine learning models based on pain-related gene signatures may identify patients at high risk of developing LPD with reasonably high accuracy. These prediction models could perhaps be integrated into clinical diagnostic tools to enhance early diagnosis and prevention.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s40122-025-00744-4>.

F. Wang · F. Meng · S. S. C. Wong (✉)
Department of Anaesthesiology, School of Clinical
Medicine, Li Ka Shing Faculty of Medicine, The
University of Hong Kong, Queen Mary Hospital,
Room 424, Block K102 Pokfulam Road, Hong Kong,
China
e-mail: wongstan@hku.hk

Keywords: Low back pain; Lumbar prolapsed disc; Gene signature; Transcriptomics; Machine learning; Risk prediction; Early prevention

Key Summary Points

Why carry out this study?

Lumbar prolapsed disc (LPD) is a leading cause of low back pain, contributing significantly to global disability and healthcare burden.

Current diagnostic methods are not sufficiently advanced to identify individuals at risk of developing LPD early, limiting the use of preventive strategies and personalized treatment plans.

This study aimed to address the unmet need for predictive clinical tools by applying transcriptomic data from peripheral blood and machine learning models to assess LPD risk and identify critical gene signatures.

What was learned from the study?

Key gene signatures associated with LPD diagnosis were identified, and machine learning models (e.g. random forest) demonstrated high predictive accuracy.

Transcriptomic data from peripheral blood, combined with machine learning techniques, showed potential for identifying high-risk individuals for LPD.

These findings may be used to develop clinical tools to aid physicians in early diagnosis, risk stratification, and implementation of preventive strategies.

INTRODUCTION

Lumbar prolapsed disc (LPD), also known as lumbar disc herniation, is a common cause of low back pain and radicular leg pain [1–3]. Approximately 39% of individuals experiencing low back pain suffer from disc-related issues [4]. LPD is a prevalent health issue that significantly impacts the quality of life for those affected, leading to substantial pain, disability, and a reduction in daily functioning [5, 6]. LPD occurs

when the nucleus pulposus of an intervertebral disc in the lumbar spine herniates through a tear in the annulus fibrosis. This condition results in inflammation and irritation of the surrounding nerves, leading to pain, discomfort, and neurological symptoms [7, 8]. The management of LPD includes conservative approaches (e.g. pharmacological analgesic medication, physical therapy), interventional pain procedures (e.g. epidural steroid injection, regenerative medicine), and surgical procedures (e.g. microdiscectomy) [7–10]. Conservative treatment typically serves as the first-line approach [11], while interventional pain procedures and surgical interventions are generally reserved for patients refractory to initial treatment [12]. However, current approaches to early diagnosis and preventive measures for LPD remain insufficient.

Machine learning has emerged as a powerful tool for identifying patterns associated with different diseases [13–15]. It was reported that machine learning models based on patient demographic and clinical characteristics were developed and validated to predict treatment outcomes related to disability and pain 1 year after lumbar disc herniation surgery [16]. However, the application of machine learning models based on gene signatures to predict the occurrence of LPD for diagnosis has not been investigated. Current diagnostic methods do not help identify individuals at risk of developing LPD, which limits the implementation of preventive strategies that can improve patient outcomes. There is an unmet need for predictive tools that can accurately assess LPD risk. Recent research has highlighted the potential role of genes involved in pain perception, signalling, and psychological processing in the susceptibility and severity of LPD [17]. These genetic factors may influence individual responses to tissue injury and inflammation [17]. A better understanding of the genetic basis of LPD could lead to personalized diagnostic approaches and targeted prevention strategies, ultimately improving patient outcomes and reducing the burden of this debilitating condition.

The objective of the study was to develop and validate machine learning models based on key pain-related gene signatures for the

identification of patients at elevated risk of LPD. We hypothesize that transcriptomic data derived from peripheral blood, when analysed using advanced machine learning models, can identify key gene signatures and predict the risk of LPD with reasonably high accuracy. Our findings could potentially lead to the development of clinical predictive tools for practical use.

METHODS

Data Source

Transcriptomic data from the peripheral blood of patients with LPD were obtained from the publicly available Gene Expression Omnibus (GEO) database (GSE150408) to train the models [18]. LPD falls under the ICD-10-CM code M51 [19]. The inclusion criteria of the dataset were patients with sciatica and LPD confirmed by magnetic resonance imaging (MRI) imaging at either the L4/5 or L5/S1 levels. The dataset used excluded individuals with other neuropathies, spinal diseases, infections, rheumatic conditions, cardiovascular or metabolic diseases, dementia, mental health disorders, a history of surgery, congenital conditions, tuberculosis, or tumours [18]. The training dataset included 17 patients and 17 healthy volunteers. To evaluate the performance of the constructed models, we used an independent testing dataset, GSE124272, consisting of eight patients and eight healthy controls [20].

Ethical Approval

This research utilized existing studies and did not involve any new experiments with human participants or animals conducted by the authors. For model training, we obtained transcriptomic data from patients with LPD using the GEO database entry GSE150408, and for model testing, we used data from GSE124272. All procedures in these referenced studies involving human participants were approved by the Ethics Committee of the Sichuan Provincial Orthopedic Hospital.

Pain-Related Genes

In developing the models, we focused on 23 genes that have been previously reported to influence the risk of chronic back pain and widespread pain syndromes (Table S1) [21]. These 23 genetic factors are involved in intervertebral disc stability, inflammation, and pain signalling, which are important in LPD and pain [22–24]. The 23 pain-related genes are grouped into three categories as follows [21]: (i) Genes affecting intervertebral disc stability: *COL9A2* (collagen type IX alpha 2 chain), *COL9A3* (collagen type IX alpha 3 chain), *COL11A1* (collagen type XI alpha 1 chain), *COL11A2* (collagen type XI alpha 2 chain), *COL1A1* (collagen type I alpha 1 chain), *ACAN* (aggrecan), *CILP* (cartilage intermediate layer protein), *VDR* (vitamin D receptor), *MMP3* (matrix metalloproteinase 3), *MMP9* (matrix metalloproteinase 9), *THBS2* (thrombospondin 2); (ii) Genes influencing inflammation: *IL1RN* (interleukin-1 receptor antagonist), *IL1A* (interleukin-1 alpha), *IL1B* (interleukin-1 beta), *IL6* (interleukin-6); and (iii) Genes involved in pain signalling: *GCH1* (GTP cyclohydrolase 1), *COMT* (catechol-O-methyltransferase), *OPRM1* (opioid receptor mu 1), *OPRD1* (opioid receptor delta 1), *MC1R* (melanocortin 1 receptor), *TRPV1* (transient receptor potential cation channel subfamily V member 1), *TRPA1* (transient receptor potential cation channel subfamily A member 1), *FAAH* (fatty acid amide hydrolase).

Feature Selection

As a result of the limited number of samples and the high dimensionality of the features, we first generated 10 sets of simulated gene expression data by adding Gaussian noise to the original training and testing datasets. This approach can ensure stability in parameter estimation and reduces the risk of overfitting [25]. Gaussian noise was added to create simulated data that mimic the inherent variability of the original data, while preserving the overall structure and relationships between variables.

Both the original and simulated datasets were used to train and test the models. To identify the most relevant features for our predictive model, we employed recursive feature elimination (RFE) using leave-one-out cross-validation (LOOCV) as the resampling method [26, 27]. RFE is a feature selection technique that recursively removes features from the dataset and evaluates the model's performance to identify the most important features. The process began with all features in the dataset and fit the model. The least important feature, as determined by the model's feature importance scores, was then removed. The model was refitted using the remaining features, and the process was repeated until a specified number of features was retained. In this study, RFE with LOOCV was used to estimate the model's performance for each feature subset. LOOCV is a cross-validation technique where one observation is used as the validation set, and the remaining observations form the training set. This process was repeated for each observation in the dataset, providing a better estimate of the model's performance [28]. The optimal set of features was selected on the basis of the highest cross-validated performance. Importance scores were further extracted from the RFE process using the `varImp` function, which determines the relevance of each feature on the basis of the model's internal criteria. This approach allowed one to identify and rank the most relevant genes contributing to the predictive power of the model.

Model Construction

In this study, multiple machine learning models were constructed on the basis of the selected gene signatures. To improve prediction performance, hyperparameter tuning was conducted using a grid search approach with fivefold cross-validation. For the support vector machine (SVM) algorithm with a radial basis function (RBF) kernel, the grid search was performed within the range of $2^{(-5:15)}$ for C and $2^{(-15:3)}$ for σ to determine the optimal parameters. The random forest model was customized to enhance performance by tuning the number of variables randomly

sampled as candidates at each split (`mtry`), the number of trees in the forest (`ntree`), and the minimum size of terminal nodes (`nodesize`): `mtry` from 1 to 8, `ntree` from 100 to 1000, and `nodesize` from 1 to 20. For the k -nearest neighbours (KNN) model, the tuning process involved exploring values for k from 1 to 100 to identify the optimal number of neighbours. The "optimal" kernel was selected to improve the weighting scheme for the neighbours' contributions, ensuring that closer neighbours have a more significant influence on the classification decision. The distance metric used was set to 2, corresponding to the Euclidean distance. In the logistic regression model, the family parameter was set to binomial to specify logistic regression, as the target variable was binary. In the XGBoost model, the learning rate (`eta`) was set to 0.01, 0.1, and 0.3 to control the boosting step size. The maximum depth of trees (`max_depth`) was set to 3, 6, and 9 to determine tree depth, while the minimum sum of instance weights (`min_child_weight`) was adjusted to 1, 3, and 5 to regulate the data required at child nodes. The subsample ratio of training data (`subsample`) and features (`colsample_bytree`) were tested at 0.5, 0.8, and 1 to prevent overfitting. The minimum loss reduction (`gamma`) was set to 0, 1, and 3. The training process involved 1000 boosting iterations with early stopping after 10 rounds to monitor and prevent overfitting.

Performance Evaluation

Each machine learning model's performance was assessed on the basis of four performance metrics: accuracy, F1 score, Matthews correlation coefficient (MCC), and area under the curve (AUC). The true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values were used to calculate these metrics.

Accuracy measures the proportion of correct predictions (true positives and true negatives) out of the total number of predictions. It is a widely used metric for classification problems but can be misleading in the case of imbalanced datasets.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

F1 score is the harmonic mean of precision and recall, which combines both false positives

and false negatives into a single metric. It is particularly useful when dealing with imbalanced datasets where the positive class is rare.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{F1 score} = 2 \times (\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$$

MCC is a metric that takes into account all four values of the confusion matrix (true positives, true negatives, false positives, and false negatives) and provides a balanced measure of classification performance. It ranges from -1 to 1 , with 1 indicating perfect classification, 0 indicating random classification, and -1 indicating complete misclassification.

$$\text{MCC} = (\text{TP} \times \text{TN} - \text{FP} \times \text{FN}) / \sqrt{((\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN}))}$$

The AUC was also calculated to evaluate the classification models. The receiver operating characteristics (ROC) curve was generated by plotting the true positive rate against the false positive rate, representing the trade-off between sensitivity and specificity. The AUC provides a single scalar value that represents the overall performance of the model, with higher values indicating better performance. The AUC ranges from 0 to 1 , with 1 being the best possible score.

Statistical Analysis

All statistical analyses were performed using R version 4.4.0 (2024, URL <https://www.r-project.org>). The feature selection was conducted using the RFE method, hyperparameter tuning, and cross-validation using the caret package [29]. The e1071 package was used to train the SVM model [30], the randomForest package was used for the random forest model [31], and the class package was used for the KNN model [32]. The function glm was used to train the logistic regression model [33]. The XGBoost model was trained using the xgboost package [34]. We used t test or Mann–Whitney U test to compare the expression levels of selected gene

signatures between two groups, with significance determined by a p value less than 0.05 . Benjamini–Hochberg false discovery rates (FDRs) were applied to correct for multiple hypotheses [35].

RESULTS

Gene Signature Selection

The expression profiles of 23 pain-related genes were extracted from both the training and testing datasets. A flowchart illustrating the entire process is presented in Fig. 1. We generated 10 sets of simulated gene expression data by introducing Gaussian noise with a standard deviation of 0.1 to the original data. To visualize the expression values of the 23 genes in both the original and simulated data for the training and testing datasets, we created a boxplot (Fig. 2a, b). The boxplot revealed that the distributions for the original and simulated data exhibited similar patterns. The addition of Gaussian noise with a standard deviation of 0.1 did not significantly alter the expression values of the 23 genes, indicating that our simulated data closely resembled real-world variations in gene expression.

The RFE technique was employed to identify the most critical features contributing to the predictive performance of the models. Eight genes were selected as important gene signatures: *MMP9*, *IL6*, *ACAN*, *IL1RN*, *MMP3*, *THBS2*, *COL11A2*, and *CILP*. The feature importance scores for each gene in the training dataset were plotted (Fig. 2c). Among all the eight genes, *MMP9* had the highest importance score, indicating that it played the most crucial role in distinguishing patients with LPD from the control group.

The differences in the expression of these eight gene signatures between individuals with LPD and those without LPD were investigated (Fig. 3a, b). In the training dataset, *ACAN*, *MMP3*, *MMP9*, and *IL1RN* were significantly upregulated in the LPD group compared to the control group (adjusted p values < 0.05). Conversely, *COL11A2*, *THBS2*, and *IL6* were significantly downregulated in the LPD group (adjusted p values < 0.05). In the testing dataset,

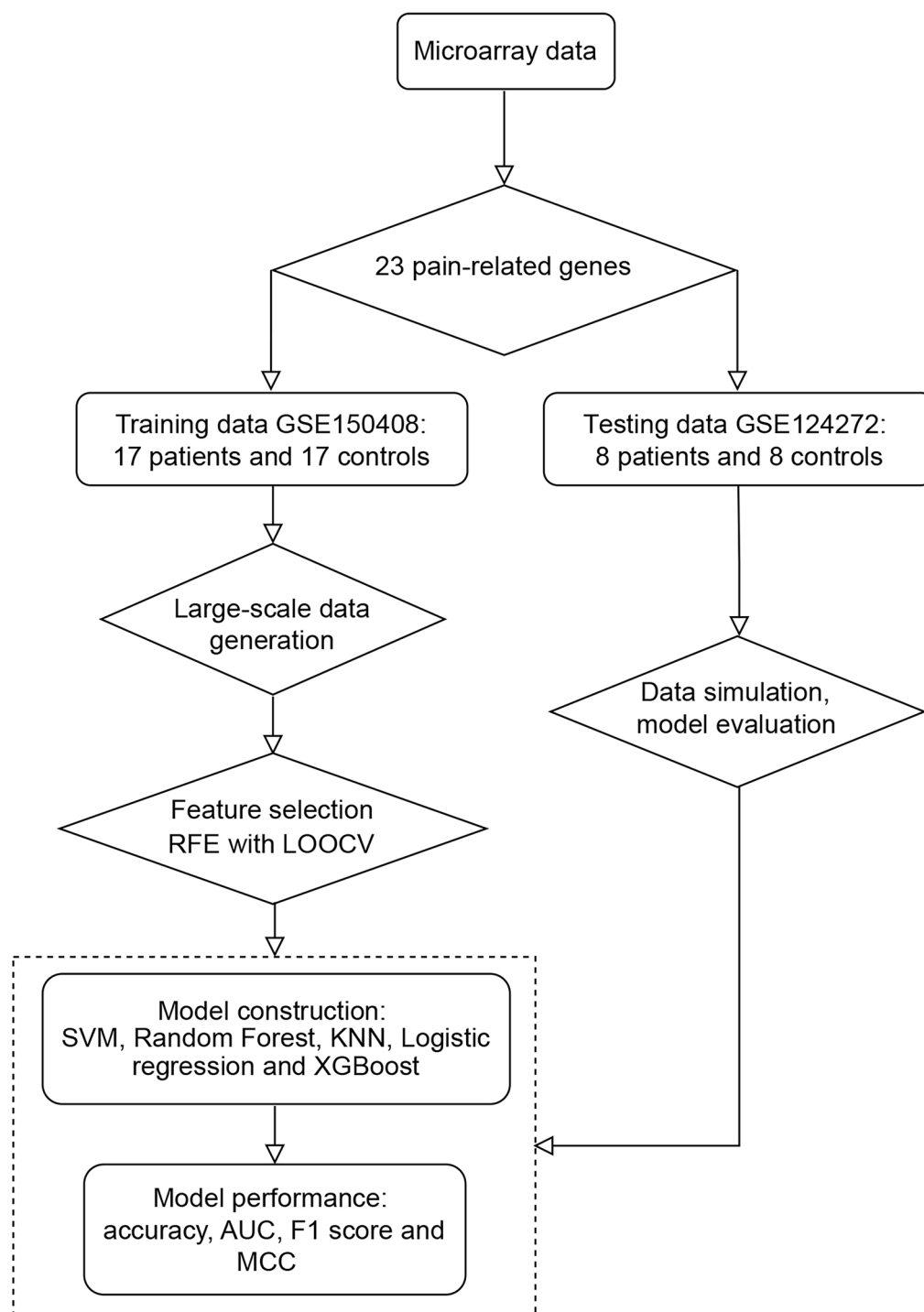


Fig. 1 Workflow of prediction model construction. *RFE* recursive feature elimination, *LOOCV* leave-one-out cross-validation, *SVM* support vector machine, *KNN* *k*-nearest

neighbours, *XGBoost* Extreme Gradient Boosting, *AUC* area under the curve, *MCC* Matthews correlation coefficient

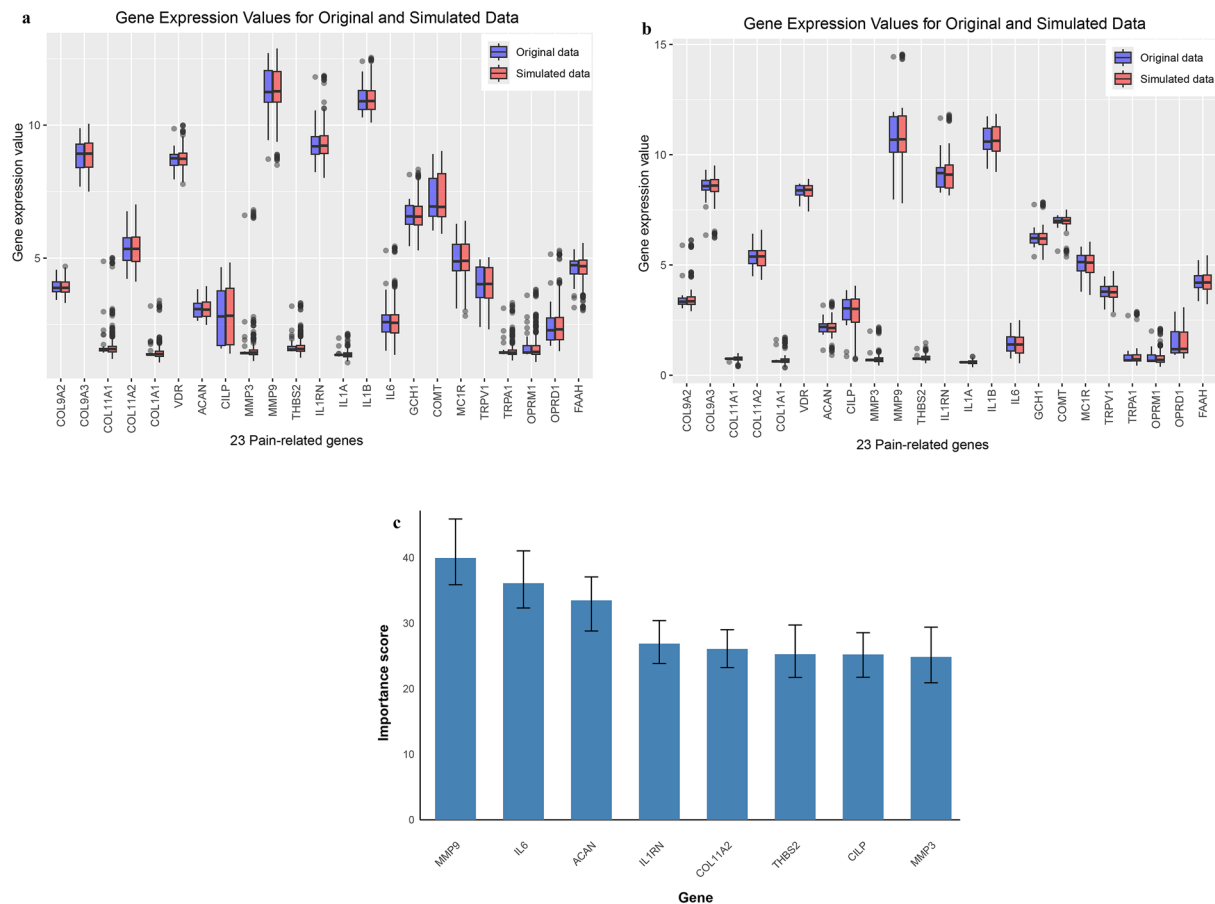


Fig. 2 Candidate gene features. **a** Boxplot of gene expression values for 23 genes in the original (purple) and 10 simulated (red) datasets in the training dataset. The x-axis represents the gene names, and the y-axis shows the expression values. The lower and upper bounds of the boxes represent the first quartile (Q1) and the third quartile (Q3) of

the data, respectively. The horizontal line inside each box represents the median (Q2) of the data. The grey dots represent outliers that fall outside the whiskers of the boxplot. **b** Boxplot of gene expression values in the testing dataset. **c** Feature importance of 8 selected gene features with 95% confidence intervals

COL11A2, *ACAN*, *MMP3*, *MMP9*, and *IL1RN* were significantly upregulated in the LPD group compared to the control group (adjusted p values < 0.05), while *CILP* was downregulated in the LPD group (adjusted p value < 0.05). Additionally, the expression levels of the genes and their correlation with the LPD group were analysed (Fig. 3c, d). The results revealed a strong association between the expression levels of the eight gene signatures and the LPD group.

Model Performance and Comparisons

The random forest model achieved the highest accuracy, with a value of 0.80 (95% CI 0.73–0.85). The KNN model followed with an accuracy of 0.71 (95% CI 0.64–0.77), while the XGBoost model had an accuracy of 0.69 (95% CI 0.62–0.75). The SVM model achieved an accuracy of 0.61 (95% CI 0.54–0.68), and the logistic regression model had the

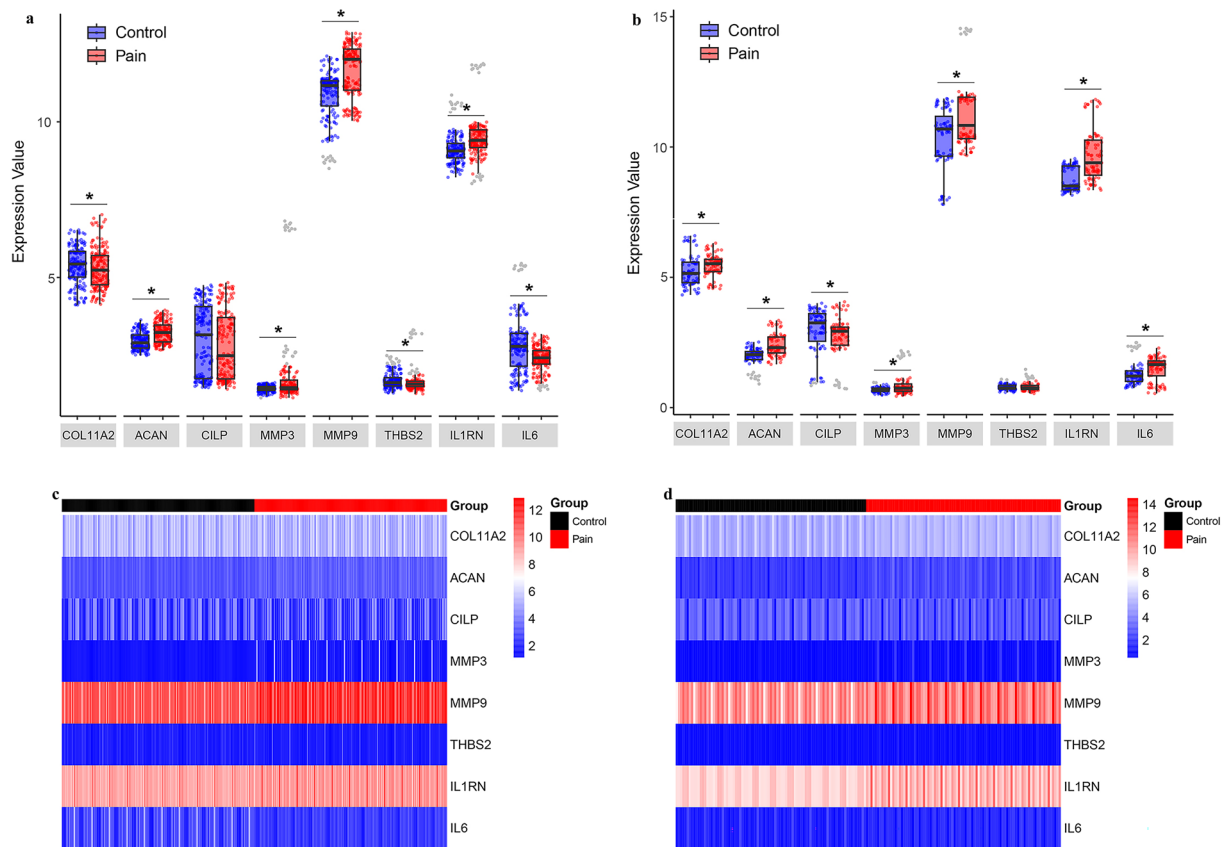


Fig. 3 Differential expression levels of selected gene features between LPD and control groups. Box plot in the **a** training set and **b** testing set. Heatmap in the **c** training set

and **d** testing set. *LPD* lumbar prolapsed disc. *Indicates $p < 0.05$ for FDR-adjusted values

Table 1 Performance metrics with 95% confidence intervals on each machine learning approach

Method	Accuracy	AUC	F1 score	MCC
Random forest	0.80 (0.73–0.85)	0.66 (0.56–0.75)	0.83 (0.77–0.88)	0.64 (0.53–0.76)
KNN	0.71 (0.64–0.77)	0.71 (0.56–0.75)	0.75 (0.68–0.81)	0.45 (0.32–0.58)
XGBoost	0.69 (0.62–0.75)	0.63 (0.54–0.71)	0.74 (0.67–0.80)	0.42 (0.28–0.55)
SVM	0.61 (0.54–0.68)	0.84 (0.79–0.90)	0.72 (0.65–0.78)	0.35 (0.21–0.49)
Logistic regression	0.56 (0.48–0.63)	0.89 (0.83–0.94)	0.69 (0.62–0.76)	0.26 (0.11–0.40)

AUC area under the curve, *MCC* Matthews correlation coefficient, *SVM* support vector machine, *KNN* *k*-nearest neighbours, *XGBoost* Extreme Gradient Boosting

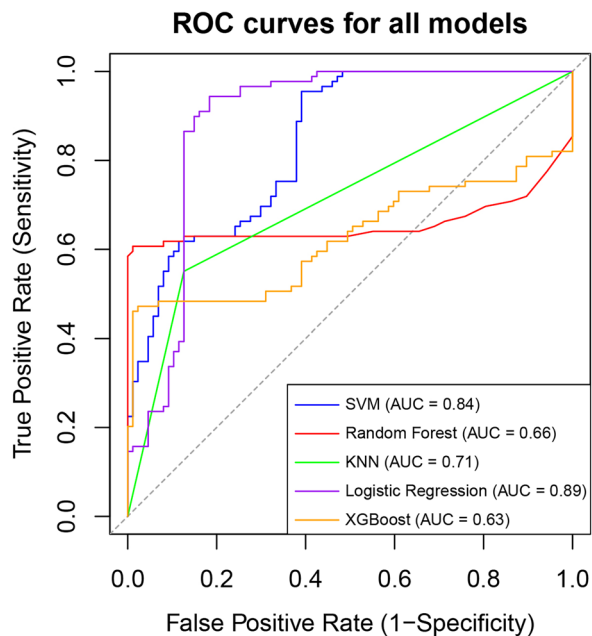


Fig. 4 ROC curves for all the predictive models. *SVM* support vector machine, *KNN* *k*-nearest neighbours, *XGBoost* Extreme Gradient Boosting, *ROC* receiver operating characteristic, *AUC* area under the curve

lowest accuracy among the approaches, at 0.56 (95% CI 0.48–0.63) (Table 1 and Fig. 4).

In terms of AUC scores, the logistic regression model achieved the highest AUC of 0.89 (95% CI 0.83–0.94), followed by the SVM model with an AUC of 0.84 (95% CI 0.79–0.90). The KNN model had an AUC of 0.71 (95% CI 0.56–0.75), and the random forest model had an AUC of 0.66 (95% CI 0.56–0.75). The XGBoost model had the lowest AUC of 0.63 (95% CI 0.54–0.71).

The F1 scores, which consider both precision and recall, were highest for the random forest model (0.83, 95% CI 0.77–0.88), suggesting a good balance between correctly identifying positive cases and avoiding false positives. The KNN model had a slightly lower F1 score of 0.75 (95% CI 0.68–0.81), followed by the SVM model at 0.72 (95% CI 0.65–0.78). The XGBoost model achieved an F1 score of 0.74 (95% CI 0.67–0.80). The logistic regression model had an F1 score of 0.69 (95% CI 0.62–0.76), indicating comparatively lower performance in balancing precision and recall.

The MCC, a measure of the quality of binary classifications, was highest for the random forest model (0.64, 95% CI 0.53–0.76), followed by the KNN model (0.45, 95% CI 0.32–0.58) and XGBoost (0.42, 95% CI 0.28–0.55). The SVM model had an MCC score of 0.35 (95% CI 0.21–0.49). The logistic regression model had the lowest MCC score at 0.26 (95% CI 0.11–0.40), indicating less optimal binary classification performance.

DISCUSSION

In this study, we developed predictive models using machine learning techniques to identify individuals at high risk of lumbar prolapsed disc based on key gene signatures. Eight gene signatures were identified for model training, with *MMP9* demonstrating the highest importance score. Among the tested models, the random forest model outperformed others in terms of accuracy, F1 score, and MCC score, highlighting its effectiveness in balancing precision and recall and providing reliable predictive performance across all classes.

The findings of this study have significant clinical implications and potential significance for researchers, clinicians, and patients. For researchers, this work provides a novel framework for using machine learning to identify critical gene signatures associated with LPD, offering new insights into the potential genetic mechanisms underlying the condition. For clinicians, the predictive models developed in this study could be incorporated into clinical tools to facilitate the early identification of high-risk patients, enabling the implementation of preventive measures and personalized management strategies. Such interventions could potentially reduce the progression of LPD, alleviate chronic pain, and improve the life quality of patients.

The potential of using machine learning models to predict the risk LPD based on specific pain-related gene signatures has not been previously investigated. Several previous studies on machine learning focused on predicting treatment outcomes for low back pain and/or were primarily based on demographic and

clinical characteristics. In contrast, our current study utilized machine learning models based on gene signatures to predict the occurrence of LPD, which is potentially useful for early diagnosis and preventive management. This current study also has several notable strengths compared to the previous machine learning studies. First, we used independent datasets for training and validation, whereas a previous study predicting treatment outcomes for low back pain relied on a single dataset for both training and testing [15]. The use of two independent datasets reduces the risk of overfitting and enhances the reliability and generalizability of the model. Second, we focused on a curated set of 23 genes linked to LPD and chronic pain, allowing us to uncover deeper insights into their roles in the development of LPD. Third, the random forest model in our study achieved a high predictive accuracy of 80%, compared to 65.6–71.6% in a previous study that used a machine learning model based on patient demographic and clinical characteristics to predict the risk of chronic low back pain [36]. The results support the feasibility of using gene signatures for risk prediction.

The selected gene signatures are biologically relevant to LPD, which provides the biological basis for the machine learning models in this study. The selected gene signatures are associated with intervertebral disc stability, inflammation, and pain signalling, which are key processes in the development of LPD [22–24]. Among the eight gene signatures, *MMP9* had the highest importance score and was consistently upregulated in the LPD group across both training and testing datasets. *MMP9* is an enzyme involved in extracellular matrix degradation, inflammation, and tissue remodeling. Previous studies have reported increased *MMP9* expression in patients with chronic pain conditions, including low back pain and lumbar disc herniation, and have linked it to structural and functional changes in the intervertebral disc [37, 38]. Our findings also support *MMP9* as a key factor in LPD development. It may have a potential role as a biomarker to predict the risk of LPD.

LIMITATIONS

This study has several limitations. First, the current sample size is relatively small, which may make it more susceptible to overfitting, potentially resulting in suboptimal performance when the model is applied to new data [39, 40]. To address this limitation, we incorporated Gaussian noise to simulate data and employed the RFE technique to reduce overfitting risks and enhance generalization capabilities [26, 41]. A further limitation of this study was its reliance on a single dataset to evaluate the model's performance. Depending on only one dataset for testing may hinder the generalizability of the results to different populations or settings [42]. To mitigate this issue and strengthen the validity of our findings, five distinct models were employed. Additionally, we used multiple performance indicators to provide a comprehensive assessment of each model's effectiveness. This multi-faceted evaluation approach helps ensure that our results are not only more reliable but also provide a broader perspective on the models' capabilities in distinguishing patients with LPD from healthy controls. Another limitation was that the original data source did not provide demographic and clinical information, such as gender, age, region, and disease course. This may limit the ability to assess their potential influence on gene expression profiles or model performance. Future studies should if possible incorporate such data to enable subgroup analyses.

CONCLUSION

The findings of our study suggest that machine learning models based on pain-related gene signatures may identify patients at high risk of developing lumbar prolapsed disc with reasonably high accuracy. These predictive models have the potential to be integrated into clinical diagnostic tools, which could assist in early diagnosis and the implementation of personalized preventive management strategies for at-risk individuals.

ACKNOWLEDGEMENTS

We thank the participants of the study.

Medical Writing, Editorial and Other Assistance. No assistance was used.

Author contributions. Fengfeng Wang and Stanley Sau Ching Wong initiated the project and contributed to the overall design. Fengfeng Wang carried out the analyses and built the models. Fengfeng Wang, Fei Meng and Stanley Sau Ching Wong were involved in interpreting the data. Fengfeng Wang was responsible for writing the manuscript. Fengfeng Wang, Fei Meng and Stanley Sau Ching Wong engaged in discussion and editing of the manuscript.

Funding. This study, including the Rapid Service Fee, was supported by the Department of Anaesthesiology, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong.

Data Availability. The datasets used and/or analysed in the current study were obtained from the publicly available Gene Expression Omnibus (GEO) database (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150408> and <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE124272>).

Declarations

Conflict of Interest. The authors (Fengfeng Wang, Fei Meng and Stanley Sau Ching Wong) declared that there are no conflicts of interests.

Ethical Approval. This research utilized existing studies and did not involve any new experiments with human participants or animals conducted by the authors. For model training, we obtained transcriptomic data from patients with lumbar prolapsed disc using the GEO database entry GSE150408, and for model testing, we used data from GSE124272. All procedures in these referenced studies involving

human participants were approved by the Ethics Committee of the Sichuan Provincial Orthopedic Hospital.

Open Access. This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc/4.0/>.

REFERENCES

1. Jensen MC, Brant-Zawadzki MN, Obuchowski N, Modic MT, Malkasian D, Ross JS. Magnetic resonance imaging of the lumbar spine in people without back pain. *N Engl J Med*. 1994;331(2):69–73.
2. Frymoyer JW. Back pain and sciatica. *N Engl J Med*. 1988;318(5):291–300.
3. Wang F, Cheung CW, Wong SSC. Regenerative medicine for the treatment of chronic low back pain: a narrative review. *J Int Med Res*. 2023;51(2):03000605231155777.
4. DePalma MJ, Ketchum JM, Saullo T. What is the source of chronic low back pain and does age play a role? *Pain Med*. 2011;12(2):224–33.
5. Ropper AH, Zafonte RD. Sciatica. *N Engl J Med*. 2015;372(13):1240–8.
6. Hoy D, March L, Brooks P, et al. The global burden of low back pain: estimates from the Global Burden of Disease 2010 study. *Ann Rheum Dis*. 2014;73(6):968–74.
7. Delitto A, George SZ, Van Dillen L, et al. Low back pain: clinical practice guidelines linked to the International classification of functioning,

- disability, and health from the orthopaedic section of the American Physical Therapy Association. *J Orthop Sports Phys Ther.* 2012;42(4):A1–57.
8. Manchikanti L, Knezevic NN, Navani A, et al. Epidural interventions in the management of chronic spinal pain: American Society of Interventional Pain Physicians (ASIPP) comprehensive evidence-based guidelines. *Pain Physician.* 2021;24(S1):27.
 9. Jn W. Surgical vs nonoperative treatment for lumbar disk herniation: the Spine Patient Outcomes Research Trial (SPORT) observational cohort. *JAMA.* 2006;296:2451–9.
 10. Orozco L, Soler R, Morera C, Alberca M, Sánchez A, García-Sancho J. Intervertebral disc repair by autologous mesenchymal bone marrow cells: a pilot study. *Transplantation.* 2011;92(7):822–8.
 11. Chou R, Qaseem A, Snow V, et al. Diagnosis and treatment of low back pain: a joint clinical practice guideline from the American College of Physicians and the American Pain Society. *Ann Intern Med.* 2007;147(7):478–91.
 12. Kreiner DS, Hwang SW, Easa JE, et al. An evidence-based clinical guideline for the diagnosis and treatment of lumbar disc herniation with radiculopathy. *The Spine Journal.* 2014;14(1):180–91.
 13. Ritchie MD. The success of pharmacogenomics in moving genetic association studies from bench to bedside: study design and implementation of precision medicine in the post-GWAS era. *Hum Genet.* 2012;131:1615–26.
 14. Wang F, Cheung CW, Wong SSC. Use of pain-related gene features to predict depression by support vector machine model in patients with fibromyalgia. *Front Genet.* 2023;14:1026672.
 15. Lian Y, Shi Y, Shang H, Zhan H. Predicting treatment outcomes in patients with low back pain using gene signature-based machine learning models. *Pain Ther.* 2025;14(1):359–73.
 16. Berg B, Gorosito MA, Fjeld O, et al. Machine learning models for predicting disability and pain following lumbar disc herniation surgery. *JAMA Netw Open.* 2024;7(2):2355024.
 17. Freidin MB, Tsepilov YA, Palmer M, et al. Insight into the genetic architecture of back pain and its risk factors from a study of 509,000 individuals. *Pain.* 2019;160(6):1361.
 18. Wang Y, Dai G, Jiang L, Liao S, Xia J. Microarray analysis reveals an inflammatory transcriptomic signature in peripheral blood for sciatica. *BMC Neurol.* 2021;21(1):1–11.
 19. ICD-10-CM Codes. <https://www.aapc.com/codes/icd-10-codes-range/>. Accessed Apr 2025.
 20. Wang Y, Dai G, Li L, et al. Transcriptome signatures reveal candidate key genes in the whole blood of patients with lumbar disc prolapse. *Exp Ther Med.* 2019;18(6):4591–602.
 21. Tegeder I, Lötsch J. Current evidence for a modulation of low back pain by human genetic variants. *J Cell Mol Med.* 2009;13(8b):1605–19.
 22. Kjaer P, Leboeuf-Yde C, Sorensen JS, Bendix T. An epidemiologic study of MRI and low back pain in 13-year-old children. *Spine.* 2005;30(7):798–806.
 23. Risbud MV, Shapiro IM. Role of cytokines in intervertebral disc degeneration: pain and disc content. *Nat Rev Rheumatol.* 2014;10(1):44–56.
 24. Kawakami M, Tamaki T, Weinstein JN, Hashizume H, Nishi H, Meller ST. Pathomechanism of pain-related behavior produced by allografts of intervertebral disc in the rat. *Spine.* 1996;21(18):2101–7.
 25. Alwosheel A, van Cranenburgh S, Chorus CG. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *J Choice Modelling.* 2018;28:167–82.
 26. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Mach Learn.* 2002;46:389–422.
 27. Kuhn M. Building predictive models in R using the caret package. *J Stat Softw.* 2008;28:1–26.
 28. Hastie T, Tibshirani R, Friedman JH, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. Cham: Springer; 2009.
 29. Kuhn M. Caret: classification and regression training. *Astrophysics Source Code Library.* 2015:ascl:1505.003.
 30. Meyer D, Dimitriadou E, Hornik K, et al. e1071: misc functions of the department of statistics, probability theory group (formerly: E1071), TU Wien. R package version. 2019;1(2).
 31. Breiman L. Random forests. *Mach Learning.* 2001;45:5–32.
 32. Fix E. Discriminatory analysis: nonparametric discrimination, consistency properties: USAF school of Aviation Medicine; 1985.
 33. Dobson AJ, Barnett AG. An introduction to generalized linear models. Chapman and Hall/CRC; 2018.

-
34. Chen T, He T, Benesty M, et al. Xgboost: Extreme Gradient Boosting. R package version 04–2. 2015;1(4):1–4.
 35. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Stat Soc: Ser B (Methodol)*. 1995;57(1):289–300.
 36. Shim J-G, Ryu K-H, Cho E-A, et al. Machine learning approaches to predict chronic lower back pain in people aged over 50 years. *Medicina*. 2021;57(11):1230.
 37. Kobayashi S, Meir A, Kokubo Y, et al. Ultrastructural analysis on lumbar disc herniation using surgical specimens: role of neovascularization and macrophages in hernias. *Spine*. 2009;34(7):655–62.
 38. Wang X, Wang H, Yang H, et al. Tumor necrosis factor- α - and interleukin-1 β -dependent matrix metalloproteinase-3 expression in nucleus pulposus cells requires cooperative signaling via syndecan 4 and mitogen-activated protein kinase–NF- κ B axis: implications in inflammatory disc disease. *Am J Pathol*. 2014;184(9):2560–72.
 39. Button KS, Ioannidis JP, Mokrysz C, et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci*. 2013;14(5):365–76.
 40. Hawkins DM. The problem of overfitting. *J Chem Inf Comput Sci*. 2004;44(1):1–12.
 41. Goodfellow I, Bengio Y, Courville A, Bengio Y. *Deep learning*. Cambridge: MIT press; 2016.
 42. Kohavi R, editor *A study of cross-validation and bootstrap for accuracy estimation and model selection*. Ijcai: Montreal, Canada; 1995.
-