AI in Orthopaedics

Wang Yui Henry To, Mohammed Jafer Ali, Tsz Shun Max Chang, Sukhraj Virdee

December 8, 2024

1 Introduction

The following report documents our solution to the task provided and consists of the steps taken towards achieving the final model code, instructions for its operation, and our strategies for clinical implementation.

2 Model Development

The goal for this competition was to produce a machine-learning algorithm that could classify orthopaedic patients into classes using six image-derived, biomechanical predictors. These are listed as follows

- Pelvic incidence
- Pelvic tilt
- Lumbar lordosis
- Sacral slope
- Pelvic radius
- Degree of spondylolisthesis

The task was further divided into two, where the main task was to differentiate patients into normal and abnormal classes, and the bonus was to further divide abnormal results into disc hernia and spondylolisthesis. The data provided contained 2 data sets. The first provided 310 samples, each containing the 6 biomechanical predictor variables and its classification of "Normal" or "Abnormal". The latter provided 309 samples of the same predictor variables but with the classifications of "Normal", "Hernia" and "Spondylolisthesis".

2.1 Main Task

The outcome variable for the main task was the classification of the orthopaedic patient, whereby the result was either "Normal" or "Abnormal", i.e. binomial in nature. Thus, the use of logistic regression analysis was decided to better suit the desired outputs. The sigmoid function produced would keep output probabilities between 2 values representing "Normal" and "Abnormal" and allow for ease of classification by the use of thresholds. The regression analysis model would also provide relatively simpler interpretation via the coefficients/weighting, confusion matrices etc. The decision to forgo the use of prebuilt regression models, such as those developed by Scikit-learn (Pedregosa et al. 2011), TensorFlow (Abadi et al. 2016) and PyTorch (Paszke et al. 2019), and to code the model from scratch was done to facilitate ease of customisation for any parameters and extraction of processed data. Our baseline model was made with reference to the resources provided by AssemblyAI (2022) and modifications were applied to suit the given task.

The logistic regression model is expressed in the following equation:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(w^T \cdot X + b)}}$$

The left-hand side represents the probability of the given patient being classed as "Abnormal" and takes any value between 0 and 1. By default, the threshold is set to 0.5, whereby any output value greater or equal to 0.5 will be classed as "abnormal". The formula below within the exponent represents how the biomechanical values given influences the output probability.

$$w^T \cdot X + b$$

w^T is the weighting vector that stores information on how much each biomechanical value influences the output probability. To further visualise how this component affects the model, we can rearrange the logistic regression equation to the logit function in the following steps

$$\begin{aligned}
\text{Odds} &= \frac{P(Y = 1|X}{1 - P(Y = 1|X)} \\
\text{Odds} &= \frac{1}{1 + e^{-(w^T \cdot X + b)}} \times \frac{1 + e^{-(w^T \cdot X + b)}}{e^{-(w^T \cdot X + b)}} \\
\text{Odds} &= e^{w^T \cdot X + b} \\
ln(Odds) &= w^T \cdot X + b \\
\text{logit}(P(Y = 1|x)) &= w^T \cdot X + b
\end{aligned}$$

Therefore, the purpose of the analysis was to identify the variables w^T and b that provides the regression model with the most accuracy and desired error levels. This process was further divided into the following steps

1. Data Preparation
2. Development of the Logistic Regression Model
3. Validation and Analysis of the Logistic Regression Model

2.1.1 Data Preparation

The first section of the code was to reorganise and prepare the data provided with the task. The data set was first split into X and y, where the first contained the biomechanical values for each patient, and the latter the outcome classifications. These are further divided into "training" and "test" sets, the latter of which is used to act as validation of the resulting logistic regression model. By convention, the split was in an 80%:20% ratio for "training" and "test" sets. The class variables "Abnormal" and "Normal" in the y data sets were mapped into numerical values of 1 and 0 respectively for the purposes of the regression analysis. Thus, the number generated for the regression model, $P(Y = 1|X)$, would be the probability that the patient is classed as abnormal.

2.1.2 Training Phase of Logistic Regression Model

The developmental stage of the regression model was divided into 3 parts. The first part comprised of processing the training data into the outcome variable using the baseline model. In the first iteration of the code, weights and bias were given baseline values of 0. The second part was used to fit our predictive

model to the "true" results provided in the dataset and train the model via the gradient descent optimisation algorithm. This method identifies the derivative of binary cross-entropy loss with respect to each weight and bias, as described in the following two equations

$$\frac{\partial L}{\partial w} = \frac{1}{n} \times X^T \cdot (\hat{y} - y)$$
$$\frac{\partial L}{\partial b} = \frac{1}{n} \times \sum_{i=1}^{n} (\hat{y} - y)_i$$

The weights and bias for the next iteration are then modified via the following equations

$$w \leftarrow w - \eta \times \frac{\partial L}{\partial w}$$
$$b \leftarrow b - \eta \times \frac{\partial L}{\partial b}$$

This algorithm was then repeated for a set number of iterations to further optimise our weight and bias values and minimise loss. The parameters of the algorithm were as follows

- Learning rate $(lr) = 0.001$
- Max epoch (iterations) $= 1000$

Finally, the resulting logistic regression model was tested with the test sets. The "X" test set was processed into probability values using the regression model, then further mapped onto either 1 and 0 based on our threshold, which was 0.5 by convention. These outcome values were compared to the given classifications in the accuracy analysis.

2.1.3 Analysis of Regression Model and Further Modifications

Our first trained regression model had the following weights and bias values

$$w = [0.0256, 0.0806, -0.0353, -0.0551, -0.0118, 0.189]$$
$$b = 0.0083$$

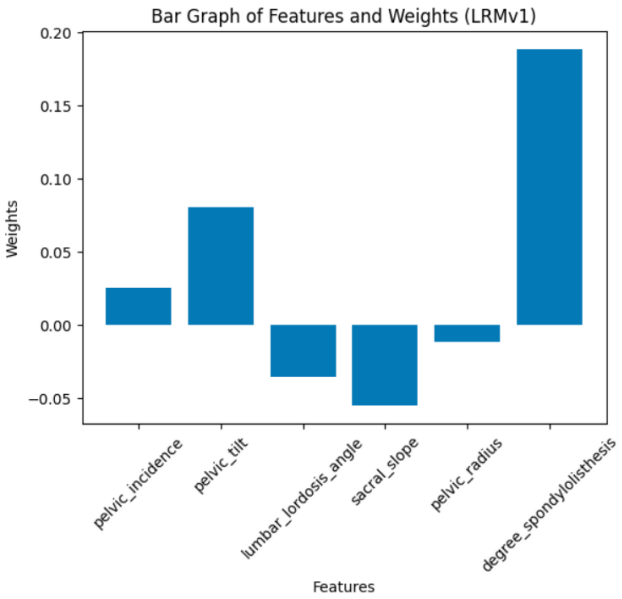

Figure 1: The bar chart provides visualisation of the weights for each biomechanical variable in the first version of the logistic regression model

The weights were for pelvic incidence, pelvic tilt, lumbar lordosis angle, sacral slope, pelvic radius and degree of spondylolisthesis respectively. As shown, the weighting of degree of spondylolistesis is particularly high. This is unsurprising given an absolute value larger than zero can already signify abnormality, as shown in the grading system of spondylolisthesis. Our testing yielded a classification accuracy of 77.4%. Breaking this down further, the model had a sensitivity value of 69.0% and specificity value of 95%.

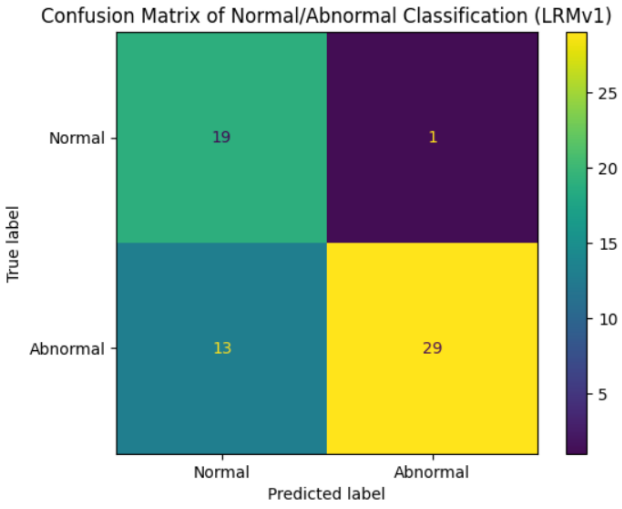

Figure 2: The confusion matrix shows a comparison between the true classes and predicted classes. The left bottom and right top quadrants represent mismatched values between the model and dataset.

Modifications to the optimisation algorithm aimed increase accuracy and decrease Type II error or number of false negatives. For the first goal, parameters for the optimisation algorithm were modified. To ensure

the minimum loss is reached, the number of iterations was increased along a decrease in learning rate to avoid overshooting or oscillations. Concurrently, an additional stopping criteria was introduced, where the cessation of the optimisation algorithm would occur if all the partial derivatives with respect to the weights of each variable and bias values were below the tolerance level. Thus, the new parameters for the second version of the model was as follows

- $lr = 0.0001$
- Max epoch $= 10000$
- Tolerance (tol) $= 0.01$

The steps for testing our previous algorithm was repeated for this version, and showed an increased accuracy of 85.5%. The confusion matrix, as shown below, shows in increased sensitivity of 88.1% but a decreased specificity of 80%.

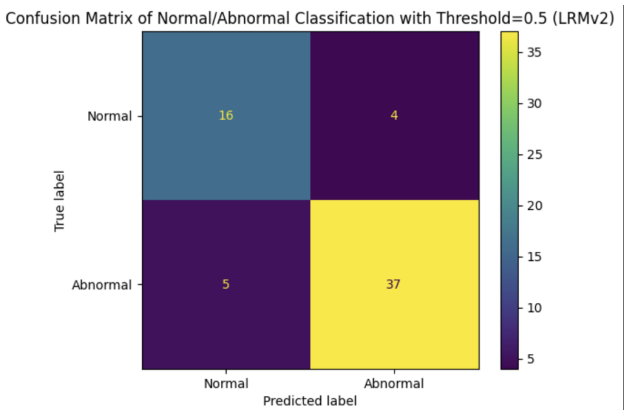

Figure 3: The confusion matrix of the model training with an updated optimisation algorithm.

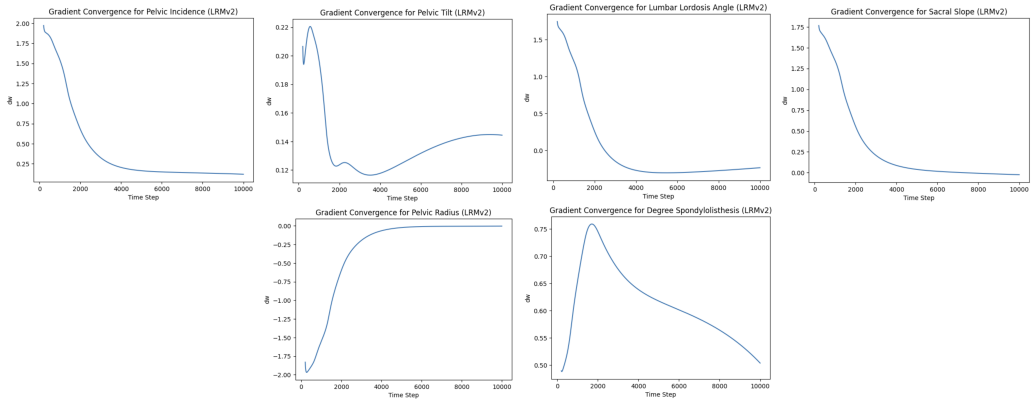

Figure 4: The graphs above allow for visualisation of the loss gradient convergence to 0 following the gradient descent algorithm. As shown, various gradient values with respect to the weight has not reached within the tolerance value.

However, it is worth noting that the loss function was yet to be minimised, as our partial derivatives did not dip below the tolerance values. The max epoch was therefore increased to a value of 30000, whereby the tolerance level was reached. Rather interestingly, the accuracy for this version of the model decreased to

83.9%. Specificity remained the same but sensitivity was slightly reduced to 85.7%. For the implementation of this model in a clinical setting, the false positive rates were favoured over false negatives, and thus the threshold for classification as "abnormal" was decreased to 0.3. The increased the sensitivity to 95.2% at the expense of decreasing specificity to 65.0%.

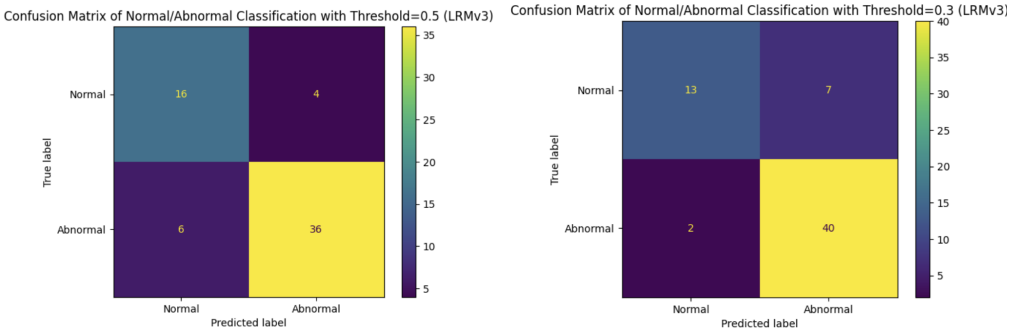

Figure 5: The figure shows a side-by-side comparison of the confusion matrices between the 0.5 and 0.3 threshold.

2.2 Bonus Challenge

The bonus challenge was to divide the samples into three classes rather than the previous two. The softmax regression model was determined to be suitable for this task and retains similarities to the logistic regression model used previously. Rather than producing a binary classification, the model is a multinomial logistic regression, allowing for multi-class classification capabilities. Where the former logistic function produces the probability value of the patient being classified as "abnormal", the softmax function works to produce a probability value for each class, where the class with the highest probability is the output. The softmax regression model is expressed in the following function:

$$P(Y = k|X) = \frac{e^{W_k} \times X + b_k}{\sum_{j=1}^{K} e^{W_j} \times X + b_j}$$

The training phase of this model also mirrored its predecessor, using the gradient descent optimisation algorithm with the same stopping criteria and parameters as listed

- $lr = 0.0001$
- Max epoch = 30000
- Tolerance = 0.01

The training softmax regression model yielded an accuracy of 91.94% on the test set. Further breakdown of this is shown in the confusion matrix below.

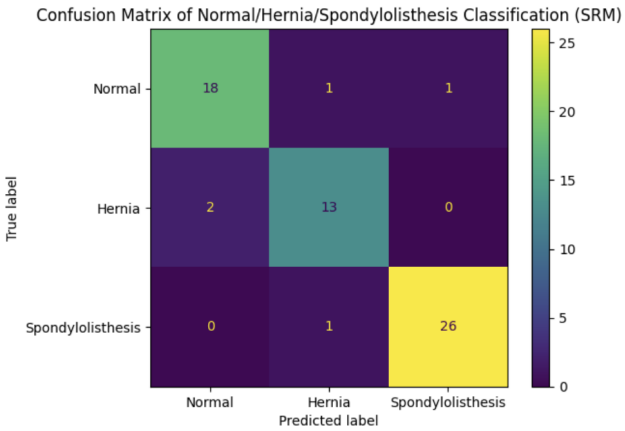

Figure 6: The confusion matrix shows the distribution and comparison bewteen predicted and true classes.

3 Discussion and Future Considerations

3.1 Regression Model

The logistic and softmax regression models were chosen due to their relative simplicity and thus an elevated suitability for the small data sets provided and taking into account the time constraints placed on the challenge. With given opportunity, the suitability of different models, e.g. decision trees, random forest, k-nearest neighbours, could have been quantitatively compared. For example, cross-validation methods could have allowed for performance testing between models to allow for attainment of the highest average accuracy.

Within our model itself, the biomechanical features themselves should be discussed. Feature selection could have shown great importance in developing our model. The pathway between the given features and classifications were not well understood and so while the logistic regression model assumed a linear relationship between the log-odds of the outcome and predictor, this may not have been the case in actuality. We identified high correlations between various biomechanical features provided in the data set. This may be resolved through making adjustments within the linear prediction model, $w^T \cdot X + b$, or by the use of another, more appropriate model. For example, the correlated variables could have been combined or selectively removed before regression analysis. In the training phase of our model, the baseline gradient descent algorithm was used and tolerance level was reached. However, it is worth noting that other optimisation algorithms may have been appropriate if the dataset was larger, either in the number of data points or features. The validation testing of our model demonstrated a prediction of the model's accuracy, though using the k-fold cross validation method could have allowed for more objective accuracy estimations. The aforementioned threshold value for binomial classification could have been more thoroughly explored using the ROC curve.

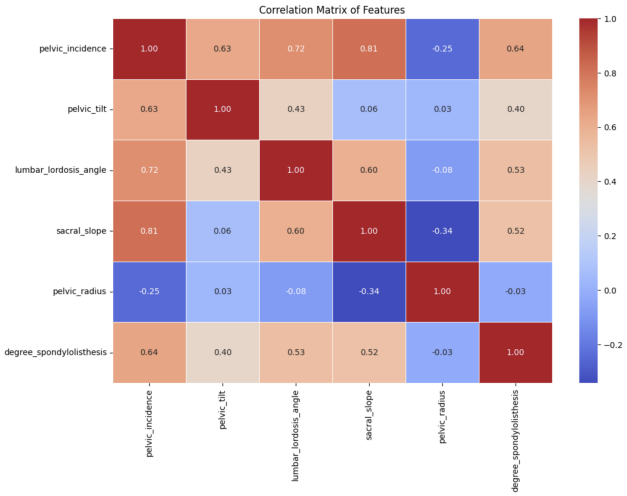

Figure 7: The figure shows a correlation matrix between the 6 biomechanical features provided in the data set. In particular, pelvic incidence seems to be highly correlated with a number of variables, and may be worth considering in the feature selection process.

3.2 The Significance of AI in Analysing Variables to Assess Lower Back Pain

Computer scientists have began to develop multiple machine learning (ML) algorithm tools to manage sensitive patient information, utilising Local Binary Pattern (LBP) (Ojala et al. 2000) and Scale-Invariant Feature Transform (SIFT) (Al-Ayyoub et al. 2018). The role of AI machine learning plays a significant role in detecting disc herniation via MRI and CT scans. In particular, ML has recently gained noticeable popularity from clinicians and involves applying algorithms to perform automated decision-making processes that have been routinely trained on large datasets of medical images yet not manually programmed. Artificial neural networks (ANN, a subset of ML) helps stimulate the structure and function of the human brain, alongside deep learning (DL). Collectively, they enable the smooth processing of large complex datasets (Kufel et al. 2023). AI during spinal surgery has opened many doors, including segmentation of spinal structures (Li et al. 2018), classifying degenerated discs, fractures identification and detection amongst several others (Murata et al. 2020).

Lumbar disc herniation is commonly affiliated with neuropathic pain due to persistent compressive effects on neural structures. The risk of developing chronic neuropathic pain is closely related to the duration and intensity of symptoms experienced. The thresholds of contextual information belonging to patients varies largely with the lack of official definitions underlying the justification of surgical intervention (Lurie et al. 2014). Despite this promising knowledge, it is challenging to distinguish which patients benefit from early surgical therapy compared to those who would avoid this conservatively due to unfounded additional benefits attached (Wirries et al. 2022). Further clinical information on the severity of neuropathic pain experienced was obtained via a machine learning approach which assessed visual analogue scale (VAS) and Oswestry Disability Index (ODI) at set periods after treatment. A decision tree regressor algorithm was used to predict quality within the limits of minimum clinically important differences for both VAS and ODI values. Correlation matrix, density distributions and histograms of many parameters were performed. ML helped define 'ODI' and 'leg pain', which manifested itself into a linear regression problem. To address this, recursive feature elimination, weighting and analysis of inter-correlating features led to certain parameters being removed to reduce complexity.

Computer aided diagnosis (CAD) is relevant in:

- Classification settings to identify and categorise pathologies

- Regression to produce numerical output as a quantitative evaluation point
- Helping with diagnosis of degenerative spinal changes from imaging data with an average accuracy rate of $> 80\%$
- Analysis of clinical, biomechanical, electrophysiological and functional imaging data (D'Antoni et al. 2022)

Specific to lower back pain, CAD systems analyse medical imaging data and automatically detect overlapping patterns to aid early detection and diagnosis. A benchmark dataset was formed by axial MRI scans instead of sagittal versions due to providing more information regarding the disc area (Ebrahim et al. 2024). Pre-trained models in the past have been programmed to perform feature extraction and fine-tuning on lumbar disc radiology dataset. There are meaningful features from images that are captured from using simple to complex architectures. Regarding machine learning classifiers, multilayer deep neural networks (DNN) consist of many layers to be trained on the extracted features, alongside help from MATLAB implementation. Collectively speaking, this speeds up the efficiency of analysing complex images from herniated disc pathology.

3.3 Clinical Implementation

The proposed algorithm can be integrated into orthopaedic pathways including prioritising and streamlining orthopaedic outpatient appointments. Back pain is a common presentation in the community and there is a significant burden of disc hernias and spondylolisthesis in the NHS. Disc hernias present in almost 5% of the general population (Pojskic et al. 2024) and spondylolisthesis can present in almost 7% (VIRTA 1992). This poses a significant challenge and burden to the outpatient orthopaedic department, and also presents an opportunity for the utilisation of AI to mitigate this.

Implementing the algorithm within UK orthopaedic outpatient settings would require several structured steps. These steps are designed to align with the operational and regulatory frameworks of the NHS clinical practices. Firstly, existing spinal imaging must be processed in real-time through current interfaces used, for example PACS (Picture Archiving and Communication Systems) (Dargan 2020). This would allow AI detection of disc hernia or spondylolisthesis in real-time post-imaging. Following this, patients would be able to be stratified according to severity and clinical urgency for intervention. This, alongside confidence scores, can be flagged to radiologists to assist in generating earlier reports, whereas imaging flagged as normal by AI can be reported with less urgency, minimising backlogs and delays in operative care for patients with disc hernias or spondylolisthesis (Chen et al. 2022). Triaged patients can be seen in specific clinics relevant to disc herniations and spondylolisthesis and have same-day anaesthetic review or prior pre-operative assessments (e.g. ECG or chest x-rays).

The AI algorithm made, seems promising in increasing the efficiency of diagnosis, by an accuracy of 83.9%. We can incorporate AI alongside standard diagnosis and clinical judgement as a 'diagnostic-aid tool'. This is due to Artificial intelligence's strength of processing speeds and pattern recognition. If used in conjunction with clinician MDT, it may notably reduce the number of rates of false-positive diagnosis. Within orthopaedics, A 2007 study mentioning, that false-positive rates were as high as 12.6% (Sharma et al. 2007), for misdiagnosis of fractures. This highlights the challenges clinicians may face when interpreting images, and helps mitigate the challenges in an emergency setting of high working pressures where the risk of misdiagnosing patients. Interoperability must also be considered, given the increasing use of surgical hubs and focus on lean systems to provide effective solutions for high volumes with low complexity pathologies.

Given the long wait times in the NHS, the AI algorithm can be used as a pre-operative planning tool to identify patients suitable for same-day surgeries or patients likely to need additional MDT involvement post-operatively. However, further cost utility and cost effectiveness analysis would need to be considered to determine what would be the cost benefit of incorporating an AI algorithm to the clinical pathway, as steps

would be required to ensure this algorithm is maintained and updated. For example adding other clinically relevant variables such as age, BMI, comorbidities and osteoporosis risk using DEXA z-score. This technology if incorporated additionally may significantly reduce NHS costs in the long term due to reduced rates of repeat imaging and can aid radiologist shortages by performing innovative reviews post interpretation to reduce interpretation bias to enhance success of diagnosis (Zhang et al. 2023).

When incorporating AI algorithms into the clinical pathways, it is important to consider ethical factors including NHS regulatory approval. This includes ensuring pilot testing if performed at local sites, audits and quality improvement projects are completed to uphold accountability and ensure all stakeholders are involved in key integration steps with any concerns addressed early. Further iteration of the algorithm may be required after results of local centres are assessed and compared. The algorithm made can innovatively used as a form of randomised checking mechanism when interpreting results, to ensure an almost 'quality improvement check' model system, to patient ensure diagnosis is to the gold standard, as a measure of clinical competence or to be used in teaching for doctors in training such as junior doctors or those in registrar positions this can potentially also help in raising any red flags in misdiagnosis, and aid clinically in preventing errors going forward and ensuring that clinicians interpreting skills are always to the gold standard. Highlighting the promising potential educational avenue this algorithm may be able to provide in an orthopaedic outpatient clinical setting.

Finally, this also requires clinician training and acceptance using a bottom-up approach with integration of early clinician adopters in the MDT including orthopaedic trainees, consultants, anaesthetists, radiologists and surgical nurses. Clinicians must be aware of the limitations of the AI algorithm in its output. Limitations of the current algorithm include the focus on identifying herniated disc, spondylolisthesis or normal, with no capability to differentiate other bone or muscular pathologies which may cause further work for clinicians in identifying abnormalities. Also, false positives must be considered by senior radiologists and a process must be put in place when clinical reporting goes against that of the AI algorithm output (Zhang et al. 2023). Ultimately, AI has the potential to provide outpatient orthopaedic clinics with rapid and reliable classification of disc herniation and spondylolisthesis to prevent delayed care and progression of disease.

4 How to Run

The following section explains the setup process and how to run the code.

First, extract the submission.zip file, and use your local terminal to navigate to the file's directory. Assuming Python and Jupyter Notebook has been installed, the next step would be to install the relevant packages. In the file directory, run the command `$pip install -r requirements.txt` in the terminal. Once this is completed, run `$jupyter notebook`. This will launch Jupyter Notebook, where the code can be accessed by navigating to the project directory. Our solution to the task is stored in `hackathon-solution.ipynb`. When running this program, select "*Restart the kernel and run all cells*" to ensure all cells are run sequentially. The seed for selecting from the dataset is set to `42`, which produced the results from this report. Different results can be generated by setting the `SEED` variable to a different number.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M. & others (2016), 'Tensorflow: Large-scale machine learning on heterogeneous distributed systems', *arXiv preprint arXiv:1603.04467* .

Al-Ayyoub, M., Al-Mnayyis, N., Alsmirat, M. A., Alawneh, K., Jararweh, Y. & Gupta, B. B. (2018), 'SIFT based ROI extraction for lumbar disk herniation CAD system from MRI axial scans', *Journal of Ambient Intelligence and Humanized Computing* .

AssemblyAI (2022), 'How to implement Logistic Regression from scratch with Python.'.

Chen, K., Stotter, C., Klestil, T. & Nehrer, S. (2022), 'Artificial Intelligence in Orthopedic Radiography Analysis: A Narrative Review.', *Diagnostics (Basel, Switzerland)* **12**(9).

Dargan, R. (2020), 'Integrating AI with PACS Key to Improving Workflow Efficiency'.

D'Antoni, F., Russo, F., Ambrosio, L., Bacco, L., Vollero, L., Vadalà, G., Merone, M., Papalia, R. & Denaro, V. (2022), 'Artificial Intelligence and Computer Aided Diagnosis in Chronic Low Back Pain: A Systematic Review', *International Journal of Environmental Research and Public Health* **19**(10), 5971.

Ebrahim, M., Alsmirat, M. & Al-Ayyoub, M. (2024), 'Advanced disk herniation computer aided diagnosis system', *Scientific Reports* **14**(1), 8071.

Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, , Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M. & Gruszczyńska, K. (2023), 'What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine', *Diagnostics* **13**(15), 2582.

Li, Y., Liang, W., Zhang, Y. & Tan, J. (2018), 'Automatic Global Level Set Approach for Lumbar Vertebrae CT Image Segmentation', *BioMed Research International* **2018**, 1–12.

Lurie, J. D., Tosteson, T. D., Tosteson, A. N. A., Zhao, W., Morgan, T. S., Abdu, W. A., Herkowitz, H. & Weinstein, J. N. (2014), 'Surgical Versus Nonoperative Treatment for Lumbar Disc Herniation', *Spine* **39**(1), 3–16.

Murata, K., Endo, K., Aihara, T., Suzuki, H., Sawaji, Y., Matsuoka, Y., Nishimura, H., Takamatsu, T., Konishi, T., Maekawa, A., Yamauchi, H., Kanazawa, K., Endo, H., Tsuji, H., Inoue, S., Fukushima, N., Kikuchi, H., Sato, H. & Yamamoto, K. (2020), 'Artificial intelligence for the detection of vertebral fractures on plain spinal radiography', *Scientific Reports* **10**(1), 20031.

Ojala, T., Pietikäinen, M. & Mäenpää, T. (2000), Gray Scale and Rotation Invariant Texture Classification with Local Binary Patterns, pp. 404–420.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J. & Chintala, S. (2019), PyTorch: An Imperative Style, High-Performance Deep Learning Library, *in* H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox & R. Garnett, eds, 'Advances in Neural Information Processing Systems', Vol. 32, Curran Associates, Inc.
URL: *https://proceedings.neurips.cc/paper$_f$iles/paper/2019/$f$ile/bdbca288$f$ee7f92f2bfa9f7012727740−Paper.pdf*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & others (2011), 'Scikit-learn: Machine learning in Python', *the Journal of machine Learning research* **12**, 2825–2830.

Pojskic, M., Bisson, E., Oertel, J., Takami, T., Zygourakis, C. & Costa, F. (2024), 'Lumbar disc herniation: Epidemiology, clinical and radiologic diagnosis WFNS spine committee recommendations', *World Neurosurgery: X* **22**, 100279.

Sharma, H., Bhagat, S. & Gaine, W. (2007), 'Reducing Diagnostic Errors in Musculoskeletal Trauma by Reviewing Non-Admission Orthopaedic Referrals in the Next-Day Trauma Meeting', *The Annals of The Royal College of Surgeons of England* **89**(7), 692–695.

VIRTA, L. (1992), 'Prevalence of isthmic lumbar spondylolisthesis in middle-aged subjects from eastern and western Finland', *Journal of Clinical Epidemiology* **45**(8), 917–922.

Wirries, A., Geiger, F., Hammad, A., Bäumlein, M., Schmeller, J. N., Blümcke, I. & Jabari, S. (2022), 'AI Prediction of Neuropathic Pain after Lumbar Disc Herniation—Machine Learning Reveals Influencing Factors', *Biomedicines* **10**(6), 1319.

Zhang, W., Chen, Z., Su, Z., Wang, Z., Hai, J., Huang, C., Wang, Y., Yan, B. & Lu, H. (2023), 'Deep learning-based detection and classification of lumbar disc herniation on magnetic resonance images', *JOR SPINE* **6**(3).