1 Introduction

The application of Artificial Intelligence (AI) has seen tremendous growth in recent times and is at the core of most operations across sectors and industries. The wide acceptance and subsequent adoption of AI stem from its exceptional performance in learning and predicting capabilities, despite the grandiose achievements, human understanding of the reasoning behind the prediction of these machine learning and deep learning models remains very limited as their internal mechanisms have become so complex and cannot be interpreted by users, such abstruse models are referred to as black boxes (Guidotti, Riccardo, et al., 2018).

These black box models are increasingly being adopted across all sectors as they can perform complex computations, make predictions, and take decisions with little or human no intervention (Arrieta, Alejandro Barredo, et al., 2020). There are however growing concerns in critical sectors where the outcome of the decision made by these models could have a direct negative impact on lives or businesses (Guidotti, Riccardo, et al., 2018), for example, in precision medicine, law, finance, and autonomous automobiles in transportation. The adoption of AI-enabled systems in these sectors has been diminished by the fear that the predictions made by the models are not justifiable, trustworthy, or legitimate and neither are the AI systems accountable for their decisions. Consequently, the rise in demand for explanation and interpretation of predictions made by these models (E. Tjoa & C. Guan, 2021).

The General Data Protection Regulation (GDPR) which was passed into law by the European Union parliament and came into force on 25th May 2018 provides rules for the processing and movement of personal data, protection of natural persons concerning their rights to the protection of personal data, and the free movement of personal data within the union (EUR-Lex, 2016). Article 4 (4) defines the term “profiling” as any form of automated processing of personal data which involves the use of such to evaluate some attributes relating to a natural person, most importantly to “predict or analyse attributes regarding the natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements” (EUR-Lex, 2016).

In furtherance to the provisions of article 4(4), articles 22(1) and 22(3) require “data subjects to have the right not to be subjected to a decision based solely on automated processing, including profiling” and for the data controller to deploy relevant measures to protect the rights, freedom, and legitimate interest of the data subject and "at least the right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision”. Articles 13(2f) and 14(2g) require that in the case of automated decision-making and profiling, meaningful information regarding the logic, significance, and consequences of the processes should be made available to the data subject. Given the foregoing, it became imperative and expedient that algorithms with the capability to resolve the black box challenge are developed and implemented.

1.1 Research objectives

This research aims to identify the best-performing explainable artificial intelligence model for adoption in occupational health and safety for the prevention of musculoskeletal symptoms and disorders. The result of this research will create trust and acceptance for the use of artificial intelligence across all sensitive fields as it seeks to address the black box challenge currently preventing the adoption of artificial intelligence in these fields.

1.2 Research Approach

The study adopts an inductive approach along with a pragmatic view. It entails the following steps.

1. **Problem identification:** This step involves a thorough understanding of the research topic, a review of available works of literature, and the identification of gaps in the research area. The project goals, scope, and limitations are defined during this phase.

2. **Action Planning:** This step details the processes involved in the execution of the project, starting with data collection, knowledge of data format and structure, any biases in the data, approval for data use, and a thorough review and understanding of the data.

3. **Implementation:** Different criteria for analysing the regression algorithm will be introduced and evaluated, followed by an introduction of the explainable aspect of the project.

5. **Evaluation:** Model prediction capability and the performance of the explainable AIs in interpreting the predictions will be measured using selected evaluation metrics.

6. **Findings:** will discuss the benefits and limitations of the study and learning points gained from the study.

2 Preliminaries and Literature Review

The chapter begins by presenting the fundamental notion of "explainability" and interpretability. It delves into various approaches used for explaining models and provides an overview of the different categorizations of explanation methods that have been proposed in scholarly works overtime. The goal of this review is to gain a comprehensive understanding of the mechanisms behind explainable AI.

2.1 Concept and motivation for explainable AI

Interpretability and “explainability” are current terms commonly used by researchers to provide an understanding of the reasoning behind the decisions of black box models. The terms though loosely used interchangeably have been identified by many researchers to have different meanings (Linardatos, Pantelis, et al., 2020). Doshi-Velez, Finale & Kim, Been, (2017) refer to interpretability as the capacity to describe in comprehensible terms to a human. Miller, (2019) presented it as the extent to which humans can comprehend the logic behind a decision. Conversely, explainability is connected to the internal workings and reasoning behind a model. Explainability makes for easier interpretability, this means the more explainable a model is the easier it is for humans to understand it. It has been argued by Gilpin, Leilani H, et al., (2018) that to unravel the “black boxness” of a machine learning model, both concepts are equally important.

The mission to understand and provide an explanation to the predictions made by a black box model has been on for over six decades, however, in 2015 the Defense Advanced Research Projects Agency (DARPA) created the Explainable Artificial Intelligence (XAI) program with the sole aim of creating understandability and trust for users of artificial intelligent systems (Gunning, David, et al., 2019).

2.2 Categorization of explainable AI methods

There has been a plethora of methods for explaining the prediction of a black box model, these methods are broadly categorized into three, as proposed by Das, Arun & Rad, Paul (2020). The categorizations are based on (1) Scope, which defines the focus

of the explanation, whether the explanation focuses on individual data instances (Local Explanation) or the explanation focuses on the whole model (Global Explanation), (2) Methodology, which defines the algorithmic approach based on the input data instance or model parameters. The methodology category is sub-divided as either achieved through (a) backpropagation which involves passing the error function backward in order to adjust the model's parameters such as weights and biases (Werbos, Paul J, 1990; Rumelhart, David E, et al., 1995) or (b)Perturbation which involves making selected changes to the input features (Das, Arun & Rad, Paul, 2020). And (3) Usage where it could be either intrinsically interpretable(ante-hoc) (Das, Arun & Rad, Paul, 2020; Kenny, Eoin M, et al., 2021) such as decision-tree and rule-based models or post hoc which defines an instance where the explanation is applied to an already trained model. Ante-hoc models are however confronted with the interpretability- versus -accuracy trade-off (Vilone, Giulia & Longo, Luca, 2020).

Post-hoc method is classified into model-agnostic and model-specific. Model-agnostic does not consider the internal mechanisms of the black box model they are interpreting and as such can work with any black box model while the model-specific method is tied to a particular type of black box model (Vilone, Giulia & Longo, Luca, 2020).

As stated by Guidotti, Riccardo, et al.(2018) an explainable model is required to possess some desirable characteristics such as (1) Interpretability which is the degree to which humans can understand the model's prediction (2) Accuracy, the degree to which the model can make accurate predictions of the observed instances and (3) Fidelity the degree to which the explainable model can accurately emulate the black box model.

2.3 Review of explainable AI evaluation methods

The concept of explainable AI has been in existence for over six decades and there have been several studies on how to assess the performance of explainers but there is a dearth of a generally acceptable systematic and standardized method for evaluating them (Guidotti, Riccardo, et al., 2018). A large number of works focus mostly on the

merits and demerits of current XAI methods (Freitas, 2014), a general overview of XAIs (Arrieta, Alejandro Barredo, et al., 2020), the impact of output format on understandability (Huysmans, Johan, et al., 2011), classification and challenges of XAIs (Arrieta, Alejandro Barredo, et al., 2020). Others emphasized the underlying problems in the field (Doshi-Velez, Finale & Kim, Been, 2017)however, Vilone, Giulia & Longo, Luca(2021) following the review of several works of literature on the evaluation measures for XAI identified two ways in which evaluation metrics for XAI can be categorized (1) objective evaluations and (2) human-centered evaluations. Studies on objective evaluation look at objective metrics and automated approaches for evaluating XAI while studies on human-centered relate to those methods with human feedback and end-user involvement. Alternative categorizations for evaluation measures were presented in (Preece, 2018; Bibal, Adrien & Frénay, Benoît, 2016) among others.

Hoffman, R.R, et al.(2018) presented the concept of explanation goodness and satisfaction as a human-centered approach to evaluating XAI performance. In their paper explanation goodness was associated with factors such the precision and clarity. A list of goodness checks was also presented as a guide to help researchers incorporate goodness in the design of XAIs and explanation satisfaction as the level to which a user comprehends the XAI or process being explained. They advocated the adoption of several scales and questionnaires as a means of assessing the measure of goodness and satisfaction for the performance of XAI.

Rosenfeld(2021) argued that the evaluation of XAI from the perspective of user studies may be flawed as the underlying hypothesis that such a study can adequately cover the intricate mechanism between user performance and explanation may be untrue. The study also presented that confirmation bias may exist in user studies, it cited an instance where users may support the benefit from an explanation that is hitherto based on a wrong hypothesis instead it suggested quantitative metrics D, R, F, and S that are independent of both the task being undertaken by the XAI and its algorithm. The metric D is the measure of change in performance between the black box model and the best-performing transparent model, it seeks to determine the performance variance between models. R is a measure of explanation simplicity, for example, in a rule-based method, the lesser the number of rules the better the explanation. This

argument is predicated on the papers by (Avi Rosenfeld & Ariella Richardson, 2019; Gerd Gigerenzer & Henry Brighton, 2009). F is determined by the number of input features and with the assumption that the fewer F is the clearer the explanation, finally, metric S measures the stability of the explanation and is connected to how well the features can deal with noise perturbations as unstable features inputs are related to poor explainability. The study believed that the metrics are complimentary in most cases though presented independently.

Darias, Jesus M, et al.(2021) reviewed some of the XAI libraries and provided a list of criteria such as usability, variability, interactivity, and other characteristics for ranking the reviewed libraries. A collection of other works has listed accuracy, consistency, fidelity, utility, and comprehensibility as a measure of evaluating the performance of XAI. So far, no quantitative approach has been developed and adopted for evaluating the explainability consequently making all suggested approaches in existing works subjective.

Notwithstanding the huge number and variety of methods already suggested for evaluating XAIs there are still some basic questions and reasonings that are yearning for answers for instance, no conformity among researchers on the term explainability and the explainable properties that influences how well users understand explanation from XAI among users. Guidotti, Riccardo, et al(2018) also argued that the general standardization, implementation, and adoption of XAI across all fields will continue to be a challenge as the concept of explainability is derived from Psychology since it is human-related and associated with terms like trust, transparency, and privacy.

2.4 Explainable AI in occupational health and safety

Researchers in the field of Occupational Health and Safety have in recent times made impressive progress in adopting XAIs to manage work related musculoskeletal symptoms and disorders. Occupational health and safety aim at the provision and maintenance of a balanced work environment that is devoid of stress, harm, and work-related injuries. Being an extensive field, it covers subjects such as workplace health and fitness, safety practices, hazardous materials, violence prevention, mental health, audit and compliance, and ergonomics. In recent times AI has found wide application

in this field and across different industries with a particular interest in the prevention of musculoskeletal symptoms and disorders.

Musculoskeletal disorders such as upper and lower back pain, neck pain, and other related work-induced body pain have been identified as one of the top reasons for poor productivity, absenteeism, and early disengagement in the workplace (Ghasemkhani, Mehdi, et al., 2008). Psychological factors such as mood, insomnia, somatic behaviors, health beliefs, and mental complications have been linked to the occurrence of musculoskeletal disorder (Chan, Victor CH, et al., 2022). Many research efforts have in recent times gone into the use of artificial intelligence in its prevention and this cuts across industries Szeto, Grace PY, et al.(2009) examined the pervasiveness of musculoskeletal symptoms among surgeons and identified the risk factors as inhibited posture, repetitive limb movement, severe exertion, and environmental factors as the major cause of musculoskeletal symptoms in surgeons. Indumathi, N & Ramalakshmi, R(2021) investigated the incidence of MSD among fireworks industry workers using the K-Nearest Neighbour algorithm to predict MSD levels and risk factors using work postures collected across departments in the industry. The model showed better prediction accuracy compared to other machine learning algorithms. Explainable artificial intelligence (XAI) models are therefore needed to provide a better understanding of the predictions of musculoskeletal symptoms and disorders made by black box models. Lee et al. (2023) worked on identifying risk factors in the workplace and how these contribute to injuries among foreign workers using Extreme gradient Boosting model and adopted SHAP to explain the model's prediction. Tang, Yue Ting & Romero-Ortuno, Roman(2022) used Random Forest and SHAP to predict falls in the older population, the result of the study revealed the cause of falls and the XAI provided an understanding of their prevention. A video-based AI system that consists of a Logistic Regression predicator and SHAP explainer was proposed in the study by G. Zhou, et al.(2021) for the assessment of weightlifting risk among workers. The proposed system was able to observe body motion, posture, and facial expression through the predictor model and assess the weight safety levels while the explainer analyses the reasoning for the output. Mollaei, Nafiseh, et al., (2022) developed a Human Centered-XAI model that could predict the next medical visit and vulnerable body parts that require protection from the stress and strain of work by analysing

employees' work capability, their occupational health profile, and work-related musculoskeletal symptoms. The study used data on functional work ability taken from 7857 participants from the automotive industry and generated between 2019 and 2021. Natural Language processing based on machine learning was used to extract relevant information from the report and analysed using regression methods, the study compared results from different models with varying Root Mean Squared Logarithmic Error (RMSLE). CatBoost regression model outperformed other models like Decision Tree, Random Forest, H2O-Gradient Boosting, Gradient Boosting, Light Gradient Boosting Machine (LGBM), and XGBRegressor. SHapley Additive exPlanations (SHAP) analysis was used to interpret the result and explain the workings of the model.

Earlier investigations of musculoskeletal symptoms and disorders in occupational health and safety used statistical methods to analyse data collected through surveys and questionnaires, this method provided little success in the prediction and prevention of MSDs (Chan, Victor CH, et al., 2022), it has been observed from several research works carried out in recent times that machine learning based methods outperform statistical methods in the analysis of occupational musculoskeletal symptoms (Lee, Ju-Yeun, et al., 2023), given that they reliably analyse huge amount of data to identify relationships which subsequently helps in making predictions. However, the reasoning behind these predictions is not explainable to humans. Hines, Brandon, et al., (2022) and Mollaei, Nafiseh, et al., (2022) tried to address this by introducing explainability using XAI, which is an improvement but results from these studies cannot be generalized across industries as the research were peculiar to specific industries, also as at the time of this review, no framework was found to be used in assessing the contribution of AI (Pishgar, M, et al., 2021) and the performance of XAIs in Occupational Health and Safety. To address these limitations, this research seeks to not only build on the existing body of knowledge but to investigate and evaluate the performance of relevant XAIs in the field.

3 Research Methodology

3.1 Introduction

This chapter presents the overall strategy and layout adopted in this study which enabled the researcher to achieve the research objectives. It reaffirms and provides an in-depth discussion of the research approach, goals, and questions already presented in the introduction of this dissertation. It also elucidates the techniques and processes employed in the identification, selection, collection, and analysis of the data used in the study.

3.2 Data description

The Ready-Made Garment (RMG) workers dataset employed in this study was obtained from a publicly available source. It was generated in similar research by (Hossain, Mohammad Didar, et al., 2018) and was adapted to meet the purpose of this research. The RMG workers dataset consists of 232 observations categorized under 165 variables, 130(79%) categorical variables, and 35(21%) numerical variables. that relate to information in three different areas.

3.3 Exploratory data analysis

The data analysis for this study involved the use of descriptive and inferential statistics, the former helps to describe the variables in the sample using statistical measures such as the measure of central tendency (mean, mode, and median), the measure of dispersion (range, quartiles, and standard deviation), and frequency distribution plots. The descriptive analysis allows for summarization of data and gives insight into its characteristics (Fisher, M.J. & Marshall, A.P, 2009). Inferential statistics, in contrast, looks at the relationships between variables and generalizes the outcomes from the analysis of a sample to the population (Shane & Cheryl, 2009). Regression and correlation techniques were used to determine the relationships between variables.

A review of the demographic characteristics of the RMG dataset (see table (1)) revealed that data consists of 46male participants with a mean age of 32 and 186 female participants with a mean age of 31, the age data however shows the presence of a few outlying ages in the female age data. The mean weight of the male participants was 58.9 with a standard deviation of ±9 and female 54.2(8.8). Overall, male, and female employees worked the same mean overtime hours per month with male employees on average, taking home a higher total monthly income. The reason for the variance in the total monthly income between male and female employees could not be directly deduced from the collected data.

		Sex	
		Male	Female
Age	Count	46	186
	Mean	32	31
	Standard Deviation	7	7
	Maximum	52	60
	Minimum	23	20
Height	Mean	1.6	1.5
	Standard Deviation	.1	.1
Weight	Mean	58.9	54.2
	Standard Deviation	9.0	8.8
BMI	Mean	22.38	23.80
	Standard Deviation	3.23	3.82
Overtime Hour per Day	Mean	2	2
Total Monthly income	Mean	15909	15007
Overall risk exposure	Mean	.5938	.5599

Table 1. Demographic characteristics of the participants.

The data also showed that male employees have higher mean overall risk exposure making them more predisposed to having musculoskeletal symptoms and disorders. Married male employees made up 75% of the entire male participants and 73% of female participants were married. Comparatively high number of the participants were from sewing department, 25% male and 45% female. The spread of gender and marital status across the departments is shown in Table 2 and Figure 1.

Analysis of work type

		Marital Status									
		Married		Unmarried		Widowed		Divorced		Separated	
		Sex		Sex		Sex		Sex		Sex	
		Male	Female	Male	Female	Male	Female	Male	Female	Male	Female
Type of work	Cutting	6	1	2	0	0	0	0	0	0	0
	Sewing	9	54	3	8	0	6	0	4	0	12
	Tailor	1	2	0	0	0	1	0	0	0	0
	Thread Cutting	0	6	0	0	0	0	0	0	0	1
	Machine Operator	6	16	4	4	0	0	0	0	0	0
	Ironing	3	0	0	0	0	0	0	0	0	0
	Finishing	6	26	1	0	0	0	0	0	0	0
	Poly	1	2	0	2	0	0	0	0	0	0
	Checking	0	13	0	1	0	1	0	1	0	2
	Mending	0	4	0	2	0	0	0	0	0	0
	Cleaning	0	3	0	0	0	2	0	0	0	0
	Folding	0	3	0	0	0	1	0	0	0	1
	Other	4	6	0	0	0	0	0	0	0	1

Table 2. Analysis of work type showing participants' gender and marital status distributions across departments

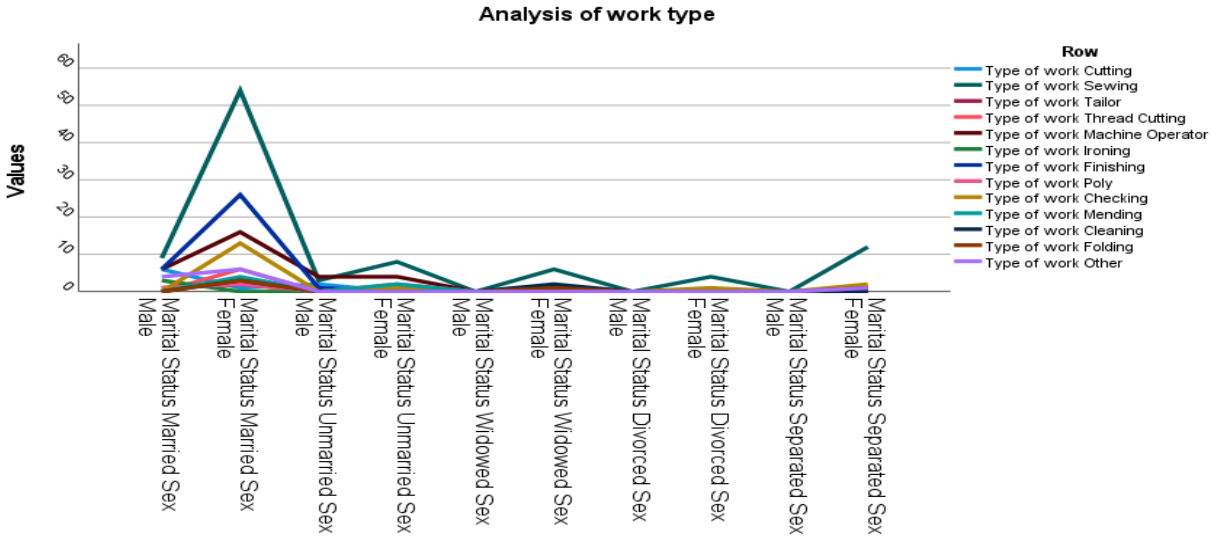


Figure 1. visual representation of participants' gender and marital status distributions across departments

A review of the information on work-related musculoskeletal symptoms and disorders collated using the NMQ questionnaire shows that neck and lower back pain are the most prevalent among participants, accounting for 24% and 21% of all the pain or injuries experienced by participants (Fig 2.). Further investigations revealed that neck (29%) and wrist/hand (8%) pain or injuries are more common with male participants in comparison to female participants with 25% and 7% respectively, other pain/injury types are more common with female participants (Table 3).

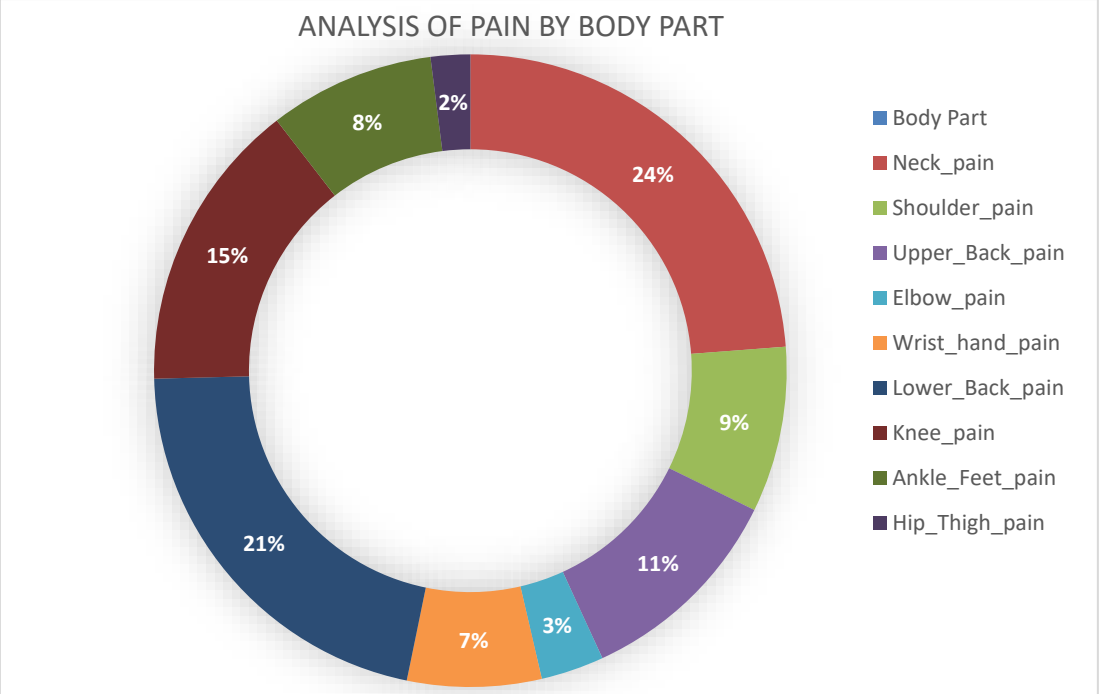


Figure 2. Analysis of the prevalence of pain by body parts

Body Part	Male	Female
Neck_pain	14(29%)	45(25%)
Shoulder_pain	2(4%)	19(10%)
Upper_Back_pain	5(10%)	22(12%)
Elbow_pain	2(4%)	6(3%)
Wrist_hand_pain	4(8%)	13(7%)
Lower_Back_pain	5(10%)	48(26%)
Knee_pain	7(15%)	30(16%)
Ankle_Feet_pain	2(4%)	19(10%)
Hip_Thigh_pain	0	5(3%)

Table 3. shows the categorization of pain/injury by gender.

Workplace productivity was examined by taking a critical look at the WMSD by body parts to determine which pain type contributes more to sick leave taken. Figure 3 reveals that though neck, back, and knee pain/injuries are the most common they contribute less to sick leave taken by employees. Hip, wrist, and elbow pains account for a higher number of sick leave taken.

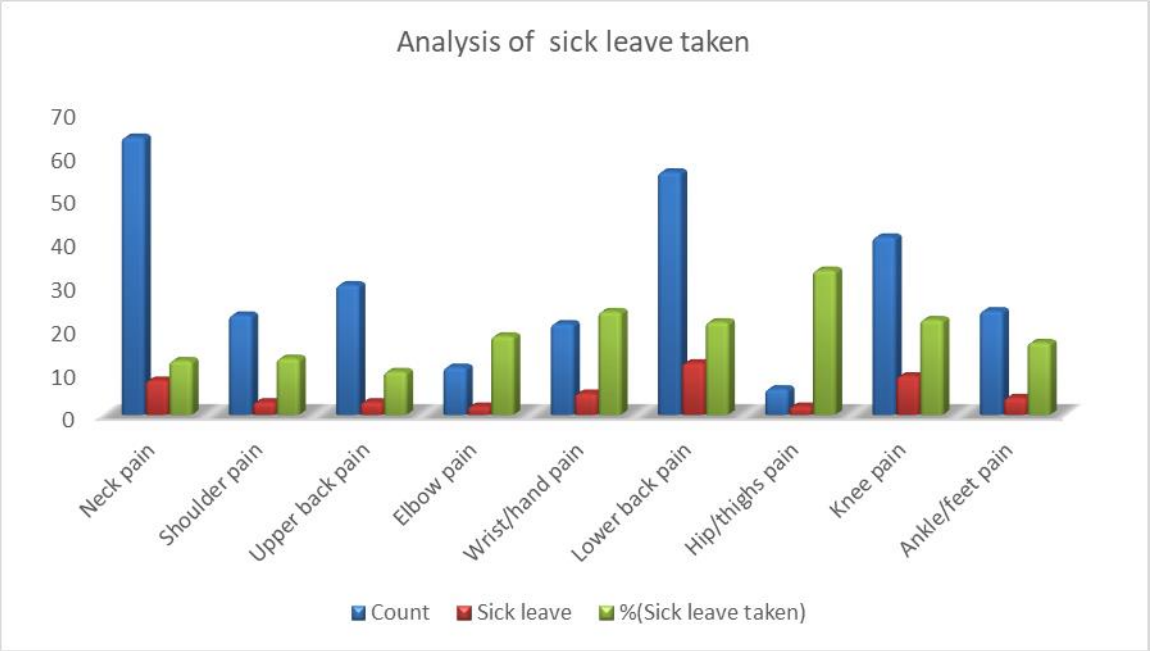


Figure 3. Analysis of sick leave taken due to injuries to different body parts.

The data pre-processing phase was initiated by data cleaning, which involves removing any irrelevant data from the dataset to reduce noise elements and inconsistencies which could adversely affect the machine learning model (Hariri, R.H, et al., 2019), to achieve this some basic cleaning techniques were adopted by visually reviewing the data, some redundant samples and features were identified which resulted in the deletion of columns and rows. The features were also investigated for the presence of outliers, which are data points with extreme values outside the expected range. This was done by using boxplots and scatterplots, outliers were detected for both male and female participants for age, height, weight, BMI, and monthly income. The identified outliers were analysed and then removed as they were not likely to be real-world values and to ensure that the remaining data is a true representation of the population. The dataset was reviewed for missing value, fourteen (14) data points were found to be missing from the dependent variable, the missingness being entirely at random (Missing at Random -MAR) and relatively small to the value of the entire sample (van der Heijden, et al., 2006), a complete case analysis also referred to as listwise deletion was used to handle the case of missing value by deleting the affected observation.

Exploring the data through univariate analysis revealed that weight, height, age, and BMI had normal or close-to-normal distribution which is an important characteristic for linear regression and support vector machine models, this was followed by performing multivariate analysis to gain insight into the relationship between independent features and dependent feature. The analysis revealed a strong positive correlation between these independent features (risk score of work pace, back injury, shoulder/arm injury, wrist/hand injury) and the dependent feature (total risk exposure), a lesser positive correlation was recorded for back dynamic, back static, stress, and vibration. This is an indication that these features are critical in predicting the independent feature.

Further investigation was done using a heatmap to confirm collinearity among the input features and to identify and select features that have the highest influence on the output variable. Collinearity refers to the linear relation among two or more independent variables, sometimes referred to as multicollinearity, it shows the degree to which the independent variables are related, and the presence could lead to a problem with trusting the estimates of the model's parameters (Alin, 2010). High collinearity can have a sizeable impact on a regression model's results (Johnston, Ron, et al., 2018) making it difficult to interpret models result in comparison with model expectation. The analysis from the map shows no evidence of high collinearity among independent features, this asserts that each independent feature has a unique characteristic that contributes to the prediction of the dependent feature.

Data normalization was achieved through standard scaling by using StandardScaler() function from the SKlearn library, this technique is also regarded as the z-score normalization. It transforms the data such that each feature has a zero mean and unit standard deviation, to achieve this, the mean value of each feature is subtracted from each data point and the resulting figure is then divided by the standard deviation of the feature. The normalization technique ensures that the weight of different features is comparable by not assigning higher weights to features with larger values. The sole aim of data normalization is to make the data have a uniform scale as it supports the assumption of linear regression and support vector machines having features that are of similar scale and are normally distributed (Poole, Michael A & O'Farrell, Patrick N,

1971). This was followed by training different machine learning models using the normalized data.

3.4 Model selection

The model-building process is a crucial step in any machine-learning project. It involves selecting appropriate algorithms, fine-tuning the hyperparameters, and then training them on the already pre-processed and normalized data.

This section highlights the steps taken to select the algorithms, their strengths and weaknesses and the challenges encountered during the model-building process.

3.4.1 Decision Tree (DT)

Decision Tree is a simple but powerful algorithm used in machine learning to solve both classification and regression problems (Fabian Pedregosa, et al., 2011). It is an effective technique that employs a tree-like structure of decisions and outcomes to predict the dependent feature, it works by recursively splitting the data into subsets based on the values of the independent features, each internal node represents a decision about a feature and the leaf node represents the prediction. The sole aim is to identify features with the highest impact on the dependent variable and the optimal split points that would reduce the prediction error (Breiman, 1984).
It uses either of the following algorithms, the ID3(Iterative Dichotomiser 3) developed in 1986 by Ross Quinlan, which creates a multiway tree for discovering each node (Quinlan, 1986), the C4.5 which allows for features to be numerical variables, thereby eliminating the restriction imposed by ID3 that permits only categorical features. This was quickly followed by an update, C5.0 which is more accurate and uses less memory (Quinlan, 2014), or CART (Classification and Regression Trees) that supports numerical output variables, does not calculate decision rules sets but creates binary trees which use features and threshold with the highest information gain for a node (Breiman, 1984).

The advantages of using DT are that it is a white box model, easy to understand and interpret, and its internal workings can be visualized. It does not require input data to be normalized, can handle both categorical and numerical data, and deal with multiple output problems. However, it can be unstable as small changes in the input data could

result in a completely different new tree. It can also create intricate trees which may not be able to adequately generalize the data (Fabian Pedregosa, et al., 2011). The decision tree regressor function from Sci-kit learn was used to develop the model, Sci-kit learn uses an optimised CART Decision Tree algorithm.

3.4.2 Random Forest

Random Forest is an ensemble of several decision trees. It combines multiple trees to improve the model's accuracy and lessen the variance. Created by Breiman, Leo., (2001), it operates by training each set of trees on a random subset of data and features through bootstrap aggregation (*known as bagging*) that helps to minimise overfitting and maximise the generalization of the model. In bagging, sets of decision trees are created with each tree in the set selected from a bootstrap sample (A bootstrap sample is created through random sampling with replacement) of the training data (Brownlee, 2016). The final prediction is made by averaging predictions (for regression) from all the trees in the ensemble and this gives a better prediction accuracy comparable to prediction from a single decision tree. The output (average prediction) is computed by the mathematical expression as denoted in (Katuwal, R, et al., 2020)

$$Z = argmax \frac{1}{T} \sum_{t=1}^{T} P_t \left(\frac{y}{x} \right)$$

where $P_t(y|x)$ is the probability distribution of individual tree (t), and x is a set of test samples.

The earliest work on a decision tree ensemble method was presented by Kwok, S.W & Carter, C, (1990), in their works, the final prediction was achieved by averaging across multiple decision trees with different structures, however, the selection of trees was neither random nor automatic but instead based on manually selecting at the top of the tree split and expanding using the ID3 technique. Kong, E.B. & Dietterich, T.G, (1995) improved on the works of Kwok, S.W & Carter, C, (1990) by introducing the randomization of the best splits at each node, this was achieved by a uniform random selection out of the best 20 splits at the given node. This approach gives a better result

than bagging in a low setting. It has been shown that bagging performs better with noise, other variants also exist (Probst, Philipp, et al., 2019).

Many advantages have been recorded for random forest compared to the single decision tree model, they include a reduction in overfitting and better generalisation, better performance of the model even in the presence of large dimensionality and non-linear relationship between the features, it can handle missing values by simply ignoring it and splitting the data along the remaining features. It is, however, likely to experience overfitting with many trees or trees too deep. These shortcomings can be managed efficiently using cross-validation or hyperparameter tuning. Another shortfall of random forest is that of reduced interpretability resulting from the combination of multiple decision trees which makes it complex and difficult to visualize.

Random Forest algorithm employs bagging and CART (Classification and Regression Trees) techniques already mentioned, in this study the RF model was built using Sci-kit random forest regressor implemented using the CART as the base estimator (Fabian Pedregosa, et al., 2011).

3.4.3 Support Vector Machine.

Support Vector Machine (SVM) is a supervised learning algorithm capable of handling both classification and regression problems (Smola, A.J & Schölkopf, B, 2004). SVM employed in regression analysis is referred to as Support Vector Regression (SVR) and could be either linear or non-linear depending on the Kernel functions used (Kavitha, S., et al., 2016). The function could be either, linear, polynomial, or radial base (RBF) or others. The kernel function acts by changing the input data into a higher dimensional feature space using a hyperplane construct to separate data points into positive and negative classes (Gunn, S.R, 1998). The model is trained to determine the hyperplane that best fits the data points by reducing the distance (margin) between the plane and the data points. The SVM seeks to maximise the margin.

SVMs can be linear if the data points are linearly separable or non-linear if the data points are not linearly separable. The non-linear SVM uses the kernel function to transform the data from a lower-dimension space to a higher-dimensional space

where it becomes linearly separable (Gunn, S.R, 1998). The SVM function can be expressed mathematically as follows:

$$W = f(y) = \Gamma\left(\sum_{i=1}^{N} y_i p_i K\left(x_i x'\right) + C\right)$$

Where $K(x_i,x')$ represents the kernel, C and p_i are the hyperplane parameters and Γ is the sign function (Jabeur, S.B, et al., 2021).

The SVM algorithm is popular among researchers because it can handle non-linear relationships between the dependent and independent features with the aid of the kernel functions. It is unaffected by outliers using the margin which ensures that outliers do not influence the decision boundary. It is flexible using different kernel functions; it can handle different relationship types and is known to be very scalable as it can manage data with high dimensions making it appropriate for solving complex regression problems and computationally efficient (Kavitha, S., et al., 2016). The algorithm, however, has some shortcomings, it is sensitive to the choice of kernel function and hyperparameters used, and the choice of KF is usually dependent on the relationship between the input and output features and the characteristics of the data. It also requires a lot of storage space to save the support vectors which is a challenge when working with big data.

3.4.4 Gradient Boosting Regressor (GBR)

Proposed by Friedman (2001, 2002) is a machine learning algorithm that belongs to the ensemble method family. Based on decision trees, it seeks to reduce the loss function (Mean Square Error-MSE) through "boosting" by iteratively building a sequence of tree models such that each new tree learns from the error of the preceding tree (Otchere, D.A., et al., 2022) and then updates the prediction based on the residual. The residuals are calculated by deducting the actual values from the predicted values. Some of the advantages that have been recorded for GBR include its ability to deal with both continuous and categorical features making it ideal for a broad ambit of regression problems. It can identify non-linear relationships between independent and dependent features and this characteristic accounts for its being a more accurate

linear regression model. It has also been found to be less perceptive to outliers in data. The model, however, suffers from overfitting with high parameters or several trees and high computational cost when large data sets are involved (Otchere, D.A., et al., 2022).

The approximation function for a GBR according to Freidman (2002) is expressed below:

$$W = f(x_i) = \sum_{i=1}^{N} \beta_i h\,(x; b_i)$$

The function h(x;b_i) represents the poor-performing base learner, (x) is the explanatory variable and β the coefficient of expansion and b_i represents the parameters of the model (Jabeur, S.B, et al., 2021)

3.4.5 Light Gradient Boosting Machine (LGBM)

LGBM is an efficient gradient boosting framework based on decision trees and the concept of combining weak learners (Fan, J, et al., 2019)developed by Microsoft in 2017((Ke, G, et al., 2017) is well suited for large datasets and efficiency and has been adopted across many fields because of its computational efficiency and outstanding ability to reduce over-fitting challenges. LGBM employs a histogram-based technique to split the data into bins to make the training process faster and reduce memory storage space, thereby reducing the overall computational cost of determining the best-split point (Fan, J, et al., 2019). The underlying calculations for LGBM are described by (Sun, X, et al., 2020) and theory by (Fan, J, et al., 2019, Ke, G, et al., 2017).

3.4.6 CatBoost Regressor

Is an open-source machine learning algorithm developed by Yandex (Safarov, R.Z., et al., 2020) that belongs to the Gradient Boosting Decision Tree (GBDT) ensemble techniques, built to effectively handle tasks that incorporate categorical and heterogeneous data with the least information loss (Hancock, J.T. & Khoshgoftaar, T.M, 2020). Proposed by Prokhorenkova, L., et al.,(2018)and Dorogush, A.V, et al., (2018), unlike other boosting algorithms, it uses a more efficient variant of the boosting algorithm known as ordered boosting, which involves using a mix of three approaches, (1) Adopts one-hot encoding to transform categorical features to binary features and (2) Decision tree to manage interactions between the features, and (3) permutation-

based feature importance computation to achieve better accuracy and reduced overfitting (Jabeur, S.B, et al., 2021).

The algorithm creates a sequence of decision trees with each working to minimise the residual error from prior trees, the residual error being the difference between the predicted and actual value. The loss function gradient is calculated for the predicted value which is subsequently used to update the weights of the decision tree.

The advantages recorded for CatBoost include its ability to efficiently handle categorical variables (Jabeur, S.B, et al., 2021), high performance (Hancock, J.T. & Khoshgoftaar, T.M, 2020), and can perform random permutations to evaluate leaf values when selecting the tree structure for the sake of overcoming overfitting (Dorogush, A.V, et al., 2018). Some of the recording shortcomings include high computational costs arising from the time required for training and storage requirements and complexity due to the number of hyperparameters needed to be tuned to achieve improved performance.

3.4.7 XGBoost(eXtreme Gradient Boosting) Regressor

is an efficient supervised learning algorithm, based on the Friedman (2001) and Fried et al (2002) gradient boosting framework with novel modifications such as the addition of a regularization term, a customized loss function, and an optimized algorithm. Developed by Chen and Guestrin (2016), it uses a regularized model formalization technique to curb overfitting and consequently improve model performance (Abdullahi & Abullahi). Xgboost is widely used by researchers as it is known to be up to 10 times faster than other common algorithms. Its popularity comes from the fact that the algorithm is scalable under all scenarios and the “scalability” property is brought about by many essential optimizations of the algorithm. The optimizations include a new tree algorithm capable of handling sparse data, parallel and distributed computing systems that allows for quicker learning by the model (Guestrin, C., 2016, August. Xgboost: A scalable).

It is adaptable to both regression and classification tasks and highly efficient in handling missing values and categorical features. Other advantages include the existence of functionalities like feature importance ranking, cross-validation, and early stopping. Conversely, the algorithm has several hyperparameters that need to be

tuned therefore making it difficult and time-consuming to find the optimal hyperparameters the optimal hyperparameter.

3.4.8 Artificial Neural Network (ANN)

Artificial Neural Network is a machine learning algorithm that is modelled to mimic the functions and structure of the human brain (A. K. Jain, Jianchang). It is made up of several layers of interconnected nodes like neurons, and each node consists of three critical elements: a node character, network structure, and a set of learning rules. The node character sets out the number of inputs/outputs, the weight for each input/output, and the activation function, the network structure determines how the nodes are connected and the number of layers involved while the learning rules define how the weights are allocated and initialized (Zou, Yi Han). The nodes in the input layer and output layers receive and send out data. Between the two external layers lies the hidden layer where complex computations and mapping of input data the output are performed (Zou, Yi Han).

ANN algorithms are well-suited for both classification and regression problems (Ray, S., 2019). For a regression model, the approach involves using the training dataset to train the network and using an optimisation function iteratively to adapt the weight between the nodes to reduce the error between the actual and predicted value The ANN model for regression problems involves training a neural network to map the input data to the output data (Ray, S., 2019).

The ANN algorithm is widely used because of its extraordinary flexibility, it can handle large datasets, missing values, and complex non-linear relationships between features. It can model complex relationships between features. It can execute feature selection and can work with different data types. However, it can be time-consuming and would require some level of expertise to deploy an optimal ANN model that would produce the best accuracy, as this will involve determining the best network structure and hyperparameters such number of nodes, layers, epochs, and the earning rate for the modeling. Other shortcomings include overfitting which can be resolved using regularizations, dropout, and early stopping during model training, and the difficulty in interpreting the model's prediction.

ANN model implementation involved defining the structure of the neural network and adopting an appropriate activation function for each layer. The rectified linear (ReLU) activation function was set for the first three layers and a linear activation function for the output layer using the Keras sequential function from the Tensorflow library, The developed model was trained using the training dataset. The model compilation was achieved using the Adam optimizer, and the loss was defined as Mean Square Error (MSE).

3.5 Model implementation

The highlighted goal of this study is to evaluate how well some XAIs perform in providing explanations to the predictions made by a black box ML model with a view to adopting these XAIs in the field of OHS for the prevention of musculoskeletal symptoms and disorders in the workplace. To achieve this, some of the machine learning algorithms that have been adopted in similar studies and are known to have optimal performance were selected. The following machine learning algorithms DT, RF, GradBoost, LightGBM, CatBoost, XGBoost, and a deep learning algorithm, ANN were employed.

To train the models, the RMG dataset was separated into two parts, the independent features represented by (X) and the dependent feature (y), this was followed by splitting both to train and test datasets. 67% of the dataset was considered as training data, consisting of 146 records relating to the RMG workers, and the remaining 72 records which make up 33% of the dataset were used as test data. The normalized data (normalization was achieved through standard scaling) was then used to train and test the model. Hyperparameter tuning was performed for some of the models to achieve better performance. For the RF model, tunning was implemented using the sci-kit learn RandomizedSearchCV function, this function randomly searches the defined parameters within a set range and creates multiple models with these parameters. The parameters, n-estimator which refers to the number of decision trees in the forest was set to (10, 50, 100, and 200), and the depth of each tree (max-depth) was set to (None,5, 10, 20), and using 5-fold cross-validation to evaluate the performance of each combination, cross-validation of 5 splits the training data into 5 equal parts, training was done using 4parts and one part was used to test the model. The prediction score

was calculated for each model and the best model was finally determined by averaging the scores. The hyperparameter grid search presented the following parameter as the best {'n_estimators': 10, 'min_samples_split': 2, 'min_samples_leaf': 1, 'max_features': 'sqrt', 'max_depth': 20}. Default parameters were used for fitting the SVR and DT models as these yielded the expected level of performance. Some of the hyperparameters like the learning rate and n-estimators were altered from the default values for ensemble models. To implement the ANN model, the Keras sequential function was deployed in structuring the network architecture. The network consisted of 4 dense (each node being fully connected) layers with the first three layers having an output shape of 128 nodes and the output layer having one node. In total, there were 50,945 trainable parameters. The model compilation was executed using an Adam optimizer and MSE as a loss function.

3.6 Measures for model evaluation

To conclude any machine learning regression project, the performance of the trained models needs to be assessed and compared using standard evaluation metrics. The metrics compare the trained model's predictions on the test dataset to the prediction with the actual data (Botchkarev, 2018).

The difference between the actual data points and the line of best fit for the models represents the model's error. For multiple data points, the error can be computed using standard regression evaluation metrics. The Mean Squared Error (MSE), the Root Mean Squared Error (RMSE), and the coefficient of determination or R-squared were initially used to evaluate the performance of the models in this study. The metric function from Scikit-learn was used to calculate the MSE, which is a measure of the average squared difference between the predicted and actual values, the R2 score, which is a measure of how well the model fits the data and lastly, the RMSE was calculated by taking the square root of the MSE.

Given that X_i and Y_i are the predicted and actual values respectively for the ith value, then the aforementioned metrics can be expressed mathematically as:

- **The Mean Squared Error (MSE)** measures the proximity of the best-fit line (regression line) to a set of data points. A smaller MSE is more desired as it tells that the data points are less dispersed around the mean, an indication of a better fit(performance)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2$$

…… (1)

- **The Root Mean Squared Error (RMSE)** is simply the square root of the MSE. It a more desired metric as it has the same unit as the independent feature making it easy to interpret the error.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - Y_i)^2}$$

…… (2)

- **The coefficient of determination or R-squared** for a regression model prediction describes the proportion of the variance in dependent features that can be predicted from the independent feature.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(X_i - Y_i)^2}{\sum_{i=1}^{N}\left(\bar{Y}_i - Y_i\right)^2}$$

…… (3)

Where $\bar{Y}$ is the mean of the actual values, defined as:

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$

In the study by Willmott, Cort J & Matsuura, Kenji(2005) to review the accuracy in the adoption of error measures for models' performance, they found MSE and its variant RMSE not to be an appropriate measure of average error, as the RMSE varies with the variability in the error distribution and the square root of the number of error and a function of three characteristic error sets. They recommended the adoption of

- **The Mean Absolute Error (MAE)** as the appropriate measure of error, this position was, however, countered by Chai & Draxler, Roland R(2014). In view

of the foregoing, the MSE which computes the absolute difference between the actual and predicted values was also adopted in this study. It is expressed mathematically as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|X_i - Y_i| \quad \text{..... (4)}$$

3.7 Measures for explainable AI evaluation

As Explainable AIs are becoming widely adopted to bring clarity to the predictions of black box models, it has become imperative to have a set of standardized metrics to evaluate their performance. The evaluation can be based on the specific need or purpose like confirming how faithful an explanation is to the black box model and how well it has served the needs of the user. The following measures which fall under the two broad categories of explainable AI measures (1) objective or computer-based evaluations and (2) human-centred evaluations have been adopted in most studies (Lopes, P, et al. 2022; Mohseni, S, et al., 2021) and provided the basis for the evaluation and subsequent adoption of XAI in this study.

Human-centred evaluation methods

Refer to evaluation measures involving human interaction with the system to confirm its suitability and performance. Users' feedback is an essential determinant of the XAI's strengths and weaknesses. Interviews, scales, think-aloud expressions, and surveys are some of the tools used in human-centred evaluations. The following metrics come under this category.

- *Trust*

Is a measure of the user's comfort in using a system and which tends to develop over time with continued usage. It is based on a user's perceived judgment that the system has the capability of performing a given task and agreement to accept the result of the system.

- *Usefulness and Satisfaction*

Usefulness measures how valuable the information provided by the system is to the user and the level to which the explanations meet the expectations of the users determines explanation satisfaction. Both qualitative measures are mostly determined using questionnaires, interviews, and surveys within a study group.

- *Understandability*

This measures how well the user can discern the workings of the model after explanation. The user's knowledge of the relationships between the input and output features, the model's behaviour, and expected prediction are pointers to understandability.

- *Task Performance:*

This focuses on user experience and confidence in the system, it confirms how explainability can lead to improved task completion. Task performance is based on factors such as success rate, accuracy of explanation, model tunning, and task completion time.

Computer-centred evaluation methods

The computer-centred evaluation method unlike the human-centred method directly assesses the performance of the XAI by using mainly qualitative means. It measures the interpretability, correctness, and completeness of the explanation presented by the AI. The computer-centred approach is broadly based on these two measures.

- *Interpretability*

Is a measure of how understandable the explanation presented by the XAI is to humans. The interpretability of an XAI can be ascertained by the absence of vagueness and the level of simplicity.

- *Fidelity*

This relates to the degree of accuracy with which a model's behaviour is depicted through an explanation. It measures how well the explanations provided by the XAI model replicate the actual decision-making process of the model. A high-fidelity metric means that the explanations are accurate and reflect the internal workings of the model.

4 Results and Discussion

4.1 Introduction

In this chapter, the performance of the models adopted in the study in predicting the overall risk exposure of workers is evaluated to determine which model is best suited for the task. The overall risk exposure level is a measure of how predisposed a worker is to WMSDs. Following the determination of the best model, four XAIs were deployed to provide explanations for the black box predictions. Finally, the effectiveness of the XAIs in providing easy and reliable explanations was assessed to determine the best XAI to adopt.

4.2 Review of model performance

The performance of the eight models used in the study was evaluated using some well-known regression metrics like the RMSE, MSE, MAE, and R2. Table 4., shows a summary of the results across the models, The results were then analysed comparatively to determine the best model for the task. Table 4 and Figure 4 shows that the Gradient Boosting Regressor and Category Boosting (CatBoost) Regressor models had the best performance in terms of both RMSE followed by Random Forest. ANN gave the worst performance with the same metric (RMSE). The smaller the RMSE value the better the performance of the model, as a small RMSE means less disparity between the actual and the predicted data points. The CatBoost Regressor gave the best MAE value of (0.008722), followed by Gradient Boosting Regressor (0.010023) and Random Forest (0.014313).

Model	MSE	RMSE	MAE	R2
Decision Tree	0.000762	0.027601	0.019965	0.903307
Random Forest	0.000348	0.018654	0.014313	0.955832
Support Vector Machine	0.002823	0.053131	0.043727	0.641708
GradientBoostingRegressor	0.000209	0.014447	0.010023	0.973506
Lightgbm	0.000498	0.022314	0.017245	0.936801
CatBoostRegressor	0.000211	0.014501	0.008722	0.973311
XGBRegressor	0.000557	0.023598	0.016773	0.929321
ANN	0.014019	0.118401	0.089896	-0.779395

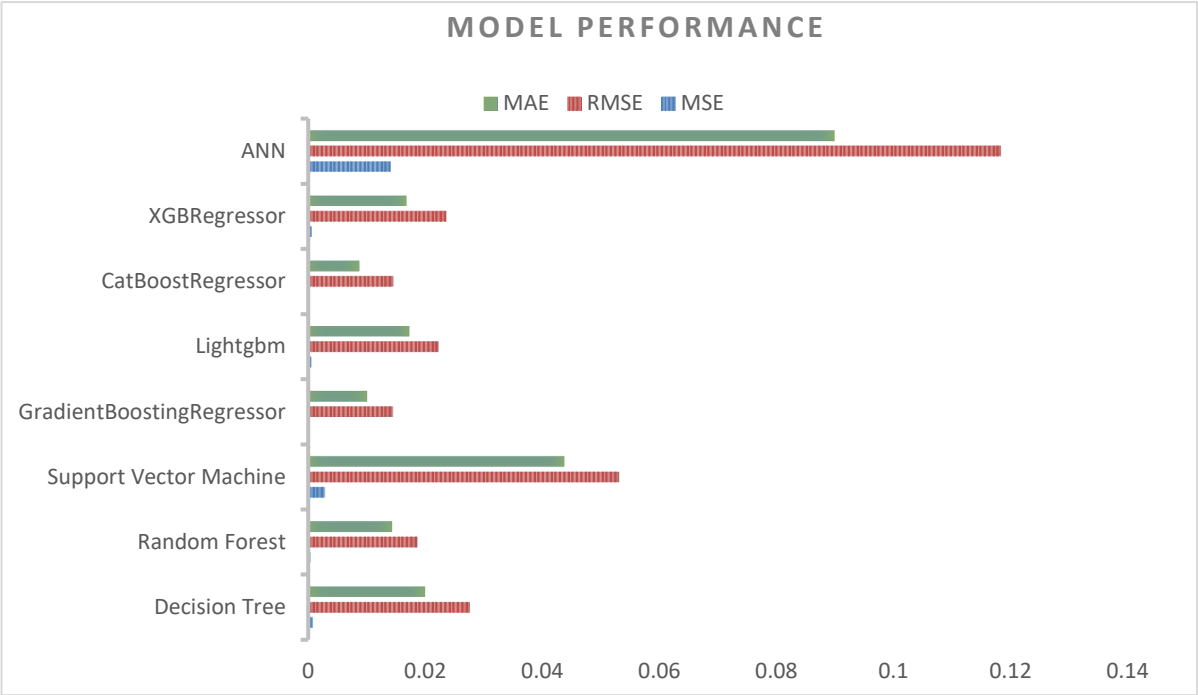
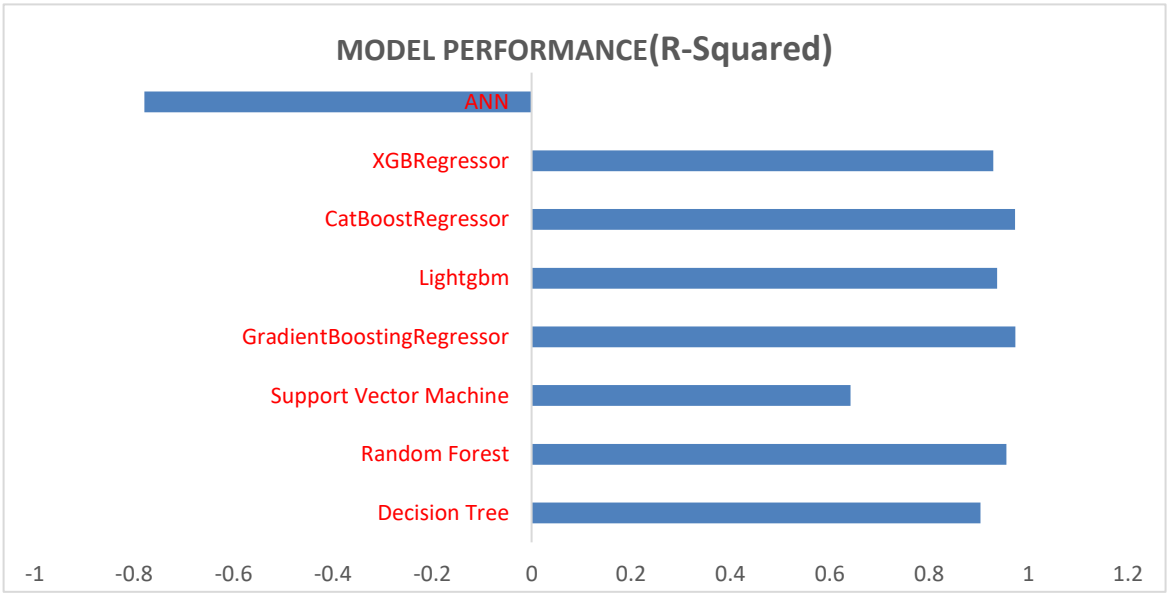


Figure 4. Performance across the models using evaluation metrics (MAE, MSE and RMSE)

Reviewing the R2 scores, Figure 5 shows the performance of the models in terms of this metric. Again, the Gradient Boosting Regressor had the performance with a score



of (0.973506) followed by CatBoost Regressor (0.973311) and Random Forest (0.955832). The R2 score indicates the measure of the variance in the target feature that is predictable from the independent features, in other words, it is a measure of how well the model predicts the observed/actual data. An R2 score close to 100% is highly desirable. The ANN model with a negative R2 score (-0.779395) again is the worst-performing model.

Figure 5. Analysis of models' performance using R2 score

Overall, the findings submit that machine learning models are effective tools for predicting musculoskeletal symptoms and diseases in occupational health settings. However, the choice model and evaluation metrics could have a substantial impact on performance.

The study revealed that the Gradient Boosting Regressor was the best-performing model, with RMSE (0.014447) and R2 score (0.973506) as the most informative metrics for assessing model performance. Figure 6(a & b) shows how closely the prediction of the Gradient Boosting Regressor model was able to mirror the actual values. Figure 6a shows the relationship between the actual and predicted values and the proximity of the points to the line of best fit(Y=X) is an indication of how well the model has performed. Figure 6b compares the distribution of the actual values and the predicted values, the closeness of the two distributions is an indication of how well the model performed.

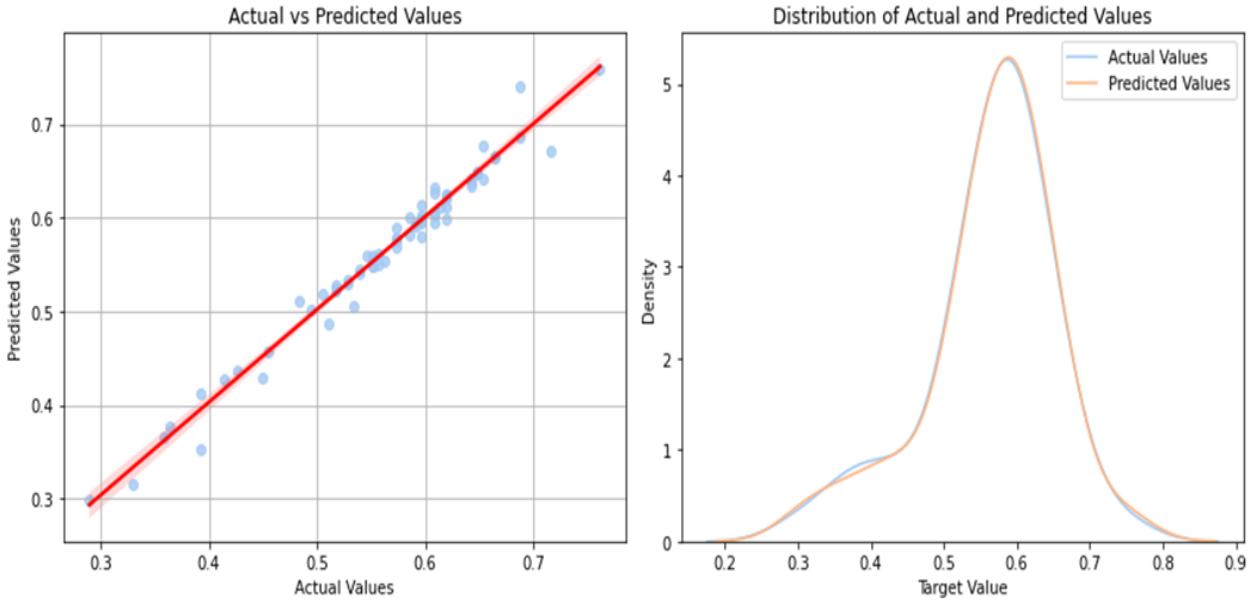


Figure 6a. scatterplot and Figure 6b. distribution plot, both showing the performance of the Gradient Boosting Regressor model.

4.3 Model analysis using post-hoc explainability techniques.

Machine learning has progressed tremendously in recent years and has led to the revival of artificial intelligence and the development of extremely sophisticated models that have found application in all fields, however, there is a huge trade-off between sophistication and explainability. To address this, a novel research field has emerged with interest in explaining the predictions of complex models (Holzinger, Andreas, et al., 2022). The field of XAI ensures that the implementation and adoption of AI technologies are safe, responsible, and ethical by developing tools and methods that are fundamental to achieving this objective. There are several explainers currently in existence and these are categorized based on the following criteria as presented in the work of (Vilone, Giulia & Longo, Luca, 2021).

Usage: This refers to the instance at which the explanation is provided. It can either be the Ante-hoc method, where the explainer is inbuilt and forms part of the predictor model, or the post-hoc method, where the black box model uses an external explainer to interpret its prediction.

Scope: This refers to the focus and element of the explanation provided, the scope can be local in which case the explanation focuses on a specific inference of a model, or Global where the whole inferential process of a model is explained

Output/Input format: Different explanation formats exist for XAI methods, they include textual, rules, numerical, visual, or mixed formats. Similarly different explainers take on different input data types such as Pictorial, categorical/numerical, time-series, or textual.

Problem types: Explainers can also be categorised based on the type of problem they are capable of handling. It can be regression, classification, or both.

Model Agnostic Approach

These are algorithms developed to provide explanations for black-box models. They are unconcerned with the internal mechanism of the model they are trying to explain but only work with the predictor's output (Holzinger, Andreas, et al., 2022) and are designed such that they can function with any learning technique, making them highly flexible and adaptable (Ribeiro, Marco Tulio, et al., 2016) but they may be limited in

the input, problem, or output types they are capable of handling (Vilone, Giulia & Longo, Luca, 2021). The mode-agnostic approach provides explanations using perturbation and alteration of the input data to obtain the sensitivity of the performance of the altered data relative to the model performance.

Model Specific Approach

In contrast to the agnostic explainer approach discussed above, the model-specific explainer approach is specific to a model or class of models. The approach is dependent on the internal structure of the model. It learns from the model to provide an explanation using reverse engineering. The model-specific approach allows for an in-depth understanding of the model's prediction (Vilone, Giulia & Longo, Luca, 2020).

Explanation methods can further be subcategorized based on the type of interpretable explainer employed. The subcategorization was presented in Guidotti, Riccardo, et al.(2018).

4.3.1 Implementation of post-hoc explainable AI

Eight models were earlier trained and tested with their performance evaluated to determine the best model using RMSE, MAE, and R2-score as evaluation metrics. The Gradient Boosting Regressor model outperformed other models with RMSE=0.014447, MAE= 0.010023, and R2 score= 0.973506 to emerge as the best model, these metrics however only show how well the model has performed in predicting values close to the actual data but do not equip the human user with the knowledge or understanding of what constitutes the model's prediction. Consequently, to provide insight into the model's prediction, four explainable artificial intelligence techniques (Local Interpretable Model-agnostic Explanation (LIME), SHapley Additive exPlanations (SHAP), Permutation Importance (PI), and Asymmetric Shapley Values (ASV), were selected to explain the prediction of the best model.

4.3.1.1 Local Interpretable Model-agnostic Explanation (LIME)

The Local Interpretable Model-Agnostic Explanation (LIME) proposed by Ribeiro, Marco Tulio, et al.(2016) is an algorithm capable of explaining the prediction of any regression or classification model through approximation. It does this by creating a local interpretable representation of the predictor model. An interpretation

representation refers to any element that is understandable to humans and can ideally symbolize the features of the model. The overall objective of LIME is to explain the prediction of a black-box model by fitting a surrogate model (interpretable representation) whose predictions are easily explainable (Holzinger, Andreas, et al., 2022). To interpret a complex model, it randomly creates samples (Nx_i) in the neighbourhood of the target feature (x_i) and evaluates them using the predictor model. The model prediction is then approximated locally using a simple linear function that is easily explainable and with their feature importance.

To implement LIME, the tabular lime explainer function from the LIME package was used to explain a data instance from the black box model, this is done by randomly generating a dataset by perturbing the original data point and creating a set of synthetic datasets. It then trains a simple interpretable model on the synthetic dataset as the local surrogate model. The trained local surrogate model imitates the attributes of the gradient-boosting regressor model (black box). It explains the prediction of the local surrogate model by analysing the model employing importance scores which represent the contribution of each feature to the prediction. Figures 7a and 7b show a set of feature importance scores that help to explain how the black box model made its prediction. Each of the contributing features have assigned weights and grouped as having either positive or negative impact on the prediction. A lower negative value for a contributing future will adversely affect the model's prediction probability. Conversely, the 'Risk_score_back', has an assigned weight of (**0.04771**) and a positive impact on the prediction outcome. A higher value for this feature would increase the prediction probability of the black box model.

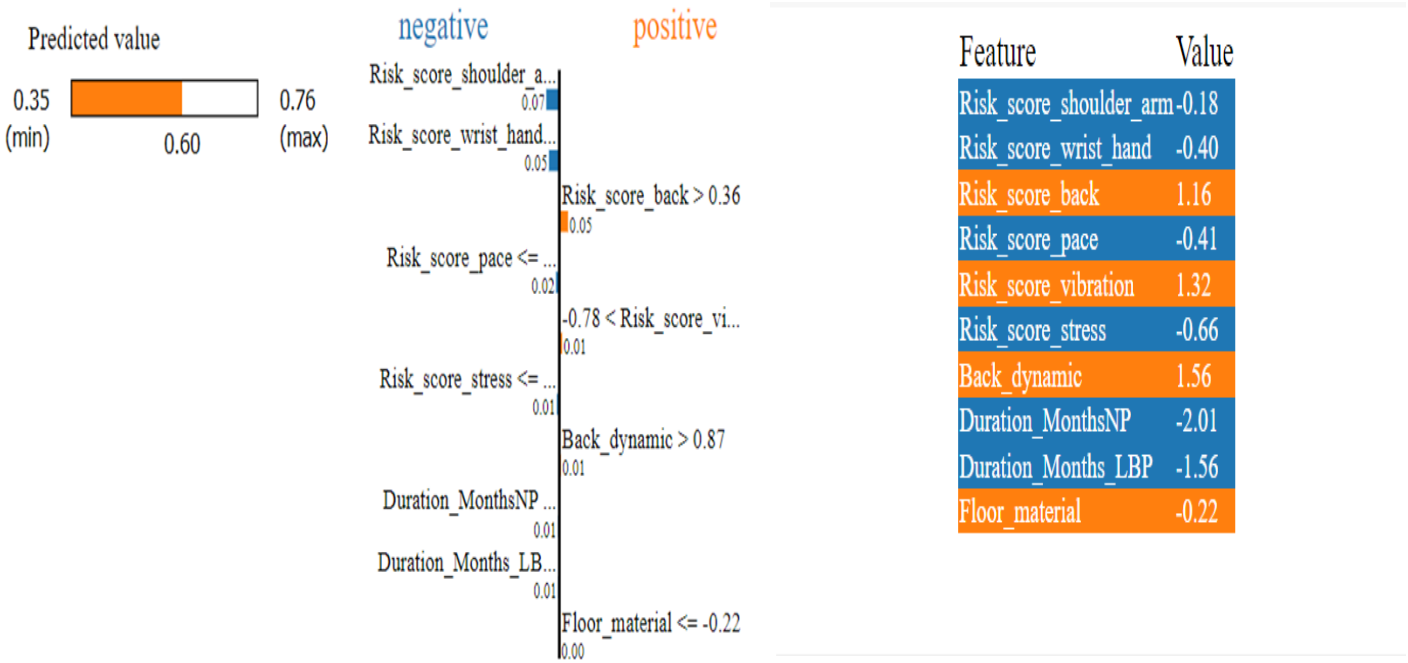


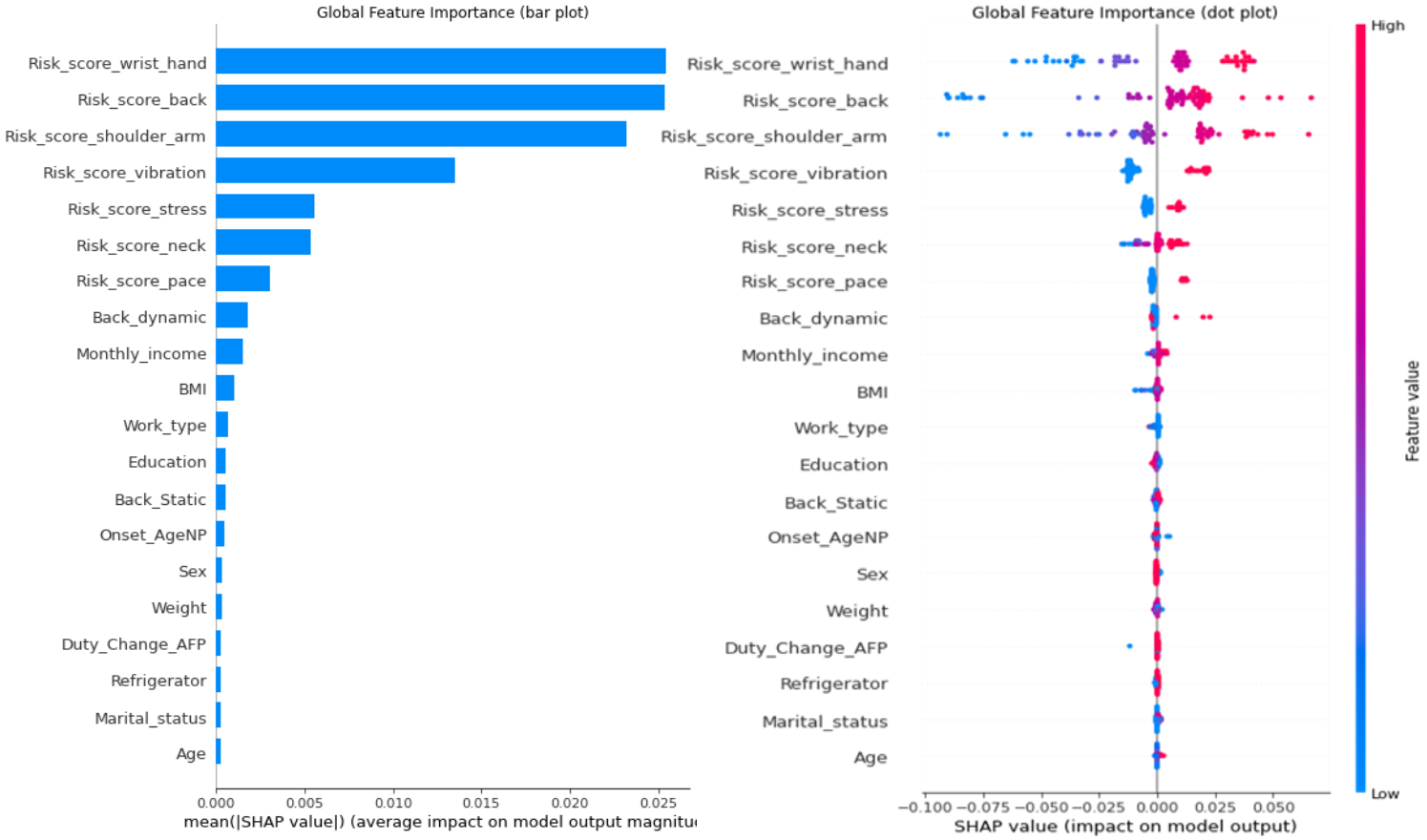
Figure 7a and 7b: LIME explanations showing prediction probability, feature weights and contribution and feature values.

4.3.1.2 SHapley Additive exPlanations (SHAP)

SHAP (SHapley Additive exPlanations) proposed by Lundberg, Scott M & Lee, Su-In (2017) is one of the feature importance agnostic models that uses the game theory to explain the prediction of any black box model. It assigns Shapley values to each feature (K. Zhang, et al., 2020) by calculating the contribution of each feature to the prediction of the black box model. Rigorous computation and approximation methods (sampling SHAP and Kernel SHAP) are required to arrive at the SHAP values (K. Zhang, et al., 2020). This study focuses on using SHAP to explain the prediction of a black box model. To achieve this, the base value for prediction from past knowledge was used to test the contribution of each feature to confirm the effect of the sum of contributions on the base value. For a regression model, the base value is computed as the mean of the dependent variable (Mollaei, Nafiseh, et al., 2022).

$$f(x) = \text{base value} + \sum(SHAP\ values)$$

The base value captured the SHAP value for all features and was subsequently used to explain the model's prediction. The SHAP explainer algorithm from the SHAP package was used to explain the black box prediction and to generate the SHAP value for each feature.



Figures 8a and 8b bar and dot plot showing the global importance of the top twenty effective input features arranged in descending order of magnitude based on their contribution to the prediction.

The summary plots in Figures 8a and 8b provided insight into the fundamental relationship between the independent features and target variable, they present a global view of the impact of each feature on the model. The plots illustrate the ranking in descending order of the top twenty features with the most impact on the total risk exposure using SHAP values. In Figure 8b the dot plot represents the magnitude of each feature for a specific data point using color coding from red(high) to blue(low). The risk score for wrist and hand was represented by blue and red horizontal dots of different sizes along the x-axis (SHAP values), this implies that the feature has a complex impact on the model prediction, a negative impact for low values(blue) and positive impact for high values(red). The size of the dots is an indication of the value of the feature for that data point.

4.3.1.3 Asymmetric Shapley Values (ASV)

The Shapley framework proposed by Lundberg, Scott M & Lee, Su-In(2017) is highly efficient in explaining the black box model. It presented a common model-agnostic language for explaining the output of the black box model using a set of mathematical theories (Frye, Christopher, et al., 2020). The framework ascribes its explanation to the contribution of the independent features to the target feature using SHAP values with the underlying assumption that the independent features are not correlated. This assumption was debated as it was found that SHAPley values were limited as the theory overlooked all the causal relationships in the data (Frye, Christopher, et al., 2020), reliant on fictitiously computed data, and expensive to compute. Frye, Christopher, et al.(2020) proposed the Asymmetric Shapley Value (ASV), an improved explainer that recognizes the causal information of all features and provides an explanation founded on the actual input features values of the black box model.

The implementation of the ASV explainer model involved taking the Shapley values as input to the explainer and outputting the ASV values. The ASV value determines the contribution of each feature to the model's prediction. To arrive at the ASV value, the negative and positive contribution for each feature was calculated and then used to calculate the ASV values.

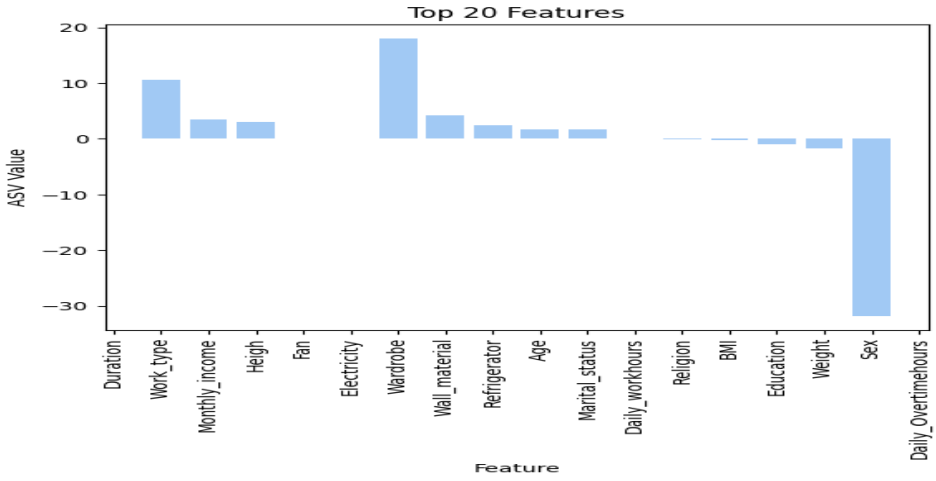


Figure 9. Asymmetric SHAPley values of the input features, showing top twenty features with the most significant contribution to the model's prediction.

Figure 9 illustrates the top twenty features based on their ASV values and impact on the target feature. It gives a simple and comprehensible representation of the relative importance of each feature in the model's prediction. Worthy of note is the feature "Sex" which has a high negative impact on the model's prediction.

4.3.1.4 Permutation Importance (PI)

Permutation Importance (PI) was proposed by Altmann, André, et al.(2010) to address the biases observed in tree-based models, such as the preference for categorial features with a high number of categories and feature importance ranking in Random Forest (RF). PI works to normalize the feature importance measure and subsequently regularise the bias (Hariharan, Swetha, et al., 2022). It achieves this by estimating the importance of a feature to a model's prediction by calculating the change in error following each iterative permutation or shuffling of the feature values. The P-value is then computed with the PI to determine the feature importance, making the model more explainable (Altmann, André, et al., 2010).

PI uses the model agnostic framework and can be applied to any model with ease and reliability without the requirement for retraining the model at each permutation of the dataset (Hariharan, Swetha, et al., 2022). The permutation importance function from the Scikit-learn library was used to implement the explainer. The approach evaluates the importance of each feature by randomly shuffling its values to confirm the effect on the model's performance. The function takes on the test data as input and performs 10 iterative permutations. Figure 10. shows the corrected feature importance using the importance score on the y-axis of the plot and features on the x-axis.

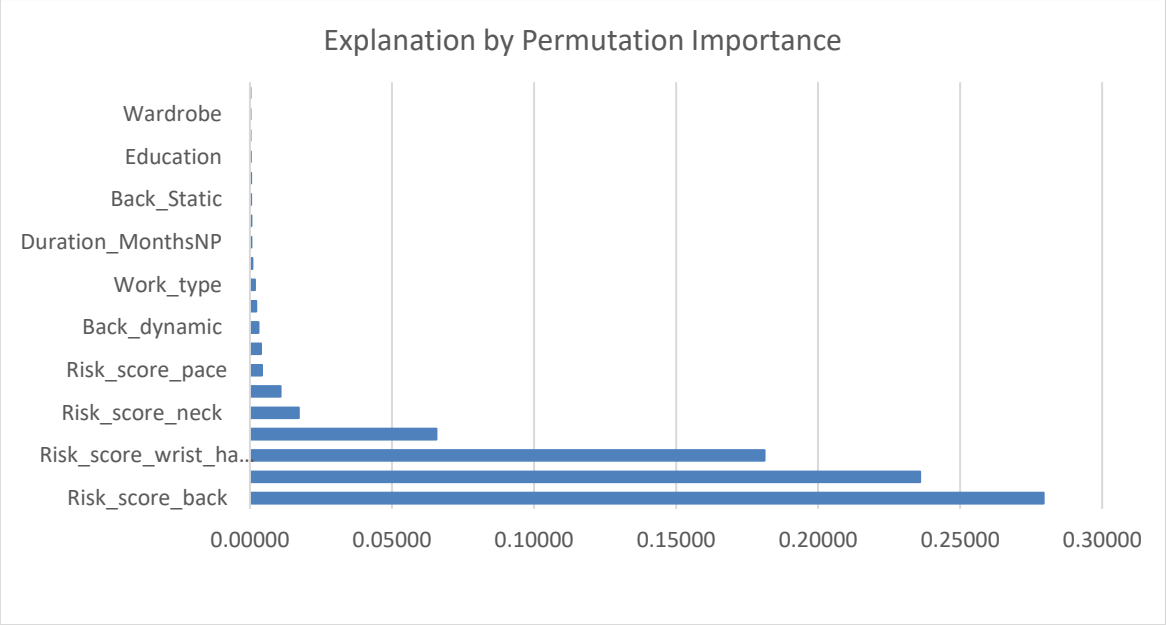


Figure 10. PI plot showing top twenty importance scores.

The importance score is a measure of the contribution of a feature to the overall performance of the model calculated using the PI method that evaluates the decline in the model's performance at the random permutation of a given feature value.

4.4 Discussion

In this study, the performance of different XAIs in interpreting the predictions of the Gradient Boosting Regressor model. The focus was on the use of a model-agnostic approach consisting of both local (LIME) and global (SHAP, ASV and PI) scope. Local explanation gives meaning to the model's decision-making process for a specific instance while the global explanation takes cognizance of the entire records.

For the global explanation, the SHAP and PI explainers were compared based on similar properties of SHAP's Global feature importance ranking and PI's importance score ranking for the top twenty impactful features, the results show a strong similarity in the top twenty important input features except for (sex, onset of neck pain, and refrigerator) for the SHAP explainer (onset of shoulder pain, duty change caused by lower back pain, and duration of neck pain) for the PI explainer. Comparing the explanation from both XAIs, the features from PI had a better impact on the model's prediction, this is observed by improvement in the model's MAE metric as against the baseline value. The features from PI gave a lower absolute error score compared to the

baseline of 0.009397/0.009433 and SHAP 0.009657/0.009433. This, however, confirms users' expectations as the important features highlighted by SHAP XAI lack clarity, for example, the feature, refrigerator, having a positive impact on the model's prediction. Both local and global explanations are also in agreement that the back, shoulder, neck, and hand/wrist have the most impact on the model's prediction and contribute to the total risk exposure of the workers.

Other qualitative measures (human-centred) were employed to evaluate the XAI explanation. SHAP though lacking in clarity and fidelity provided better user understandability following its simple and compact explanation, with easy-to-grasp visualizations and computations.

5 General Discussion and Conclusion

5.1 Introduction

This study set out to evaluate and adopt explainable artificial intelligence in occupational health and safety, with a particular focus on its potential for preventing work-related musculoskeletal symptoms and disorders. This was initiated by examining the application of state-of-art artificial intelligence in the field and reviewing the challenges facing the adoption in a sensitive domain like occupational health and safety. Subsequently, the effectiveness of relevant XAIs was evaluated. The overall aim of the evaluation is to mitigate the existing limitation currently facing the acceptance of artificial intelligent models and to highlight the benefits associated with incorporating explainable AI into the prevention of work-related musculoskeletal disorders.

5.2 Evaluation of main research approach and findings

Project implementation involved a thorough review of several pieces of literature and an examination of machine and deep learning algorithms to identify suitable algorithms for the project. Eight models were implemented, and their evaluation revealed certain models performed well but with the gradient-boosting regressor model outperforming the others. SHAP, LIME, ASV and PI were employed to explain the model's prediction and subsequently, the performance of XAIs was evaluated using human-centred and computer-centred approaches. SHAP and PI provided a better global explanation compared to ASV and LIME was employed to provide the local explanation.

Findings from this study revealed that explainable AIs can greatly enhance the prevention of WMSDs, these systems take advantage of the advanced machine learning algorithms which can not only analyse large datasets but identify patterns and risk factors associated with WMSDs. Explainable AI allows health and safety professionals and other users at all levels to understand and develop trust in the decision-making process of models and by this means improve the probability of successful intervention and prevention strategies, for example, XAI can build the trust of

management and occupational health personnel in the prediction of the machine learning model as they can predict the occurrence of WSMD to a high degree and also identify the contributing risk factors. Armed with this knowledge, management can adequately tailor programmes, restructure the work environment and put in place procedures that would eliminate the risk factors to a very large extent. Explainable AIs can also enable communication and collaboration between humans and intelligent systems thereby creating a symbiotic relationship that exploits the strengths of both parties.

Despite the promising benefits, the adoption of explainable AI in the workplace for the prevention of WMSDs may be faced with some challenges. First, there is a need for the effective integration of AI systems into occupational health and safety frameworks and standards. This will require AI experts and domain specialists to work together to ensure that the AI models align with the unique needs of the workplace. Second, being an innovation the adoption of XAI may require change management. Employees and occupational health officers have to be trained and carried along in its implementation. Finally, the adoption of XAI will also require extensive training for health and safety professionals to enable them to understand and utilize XAI systems effectively.

5.3 Ethical, legal, social, professional, and security issues

Explainable AI systems work with large amounts of data and in the field of occupational health and safety, the data collected for analysis often include sensitive personal information and this raises concerns about data privacy and security issues that must be addressed. It is therefore of great importance to put in place robust data governance protocols in line with already existing relevant data protection legislation and regulations to ensure strict compliance with these regulations.

Explainable AI promotes compliance with ethical and legal requirements through (1) the provision of transparency and understandability of the decision-making process that births the prediction of the model. (2) detecting and eliminating biases that may be perpetuated by models from the training data, this is achieved by the XAI analysing and explaining the model prediction and presenting the features employed in the decision process, making it easy to detect and eliminate biases (3)enabling audit and accountability through logs of decision steps and reasoning behind each decision, this

property can help meet legal and compliance requirements (4) supports ethical impact evaluation by providing insightful information on the impact of the AI model on users groups, this sort of evaluations helps to confirm if the model complies with ethical requirements such as fairness, privacy protection and non-discrimination.

5.4 Personal reflection

Some useful insights have been gained from the entire process of evaluating and adopting explainable AI techniques for the prevention of musculoskeletal disorders such as the need to preserve data quality through the entire process as this will ensure that the model predictions are devoid of bias and are applicable. For generalization of the result obtained data size must be adequately representative of the population both in size and composition.

Having few or too many features for model training may be problematic as the presence of input features with high feature-feature correlation can lead to multicollinearity which reduces the statistical power of a regression model by making it difficult for the model to evaluate the relationship between the independent and target features the, also, selecting only a few input variables with high correlation while omitting those with less correlation may lead to data bias, it is therefore important that suitable and thorough feature selection methods are adopted.

In the area of explaining model prediction, global explanation using SHAP, ASV and PI results in a simple representation of the machine learning model which may not be adequate to completely interpret the model on the other hand local explanation by LIME which employs surrogate methods may also not adequately interpret the original model, with this in mind, researchers need to thoroughly evaluate explanation methods to confirm that the method selected can provide a comprehensive and interpretation of the model.

It is also important to highlight the limitation regarding the non-availability of standardized quantifiable measures for evaluating explainable AIs, current studies have proposed different measures both qualitative and quantitative, however, there is a need for the AI community and professionals to review and evaluate existing

proposed metrics with a view to adopting a common set of measures for evaluating the performance explainable AI techniques.

5.5 Conclusion

In conclusion, this study has shown that explainable AI can greatly improve workplace safety by identifying and preventing risk factors that are linked to musculoskeletal symptoms and disorders. Despite its limitations, the study certainly adds to the existing body of knowledge and deepens our understanding of the application of explainable AI in the field of occupational health and safety. By highlighting the benefits and limitations associated with adopting explainable AI the study seeks to encourage further works and collaborations among stakeholders.

6 References

Alaca, Nuray, Safran, Elif Esma, Karamanlargil, Asli İrem & Timucin, Emel, 2019. Translation and cross-cultural adaptation of the extended version of the Nordic musculoskeletal questionnaire into Turkish. *Journal of musculoskeletal & neuronal interactions,* 19(4), p. 472.

Alin, A., 2010. Multicollinearity. *Wiley interdisciplinary reviews: computational statistics,* 2(3), pp. 370-374.

Altmann, André, Toloşi, Laura, Sander, Oliver & Lengauer, Thomas, 2010. Permutation importance: a corrected feature importance measure. *Bioinformatics,* 26(10), pp. 1340-1347.

Arrieta, Alejandro Barredo, et al., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion,* Volume 58, pp. 82-115.

Avi Rosenfeld & Ariella Richardson, 2019. Explainability in human-agent systems. *Auton. Agents Multi Agent Syst,* 33(6), pp. 673-705.

Avnimelech, R. & Intrator, N., 1999. Boosting regression estimators. *Neural computation,* 11(2), pp. 499-520.

Bibal, Adrien & Frénay, Benoît, 2016. *Interpretability of machine learning models and representations: an introduction.* s.l., s.n., pp. 77-81.

Botchkarev, A., 2018. Performance metrics (error measures) in machine learning regression, forecasting and prognostics: Properties and typology. *arXiv preprint arXiv:1809.03006.*

Breiman, Leo., 2001. Random forests. *Machine learning,* Volume 45, pp. 5-32.

Breiman, L., 1984. *Classification and Regression Trees.* 1st Edition ed. New York: Routledge.

Breiman, L., 1996. Bagging predictors. *Machine learning,* Volume 24, pp. 123-140.

Brownlee, J., 2016. *Machine learning algorithms from scratch with Python. Machine Learning Mastery..* s.l.:s.n.

Bühlmann, Peter & Hothorn, Torsten, 2007. Boosting algorithms: Regularization, prediction and model fitting..

Chai, T. & Draxler, Roland R, 2014. Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific model development discussions,* 7(1), pp. 1525-1534.

Chan, Victor CH, et al., 2022. The role of machine learning in the primary prevention of work-related musculoskeletal disorders: A scoping review. *Applied Ergonomics,* 98(0003-6870), p. 103574.

Creswell, John W & Poth, Cheryl N, 2016. *Qualitative inquiry and research design: Choosing among five approaches.* s.l.:Sage publications.

Darias, Jesus M, Díaz-Agudo, Belén & Recio-Garcia, Juan A, 2021. *A Systematic Review on Model-agnostic XAI Libraries.* s.l., s.n., pp. 28-39.

Das, Arun & Rad, Paul, 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371.*

Dawson, Anna P, Steele, Emily J, Hodges, Paul W & Stewart, Simon, 2009. Development and Test–Retest Reliability of an Extended Version of the Nordic Musculoskeletal Questionnaire (NMQ-E): A Screening Instrument for Musculoskeletal Pain. *The Journal of Pain,* 10(5), pp. 517-526.

De Barros, ENC & Alexandre, Neusa Maria C, 2003. Cross-cultural adaptation of the Nordic musculoskeletal questionnaire. *International nursing review,* 50(2), pp. 101-108.

Dorogush, A.V, Ershov, V & Gulin, A, 2018. CatBoost: gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363.*

Doshi-Velez, Finale & Kim, Been, 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608.*

E. Tjoa & C. Guan, 2021. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning System,* 32(11), pp. 4793-4813.

EUR-Lex, 2016. *General Data Protection Regulation (GDPR),* Luxembourg: EUR-Lex.

Fabian Pedregosa, et al., 2011. *Scikit-learn: Machine Learning in Python.* [Online] Available at: http://jmlr.org/papers/v12/pedregosa11a.html

Fan, J, et al., 2019. Light Gradient Boosting Machine: An efficient soft computing model for estimating daily reference evapotranspiration with local and external meteorological data. *Agricultural water management,* Volume 225, p. 105758.

Fisher, M.J. & Marshall, A.P, 2009. Understanding descriptive statistics. *Australian critical care,* 22(2), pp. 93-97.

Freitas, A. A., 2014. Comprehensible classification models: a position paper. *ACM SIGKDD explorations newsletter,* 15(1), pp. 1-10.

Friedman, J., 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics,* pp. 1189-1232.

Friedman, J., 2002. Stochastic gradient boosting. *Computational statistics & data analysis,* 38(4), pp. 367-378.

Frye, Christopher, et al., 2020. Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272.*

Frye, Christopher, Rowat, Colin & Feige, Ilya, 2020. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems,* Volume 33, pp. 1229-1239.

G. Zhou, V. Aggarwal, M. Yin & D. Yu, 2021. *Video-based AI Decision Support System for Lifting Risk Assessment.* s.l., s.n., pp. 275-282.

Gerd Gigerenzer & Henry Brighton, 2009. Homo heuristicus: Why biased minds make better inferences. *Topics in cognitive science,* 1(1), pp. 107-143.

Ghasemkhani, Mehdi, Mahmudi, Elham & Jabbari, Hossain, 2008. Musculoskeletal symptoms in workers. *International Journal of Occupational Safety and Ergonomics,* 14(4), pp. 455-462.

Gilpin, Leilani H, et al., 2018. *Explaining explanations: An overview of interpretability of machine learning.* s.l., IEEE, pp. 80-89.

Guidotti, Riccardo, et al., 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR),* 51(5), pp. 1-42.

Gunn, S.R, 1998. Support vector machines for classification and regression. *ISIS technical report,* 14(1), pp. 5-16.

Gunning, David, et al., 2019. XAI—Explainable artificial intelligence. *Science robotics,* 4(37), p. eaay7120.

Gunning, D., Vorm, E., Wang, Yunyan & Turek, Matt, 2021. DARPA's explainable AI (XAI) program: A retrospective. *Authorea Preprints.*

Gupta, G., 2013. Prevalence of musculoskeletal disorders in farmers of Kanpur-Rural. *India. J Community Med Health Educ,* 3(249), pp. 2161-0711.1000249.

Hancock, J.T. & Khoshgoftaar, T.M, 2020. CatBoost for big data: an interdisciplinary review. *Journal of big data,* 7(1), pp. 1-45.

Hariharan, Swetha, et al., 2022. XAI for intrusion detection system: comparing explanations based on global and local scope. *Journal of Computer Virology and Hacking Techniques.*

Hariri, R.H, Fredericks, E.M & Bowers, K.M, 2019. Uncertainty in big data analytics: survey, opportunities, and challenges. 6(1), pp. 1-16.

Hines, Brandon, Talbert, Douglas & Anton, Steven, 2022. Improving Trust via XAI and Pre-Processing for Machine Learning of Complex Biomedical Datasets. *The International FLAIRS Conference Proceedings,* 05 April.Volume 35.

Hoeting, J.A, Madigan, D, Raftery, A.E & Volinsky, C.T, 1999. Bayesian model averaging: a tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors. *Statistical science,* 14(4), pp. 382-417.

Hoffman, R.R, Mueller, S.T, Klein, G & Litman, J., 2018. *Metrics for explainable AI: Challenges and prospects.* s.l., arXiv preprint arXiv:1812.04608..

Holzinger, Andreas, et al., 2022. *Explainable AI methods-a brief overview.* s.l., Springer, pp. 13-38.

Hossain, Mohammad Didar, et al., 2018. Prevalence of work related musculoskeletal disorders (WMSDs) and ergonomic risk assessment among readymade garment workers of Bangladesh: A cross sectional study. *PLOS ONE,* 13(7), p. e0200122.

Huysmans, Johan, et al., 2011. An empirical evaluation of the comprehensibility of decision table, tree and rule based predictive models. *Decision Support Systems,* 15(1), pp. 141-154.

Indumathi, N & Ramalakshmi, R, 2021. *An evaluation of work posture and musculoskeletal disorder risk level identification for the fireworks industry worker's.* s.l., IEEE, pp. 1-5.

Jabeur, S.B, Gharib, C., , Mefteh-Wali, S & Arfi, W.B., , 2021. CatBoost model and artificial intelligence techniques for corporate failure prediction. *Technological Forecasting and Social Change,* Volume 166, p. 120658.

Jain, V., et al., 2022. Ambient intelligence-based multimodal human action recognition for autonomous systems. *ISA Trans.*

Jaiswal, J.K. & Samikannu, R, 2017. *Application of random forest algorithm on feature subset selection and classification and regression.* s.l., IEEE, pp. 65-68.

Johnston, Ron, Jones, Kelvyn & Manley, David, 2018. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. *Quality & quantity,* Volume 52, pp. 1957-1976.

K. Zhang, P. Xu & J. Zhang, 2020. *Explainable AI in Deep Reinforcement Learning Models: A SHAP Method Applied in Power System Emergency Control.* s.l., s.n., pp. 711-716.

Katuwal, R, Suganthan, P.N & Zhang, L, 2020. *Heterogeneous oblique random forest. Pattern Recognition.* s.l.:Elsevier.

Kavitha, S., Varuna, S & Ramya, R, 2016. *A comparative analysis on linear regression and support vector regression.* s.l., IEEE, pp. 1-5.

Ke, G, et al., 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems,* Volume 30.

Kenny, Eoin M, Ford, Courtney, Quinn, Molly & Keane, Mark T, 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence,* p. 103459.

Kong, E.B. & Dietterich, T.G, 1995. *Error-correcting output coding corrects bias and variance.* s.l., s.n., pp. 313-321.

Kothari, C., 2020. *Research methodology methods and Techniques.* s.l.:s.n.

Kuorinka, Ilkka, et al., 1987. Standardised Nordic questionnaires for the analysis of musculoskeletal symptoms. *Applied ergonomics,* 18(3), pp. 233-237.

Kwok, S.W & Carter, C, 1990. Multiple decision trees. *In Machine intelligence and pattern recognition,* Volume 9, pp. 327-335.

Lee, Ju-Yeun, Lee, Woojoo & Cho, Sung-il, 2023. Characteristics of Fatal Occupational Injuries in Migrant Workers in South Korea: A Machine Learning Study. *Available at SSRN 4314078.*

Linardatos, Pantelis, Papastefanopoulos, Vasilis & Kotsiantis, Sotiris, 2020. Explainable ai: A review of machine learning interpretability methods. *Entropy,* 23(1), p. 18.

Lopes, P, et al., 2022. XAI Systems Evaluation: A Review of Human and Computer-Centred Methods.. *Applied Sciences,* 12(19), p. 9423.

Lundberg, Scott M & Lee, Su-In, 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems,* Volume 30.

Miller, T., 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence,* 01 February, Volume 267, pp. 1-38.

Mohseni, S, Zarei, N. & Ragan, E.D, 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS),* 11(3-4), pp. 1-45.

Mollaei, Nafiseh, et al., 2022. Human-Centered Explainable Artificial Intelligence: Automotive Occupational Health Protection Profiles in Prevention Musculoskeletal Symptoms. *International Journal of Environmental Research and Public Health,* 11 August, 19(15), p. 9552.

Ng, A., 2004. *Feature selection, L 1 vs. L 2 regularization, and rotational invariance.* s.l., In Proceedings of the twenty-first international conference on Machine learning, p. 78.

Nur, Nurhayati Mohd, Dawal, Siti Zawiah & Dahari, Mahidzal, 2014. *The prevalence of work related musculoskeletal disorders among workers performing industrial repetitive tasks in the automotive manufacturing companies.* s.l., s.n., pp. 1-8.

Otchere, D.A., et al., 2022. Application of gradient boosting regression model for the evaluation of feature selection techniques in improving reservoir characterisation predictions. *Journal of Petroleum Science and Engineering,* p. 109244.

Piranveyseh, Peyman, et al., 2016. Association between psychosocial, organizational and personal factors and prevalence of musculoskeletal disorders in office workers. *International Journal of Occupational Safety and Ergonomics,* 22(2), pp. 267-273.

Pishgar, M, Issa, S. F., Pratap, P & Darabi, H., 2021. REDECA: A Novel Framework to Review Artificial Intelligence and Its Applications in Occupational Safety and Health. *Int J Environ Res Public Health,* 18(13).

Poole, Michael A & O'Farrell, Patrick N, 1971. The assumptions of the linear regression model. *Transactions of the Institute of British Geogrraphers.*

Preece, A., 2018. Asking ‘Why’in AI: Explainability of intelligent systems–perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management,* 25(2), pp. 63-72.

Probst, Philipp, Wright, Marvin N & Boulesteix, Anne-Laure, 2019. Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: data mining and knowledge discovery,* 9(3), p. e1301.

Prokhorenkova, L., et al., 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems,* Volume 31.

Pugh, Judith D, et al., 2015. Validity and reliability of an online extended version of the Nordic Musculoskeletal Questionnaire (NMQ-E2) to measure nurses' fitness. *Journal of clinical nursing,* 24(23-24), pp. 3550-3563.

Quinlan, J., 1986. Induction of decision trees. *Machine learning.,* Volume 1, pp. 81-106.

Quinlan, J., 2014. *C4. 5: programs for machine learning.* s.l.:Elsevier.

Ramdan, Iwan Muhamad, Duma, Krispinus & Setyowati, Dina Lusiana, 2019. Reliability and validity test of the Indonesian version of the Nordic musculoskeletal questionnaire (NMQ) to measure musculoskeletal disorders (MSD) in traditional women weavers. *Glob Med Health Commun,* 7(2), pp. 123-130.

Ray, S., 2019. *A quick review of machine learning algorithms.* In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), IEEE, pp. 35-39.

Rehman, Bushra, Aslam, Ayesha, Ali, Afsheen & Tariq, Anum, 2016. Ergonomic hazards to dental surgeons: A cross-sectional study. *Pakistan Oral & Dental Journal,* 36(1).

Ribeiro, Marco Tulio, Singh, Sameer & Guestrin, Carlos, 2016. *" Why should i trust you?" Explaining the predictions of any classifier".* s.l., s.n., pp. 1135-1144.

Ribeiro, Marco Tulio, Singh, Sameer & Guestrin, Carlos, 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386.*

Robb, M. J. & Mansfield, Neil J, 2007. Self-reported musculoskeletal problems amongst professional truck drivers. *Ergonomics,* 50(6), pp. 814-827.

Rosecrance, John C, et al., 2002. Test-retest reliability of a self-administered musculoskeletal symptoms and job factors questionnaire used in ergonomics research. *Applied occupational and environmental hygiene,* 17(9), pp. 613-621.

Rosenfeld, A., 2021. *Better metrics for evaluating explainable artificial intelligence.* s.l., s.n., pp. 45-50.

Rumelhart, David E, Durbin, Richard, Golden, Richard & Chauvin, Yves, 1995. Backpropagation: The basic theory. *Backpropagation: Theory, architectures and applications,* pp. 1-35.

Safarov, R.Z., et al., 2020. Solving of classification problem in spatial analysis applying the technology of gradient boosting catboost. *Folia Geographica,* 62(1), p. 112.

Sagi, O. & Rokach, L, 2018. Ensemble learning: A survey. Wiley Interdisciplinary Reviews. *Data Mining and Knowledge Discovery,* 8(4), p. e1249..

Schapire, R., 1990. The strength of weak learnability. *Machine learning,* Volume 5, pp. 197-227.

Schmidt, M., 2005. Least squares optimization with L1-norm regularization. *CS542B Project Report,* Volume 504, pp. 195-221.

Sewell, M., 2008. Ensemble learning. *RN,* 11(02), pp. 1-34.

Shane, A. & Cheryl, B. T., 2009. Inferential Statistics. *Air Medical Journal,* 28(4), pp. 168-171.

Smola, A.J & Schölkopf, B, 2004. A tutorial on support vector regression. *Statistics and computing,* Volume 14, pp. 199-222.

Sun, X, Liu, M & Sima, Z, 2020. A novel cryptocurrency price trend forecasting model based on LightGBM. *Finance Research Letters,* Volume 32, p. 101084.

Szeto, Grace PY, et al., 2009. Work-related musculoskeletal symptoms in surgeons. *Journal of occupational rehabilitation,* Volume 19, pp. 175-184.

Tang, Yue Ting & Romero-Ortuno, Roman, 2022. Using Explainable AI (XAI) for the Prediction of Falls in the Older Population. *Algorithms,* 15(10), p. 353.

van der Heijden, et al., 2006. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: A clinical example. 59(10), pp. 1102-1109.

Vilone, Giulia & Longo, Luca, 2020. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093.*

Vilone, Giulia & Longo, Luca, 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion,* Volume 76, pp. 89-106.

Werbos, Paul J, 1990. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE,* 78(10), pp. 1550-1560.

Willmott, Cort J & Matsuura, Kenji, 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research,* 30(1), pp. 79-82.

Wolpert, D.H, 1992. Stacked generalization. *Neural networks,* 5(2), pp. 241-259.