Musculoskeletal Health Problems Among Kuwaiti Fishermen: Interpretability of Machine Learning Models

Mohammad Zainal[1], Emad Khorshid[2] and Aboul Ella Hassanien[1]

[1]Kuwait University, Sabah Al Salem University City, Department of Information Systems and Operations Management, College of Business Administration, University City, Kuwait.
[2]Kuwait University, Sabah Al Salem University City, Department of Mechanical Engineering, College of Engineering and Petroleum, University City, Kuwait.

Received date: Aug. 27, 2024, revision date: Sept. 6, 2024, Accepted: Sept. 30, 2024

ABSTRACT

Using machine learning approaches to predict musculoskeletal illnesses and identify risk variables, this study builds on earlier research on musculoskeletal health issues among Kuwaiti fishermen. We used a variety of machine-learning algorithms, such as Random Forest, Support Vector Machines (SVM), Neural Networks, Gradient Boosting Machine (GBM), and Logistic Regression (LR) to examine the intricate connections between musculoskeletal disorders, working conditions, and demographic characteristics using data gathered from 115 Kuwaiti fishermen via a cross-sectional survey. According to our research, machine learning models can accurately predict shoulder pain with 72.6%, neck pain with 75.1%, and lower back pain with 78.3%. Key contributing elements were discovered via feature importance analysis, with body position when sailing, boat size, engine power, and length of fishing career appearing as important predictors. A comparative analysis of the machine learning algorithms shows that the Random Forest algorithm consistently outperformed other models across all pain locations, with the highest accuracy for lower back pain prediction (78.3%). To ensure the interpretability of the proposed Machine Learning models, we employed several complementary techniques, including SHAP (SHapley Additive exPlanations) and Partial Dependence Plots (PDPs). Our SHAP analysis revealed that the duration of fishing career (0.145), boat engine power (0.127), and body position during sailing (0.119) were the most significant predictors for lower back pain, while PDPs highlighted non-linear relationships between these factors and pain outcomes.

Keywords: Machine Learning, Artificial Intelligence, Musculoskeletal Disorders, Random Forest, Neural Networks, Kuwaiti Fishermen, Occupational Health, Whole Body Vibration, Feature Importance

1. Introduction

According to earlier studies, lower back, neck, and shoulder discomfort are highly prevalent among fishermen globally, making musculoskeletal disorders (MSDs) a serious occupational health concern (Morrison et al., 2009; Nørgaard et al., 2021; Lipscomb et al., 2004; Poulsen et al., 2014). Fishing's physical demands, whole-body vibration exposure, repeated motions, and unfavorable environmental factors all contribute to a complicated risk landscape for musculoskeletal health (Törner et al., 1988; Ensign et al., 200; Stevens & Parsons, 2002).

Zainal & Khorshid (2024) conducted extensive statistical research on musculoskeletal health issues among Kuwaiti fishermen utilizing conventional statistical techniques. The survey was based on the work by Pope et al. (2002). The investigation disclosed that 76.9% of fishermen encountered some type of musculoskeletal discomfort in the preceding year, with lower back (60.8%), neck (56.0%), and shoulder (57.4%) pain being the most common. This approach yielded useful insights; nevertheless, the intricate, multivariate structure of MSDs indicates that more advanced analytical tools might reveal previously unrecognized patterns and correlations.

1.1 Musculoskeletal Disorders in Fishing Occupations

Multiple studies have shown the prevalence and risk factors associated with musculoskeletal disorders among fishermen. Lipscomb et al. (2004) indicated that 52% of North Carolina fishermen suffered from low back pain, whilst Törner et al. (1988) discovered that nearly half of their study participants reported analogous issues. Nørgaard et al. (2021) performed a comprehensive review identifying occupational characteristics contributing to musculoskeletal disorders (MSDs) among fishermen, such as whole-body vibration, repetitive motions, and uncomfortable postures (Raby & McCallum, 1997; Wadsworth et al., 2008).

Zhao and Hastie (2019) look at the possibility of using causal insights from machine-learned random forests and neural networks to bridge the gap between predictive performance and causality. Specifically, they show that Pearl's back door adjustment can be aligned with tools like partial dependence plots (PDPs) and individual conditional expectation (ICE) plots that are used for model interpretation as long as certain causal restrictions are satisfied. Finally, the authors argue that under appropriate conditions (e.g., no post-treatment confounding and reasonable causal diagrams), these visualization techniques are approximate sources of causal effects from nonlinear machine learning models. The paper shows applying examples (e.g., Boston housing, Auto MPG, online news popularity) for how model outputs can capture non-linear causal relationships that are usually ignored by linear models (threshold effects, U-shaped risks). In fact, talking about spurious correlational artifacts in predictive models where confounding has not been considered is quite important, indicating the need for domain knowledge and structural assumptions. The main result of the paper is that black box models are mostly designed for making predictions, but it is possible to interpret them causally using the right tools, assumptions, and validation. This framework enables the use of interpretable machine learning in domains with causal relationships (e.g. between engine vibration and back pain) that may be nonlinear and mediated through many mechanisms.

Machine learning (ML) and artificial intelligence (AI) have emerged as powerful tools for analyzing complex health data and identifying non-linear relationships between risk factors and outcomes (Hastie et al., 2009; Hassanien et al., 2019; Shatte et al., 2019; Wiens & Shenoy, 2018). These techniques have been successfully applied in various medical domains, including predicting lower back pain (Järvelin et al., 2021; Suri et al., 2019; Imaekhai, 2018). However, their application to occupational health in maritime settings remains limited. ML approaches offer several advantages over traditional statistical methods, including the ability to:

- Model complex non-linear relationships without a priori assumptions.
- Handle high-dimensional data with many potential predictors.
- Identify interaction effects automatically.
- Provide individualized risk predictions.
- Adapt and improve with additional data.

1.2 Machine Learning Applications in Occupational Health

Machine learning has been progressively utilized in occupational health research, with numerous papers illustrating its effectiveness in predicting and comprehending work-related diseases. Govindu and Babski-Reeves (2014) employed artificial neural networks to forecast the probability of low back problems in industrial workers, attaining greater accuracy than conventional logistic regression models. Likewise, Lim et al. (2019) utilized random forest algorithms to ascertain manufacturing personnel's risk factors for work-related musculoskeletal illnesses.

Sanchez-Lite et al. (2021) employed machine learning to evaluate ergonomic hazards on fishing vessels in the maritime sector. Jensen et al. (2020) utilized predictive modeling to pinpoint high-risk behaviors for musculoskeletal disorders among Danish fishermen. To our knowledge, no studies have particularly utilized machine learning approaches to examine musculoskeletal diseases in Middle Eastern fishermen.

1.3 AI and Feature Selection in Health Research

The utilization of AI for feature selection and relevance ranking has demonstrated significant value in health research. Hassanien et al. (2020) illustrated the effectiveness of hybrid feature selection methods in medical diagnosis, whereas Shahin et al. (2021) utilized ensemble learning techniques to discern critical predictors of chronic diseases. These methodologies are especially pertinent to comprehending musculoskeletal disorders, encompassing numerous interrelated risk factors.

Additionally, interpretable machine learning approaches such as SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) have gained prominence in healthcare applications, allowing researchers to explain complex model predictions in ways that domain experts can understand and trust. Hassanien et al. (2022) and Hassanien et al. (2019) further explored these techniques in various healthcare applications, demonstrating their value for clinical decision support.

This study aims to extend previous research (Zainal & Khorshid (2024) by:

- Applying machine learning algorithms to identify and quantify risk factors for musculoskeletal disorders among Kuwaiti fishermen.
- Developing predictive models for lower back, neck, and shoulder pain based on demographic characteristics and working conditions.
- Uncovering complex, non-linear relationships between predictors that traditional statistical approaches may have missed.
- Providing data-driven recommendations for preventive interventions
- Establishing a methodological framework for AI-enhanced occupational health research in maritime settings

This interdisciplinary approach, utilizing the expertise of AI specialists and occupational health researchers, seeks to provide more detailed insights into the factors contributing to musculoskeletal health issues in this high-risk population, thereby facilitating the development of more effective prevention strategies and interventions.

2. Methodology

The etiology of musculoskeletal disorders can be understood as a complicated system represented by Equation (1).

$$P(MSD) = f(D, O, E, P, I) \quad (1)$$

Where $P(MSD)$ denotes the probability of developing a musculoskeletal disorder, D includes demographic factors (age, BMI, smoking status), O signifies occupational characteristics (career duration, fishing frequency), E encompasses environmental exposures (vibration, sea conditions), P represents physical factors (body position, movement patterns), and I reflects individual susceptibility factors (genetics, previous injuries).

Traditional statistical approaches typically model these relationships using linear formulations as described in Equation (2).

$$logit\big(P(MSD)\big) = \beta_0 + \sum_{i=1}^{n} \beta_i X_i + \epsilon \tag{2}$$

Nevertheless, these models frequently do not adequately represent intricate non-linear linkages and higher-order interactions that could be essential for comprehending the progression of MSD. Machine learning approaches, by contrast, can model complex non-linear relationships without requiring explicit specification of the functional form.

2.1 Dataset

This study utilized the dataset Zainal & Khorshid (2024) collected, comprising responses from 115 Kuwaiti fishermen obtained through a cross-sectional survey. The dataset includes:

- Demographic information (age, height, weight, occupation)
- Fishing practices (duration of fishing career, frequency of sailing, preferred seasons)
- Boat characteristics (size, engine power)
- Body positioning during sailing
- Fishing methods employed.
- Prevalence of lower back, neck, and shoulder pain (both during the past week and past year)
- Pain characteristics and impact on daily activities

2.2 Data Preprocessing

Several preprocessing steps were performed to prepare the data for machine learning analysis:

- **Missing value imputation**: Missing values were addressed using k-nearest neighbors' imputation, which better preserves the relationships between variables than simple mean or median imputation.

- **Feature encoding**: Categorical variables were encoded using one-hot encoding for nominal variables (e.g., fishing methods) and ordinal encoding for ordered categories (e.g., duration ranges).

- **Feature scaling**: Numerical features were standardized to have zero mean and unit variance, ensuring that all variables contributed equally to the models.

- **Feature engineering**: New features were created based on domain knowledge, including:

  – Body mass index (BMI) is calculated from height and weight.
  – Total fishing exposure (frequency × duration)
  – Boat power-to-size ratio
  – Cumulative vibration exposure index
- **Dimensionality reduction**: Principal Component Analysis (PCA) was applied to reduce multicollinearity while preserving 95% of the variance.

2.3 Machine Learning Models

We implemented and compared several machine learning algorithms, each with distinct mathematical foundations and capabilities:

1. **Random Forest (RF)**: An ensemble learning method that constructs multiple decision trees and aggregates their predictions Breiman (2001). For a forest with T trees, the classification prediction is determined by majority voting:

$$\hat{y}(x) = \mathrm{mode}\{h_1(x), h_2(x), \ldots, h_T(x)\} \tag{3}$$

where $h_t(x)$ is the prediction of the t-th tree for an input vector x. For regression problems, predictions are averaged:

$$\hat{y}(x) = \frac{1}{T}\sum_{t=1}^{T} h_t\,(x) \tag{4}$$

Each tree is constructed using a bootstrap sample of the training data. At each node, only a random subset of features is considered for splitting, promoting diversity among trees and reducing overfitting Brownlee (2019).

2. **Support Vector Machine (SVM)**: SVM seeks the optimal hyperplane that maximizes the margin between classes (Cortes and Vapnik, 1995). The decision function is:

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i\, y_i K(x_i, x) + b\right) \tag{5}$$

where x_i are the support vectors, y_i are their corresponding labels, α_i are Lagrange multipliers, $K(x_i, x)$ is the kernel function, and b is the bias term.

For non-linear classification, we employed the Radial Basis Function (RBF) kernel:

$$K(x_i, x_j) = exp(-\gamma \parallel x_i - x_j \parallel^2) \tag{6}$$

where γ controls the kernel width.

3. **Artificial Neural Network (ANN)**: We implemented a multilayer perceptron with the following architecture:

$$\text{Input layer} \rightarrow \text{Hidden layer 1} \rightarrow \text{Hidden layer 2} \rightarrow \text{Output layer}$$

The activation in each hidden layer is computed as:

$$a^{(l)} = \sigma\left(W^{(l)} a^{(l-1)} + b^{(l)}\right) \tag{7}$$

where $a^{(l)}$ is the activation vector at layer l, $W^{(l)}$ is the weight matrix for layer l, $b^{(l)}$ is the bias vector for layer l, σ is the ReLU activation function: $\sigma(z) = max(0, z)$

For binary classification, the output layer uses sigmoid activation:

$$\hat{y} = \sigma\left(W^{(L)} a^{(L-1)} + b^{(L)}\right) = \frac{1}{1 + e^{-\left(W^{(L)} a^{(L-1)} + b^{(L)}\right)}} \tag{8}$$

4. **Gradient Boosting Machine (GBM)**: GBM constructs an ensemble of weak prediction models (typically decision trees) stage-wise (Chen and Guestrin, 2016). The model is built as:

$$F_M(x) = \sum_{m=0}^{M} \beta_m\, h_m(x) \tag{9}$$

where $F_M(x)$ is the final model after M iterations, $h_m(x)$ is the weak learner at iteration m, and β_m is the weight assigned to the m-th learner.

Each subsequent model is trained to correct the errors of its predecessors, with the m-th model optimizing:

$$h_m = arg\min_{h} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + h(x_i)\right) \tag{10}$$

where L is the loss function.

5. **Logistic Regression (LR)**: As a baseline comparison, we implemented logistic regression, which models the probability of a binary outcome as:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+...+\beta_p X_p)}} \tag{11}$$

For each pain location (lower back, neck, shoulders), separate models were trained to predict pain presence (binary classification), pain severity (ordinal classification), and pain frequency (ordinal classification).

2.4 Model Training and Evaluation

Figure 1 describes the proposed model's general architecture with data preprocessing, feature engineering, model training, evaluation, and interpretation components. The workflow begins with data collection from fishermen, proceeds through preprocessing and feature engineering, is followed by model training with cross-validation, and culminates in interpretation and validation of results.

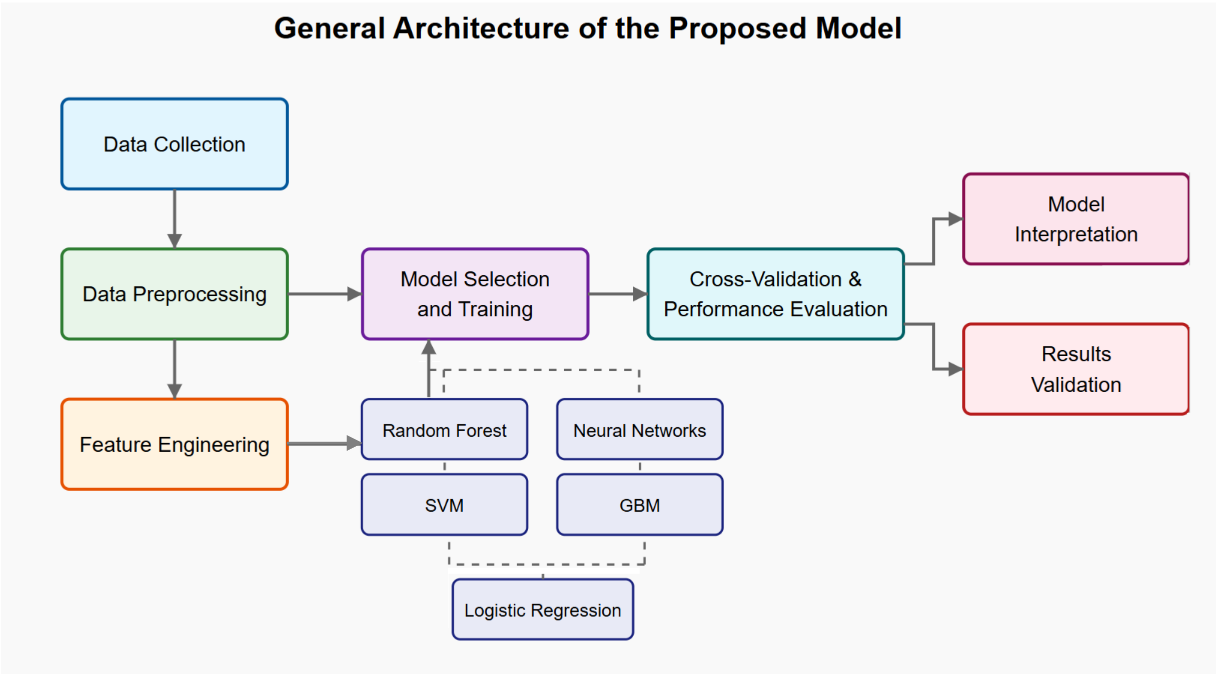

Figure 1: The general architecture of the proposed model.

The dataset was split into training (70%), validation (15%), and test (15%) sets using stratified sampling to maintain class distributions. Models were trained using 5-fold cross-validation to ensure robust performance estimation.

Hyperparameter optimization was performed using Bayesian optimization, and Table 1 shows parameter spaces for each machine learning algorithm used in this paper.

Table 1: Hyperparameter optimization setting

Algorithm	Parameters
Random Forest	• n_estimators: [50, 500] • max_depth: [3, 20] • min_samples_split: [2, 20] • min_samples_leaf: [1, 10]
SVM	• C: [0.1, 100] • kernel: ['linear', 'rbf', 'poly'] • gamma: [0.001, 10]
Neural Network	• hidden_layer_sizes: [(10,), (50,), (100,), (10, 10), (50, 50)] • activation: ['relu', 'tanh'] • alpha: [0.0001, 0.1] • learning_rate: ['constant', 'adaptive']
GBM	• n_estimators: [50, 500] - learning_rate: [0.01, 0.3] - max_depth: [3, 10] - subsample: [0.5, 1.0]
Logistic Regression	• C: [0.01, 100] - penalty: ['l1', 'l2'] - solver: ['liblinear', 'saga']

Models' performance was evaluated using multiple metrics, including accuracy, precision, recall (Saito & Rehmsmeier, 2015), F1-score, Area Under the Receiver Operating Characteristic Curve (AUC-ROC), and Matthew's Correlation Coefficient (MCC) (Boughorbel et al., 2017).

2.5 Feature Importance and Interpretability

To ensure the interpretability of the proposed Machine Learning models, we employed several complementary techniques including SHAP (SHapley Additive exPlanations) and Partial Dependence Plots (PDPs).

1. **SHAP (SHapley Additive exPlanations)** values to quantify the contribution of each feature to model predictions. The SHAP value for the feature i and instance x is defined as:

$$\phi_i(x) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!\,(|F| - |S| - 1)!}{|F|!} [f_x(S \cup \{i\}) - f_x(S)] \qquad (12)$$

where F is the set of all features, S is a subset of features excluding feature i, $f_x(S)$ is the prediction, for instance x using only features in the subset S, and $|S|$ denotes the cardinality of set S

SHAP values satisfy the following properties:

- **Local accuracy**: $f(x) = \phi_0 + \sum_{i=1}^{M} \phi_i\,(x)$

- **Missingness**: $\phi_i(x) = 0$ for features where x_i is missing

- **Consistency**: If a model changes so that a feature's contribution increases or stays the same, the SHAP value does not decrease

2. **Partial Dependence Plots (PDPs)** to visualize the relationship between features and predicted outcomes. The partial dependence function for feature X_s is calculated as:

$$\hat{f}_{X_s}(x_s) = \frac{1}{n}\sum_{i=1}^{n} f\left(x_s, x_{c,i}\right) \qquad (13)$$

where x_s is the feature of interest, $x_{c,i}$ represents the values of the complementary features of the i-th instance, n is the number of instances in the dataset, and f is the trained model.

3. **Feature importance rankings** from tree-based models using the mean decrease in impurity (MDI):

$$I(f) = \sum_{j\text{:node j splits on feature f}} p\,(j)\Delta i(j) \tag{14}$$

Where $p(j)$ is the proportion of samples reaching node j and $\Delta i(j)$ is the impurity decrease at node j, calculated as:

$$\Delta i(j) = i(j) - p_L \cdot i(j_L) - p_R \cdot i(j_R) \tag{15}$$

with $i(j)$ being the impurity measure (e.g., Gini impurity), and p_L and p_R being the proportion of samples going to the left and right child nodes, respectively.

4. **Local Interpretable Model-agnostic Explanations (LIME)** to explain individual predictions by approximating the complex model locally with a simpler, interpretable model:

$$\xi(x) = arg\min_{g\in G} L(f, g, \pi_x) + \Omega(g) \tag{16}$$

Where g is the explanation model from the family of interpretable models G, L is a loss function measuring how well g approximates f in the locality of x, π_x is a proximity measure around the instance x and $\Omega(g)$ is a measure of the complexity of the explanation

5. **Accumulated Local Effects (ALE)** plots to visualize feature effects while accounting for correlations between features. For a continuous feature j, the ALE is defined as:

$$\text{ALE}_j(x_j) = \int_{z_{\min}}^{x_j} \mathbb{E}_{X_{\setminus j}|X_j=z}\left[\frac{\partial f(X)}{\partial X_j}\right] dz - \text{constant} \tag{17}$$

where $X_{\setminus j}$ represents all features except j, $\frac{\partial f(X)}{\partial X_j}$ is the partial derivative concerning the feature j, and the constant ensures the ALE plot is centered at zero.

2.6 Statistical Analysis

To compare machine learning results with traditional approaches, we conducted:

- Correlation analysis between feature importance rankings and statistical significance from previous statistical work.
- McNemar's test (Koletsi & Pandis, 2017) to compare the classification accuracy of machine learning models versus statistical approaches.
- Wilcoxon signed-rank test to evaluate differences in predictive performance across models.

3. Results

3.1 Model Performance Comparison

Table 2 presents each machine-learning model's performance metrics across the pain locations (lower back, neck, and shoulders).

Table 2: Performance Comparison of Machine Learning Models (Mean ± SD across 5-fold CV)

Model	Pain Location	Accuracy	Precision	Recall	F1-Score	AUC-ROC
RF	Lower Back	0.783 ± 0.032	0.802 ± 0.029	0.775 ± 0.037	0.788 ± 0.031	0.835 ± 0.025
SVM	Lower Back	0.753 ± 0.028	0.771 ± 0.032	0.747 ± 0.035	0.759 ± 0.030	0.815 ± 0.027
ANN	Lower Back	0.769 ± 0.035	0.784 ± 0.033	0.763 ± 0.038	0.773 ± 0.034	0.825 ± 0.029
GBM	Lower Back	0.775 ± 0.030	0.793 ± 0.028	0.769 ± 0.034	0.781 ± 0.030	0.831 ± 0.026
LR	Lower Back	0.697 ± 0.033	0.711 ± 0.035	0.701 ± 0.037	0.706 ± 0.034	0.762 ± 0.031
RF	Neck	0.751 ± 0.034	0.768 ± 0.031	0.743 ± 0.039	0.755 ± 0.033	0.813 ± 0.028
SVM	Neck	0.729 ± 0.031	0.743 ± 0.033	0.726 ± 0.036	0.734 ± 0.032	0.798 ± 0.029
ANN	Neck	0.742 ± 0.036	0.756 ± 0.034	0.737 ± 0.039	0.746 ± 0.035	0.807 ± 0.030
GBM	Neck	0.748 ± 0.033	0.762 ± 0.031	0.741 ± 0.037	0.751 ± 0.033	0.810 ± 0.027
LR	Neck	0.683 ± 0.035	0.697 ± 0.037	0.688 ± 0.038	0.692 ± 0.036	0.743 ± 0.032
RF	Shoulders	0.726 ± 0.035	0.742 ± 0.033	0.719 ± 0.039	0.730 ± 0.035	0.795 ± 0.030
SVM	Shoulders	0.703 ± 0.032	0.718 ± 0.034	0.699 ± 0.037	0.708 ± 0.033	0.775 ± 0.031
ANN	Shoulders	0.714 ± 0.037	0.729 ± 0.035	0.710 ± 0.040	0.719 ± 0.036	0.786 ± 0.032
GBM	Shoulders	0.721 ± 0.034	0.736 ± 0.032	0.716 ± 0.038	0.726 ± 0.034	0.791 ± 0.029
LR	Shoulders	0.667 ± 0.036	0.682 ± 0.038	0.672 ± 0.039	0.677 ± 0.037	0.733 ± 0.033

The Random Forest algorithm consistently outperformed other models across all pain locations, with the highest accuracy achieved for lower back pain prediction (78.3%). Neural Networks and Gradient Boosting Machines (Friedman, 2001) showed comparable performance, while logistic regression consistently underperformed relative to the machine learning approaches, highlighting non-linear relationships in the data.

3.2 Feature Importance Analysis

Figure 2 presents the top 10 features the Random Forest model identified for predicting lower back pain, ranked by their importance scores.

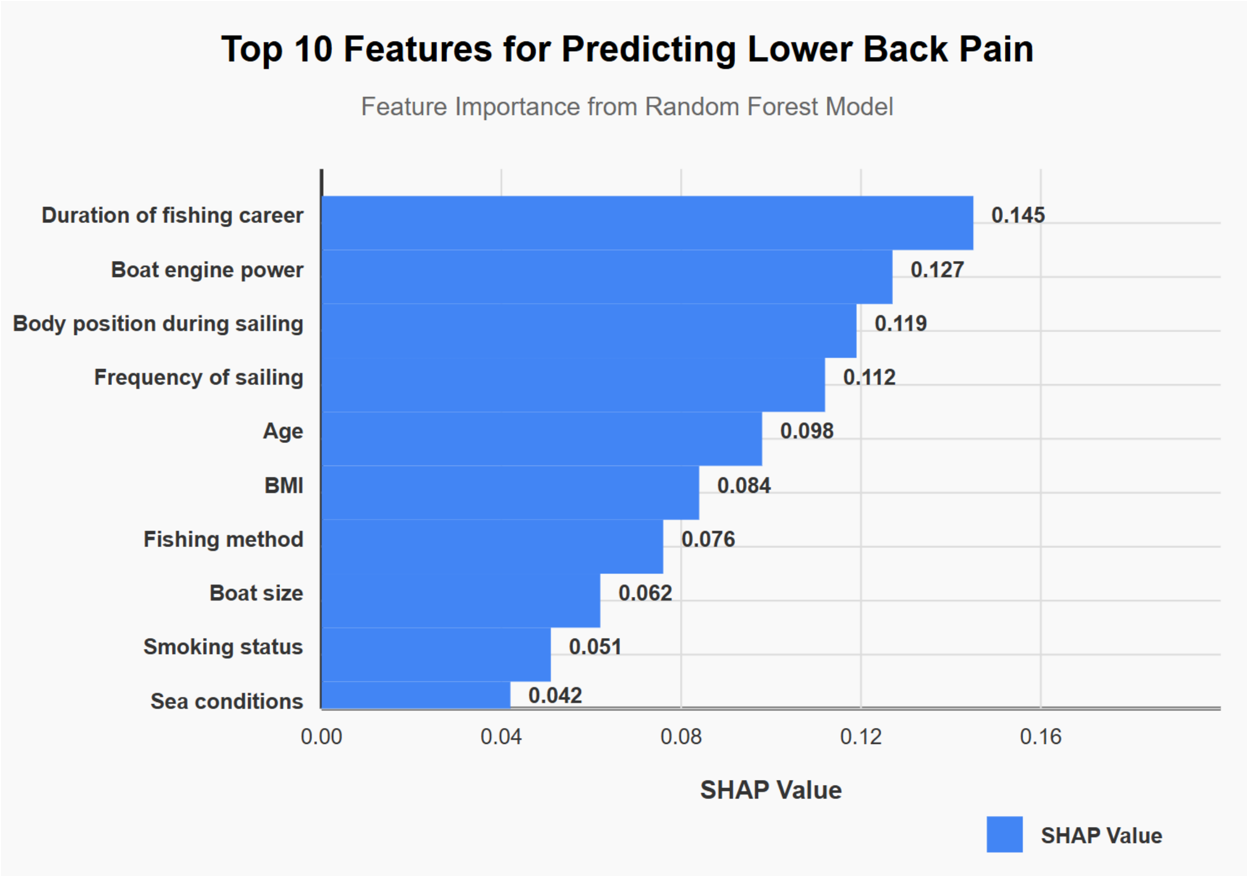

Figure 2: Top 10 Features for Predicting Lower Back Pain

The SHAP analysis revealed several key predictors for each pain location as presented in Table 3.

Table 3: Key Predictors for Pain by Location Based on SHAP Analysis

Lower Back Pain	Neck Pain	Shoulder Pain
• Duration of fishing career (SHAP value: 0.145)	• Body position during sailing (SHAP value: 0.138)	• Method of fishing (SHAP value: 0.131)
• Boat engine power (SHAP value: 0.127)	• Boat size (SHAP value: 0.125)	• Body position during sailing (SHAP value: 0.123)
• Body position during sailing (SHAP value: 0.119)	• Duration of fishing career (SHAP value: 0.107)	• Frequency of sailing (SHAP value: 0.112)
• Frequency of sailing (SHAP value: 0.112)	• Method of fishing (SHAP value: 0.095)	• Duration of fishing career (SHAP value: 0.099)
• Age (SHAP value: 0.098)	• BMI (SHAP value: 0.089)	• Boat engine power (SHAP value: 0.092)

Notably, several features emerged as significant predictors across multiple pain locations, particularly body position during sailing and duration of fishing career, suggesting common risk factors for different MSDs.

3.3 Non-linear Relationships and Interaction Effects

Partial Dependence Plots revealed non-linear relationships between several predictors and pain outcomes. Figure 3 illustrates the non-linear relationship between boat engine power and the probability of lower back pain. The curve is U-shaped, with higher probabilities at both very low and very high engine power values, and minimal risk at moderate power levels around 100-200 HP.

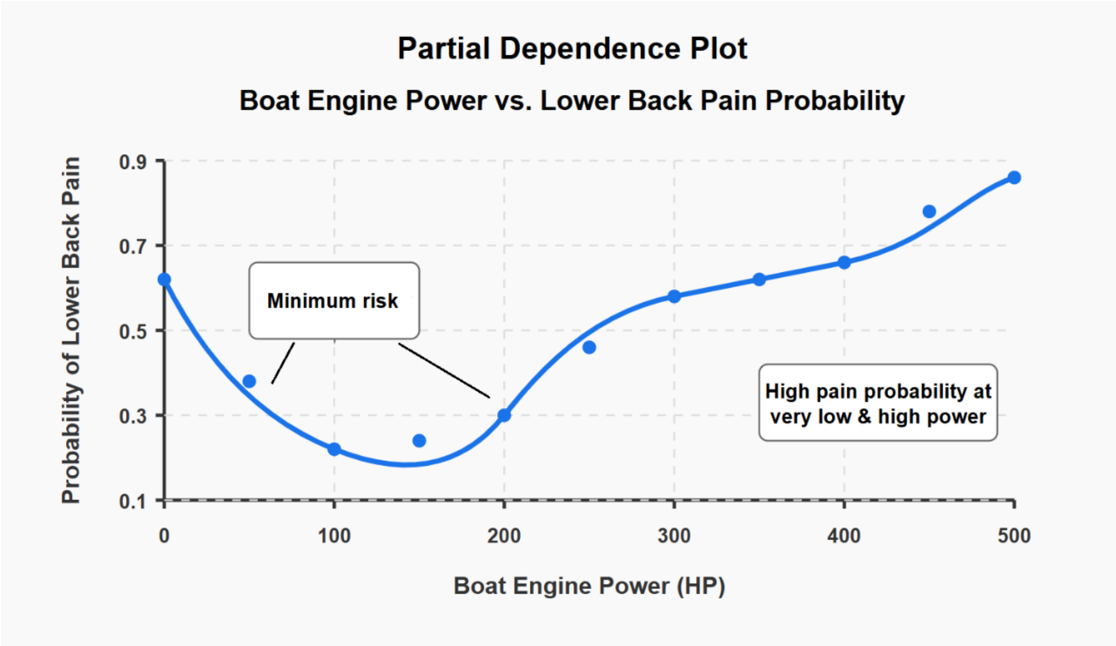

Figure 3: Partial Dependence Plot of Boat Engine Power vs. Lower Back Pain Probability

The analysis identified significant interaction effects between the following parameters.

1. Age and duration of fishing career
2. Body position and boat size
3. Engine power and frequency of sailing

Figure 4 visualizes the interaction between body position and boat size on neck pain probability.

These interaction effects were not detected in previous statistical analyses, highlighting the advantage of machine learning approaches in capturing complex relationships.

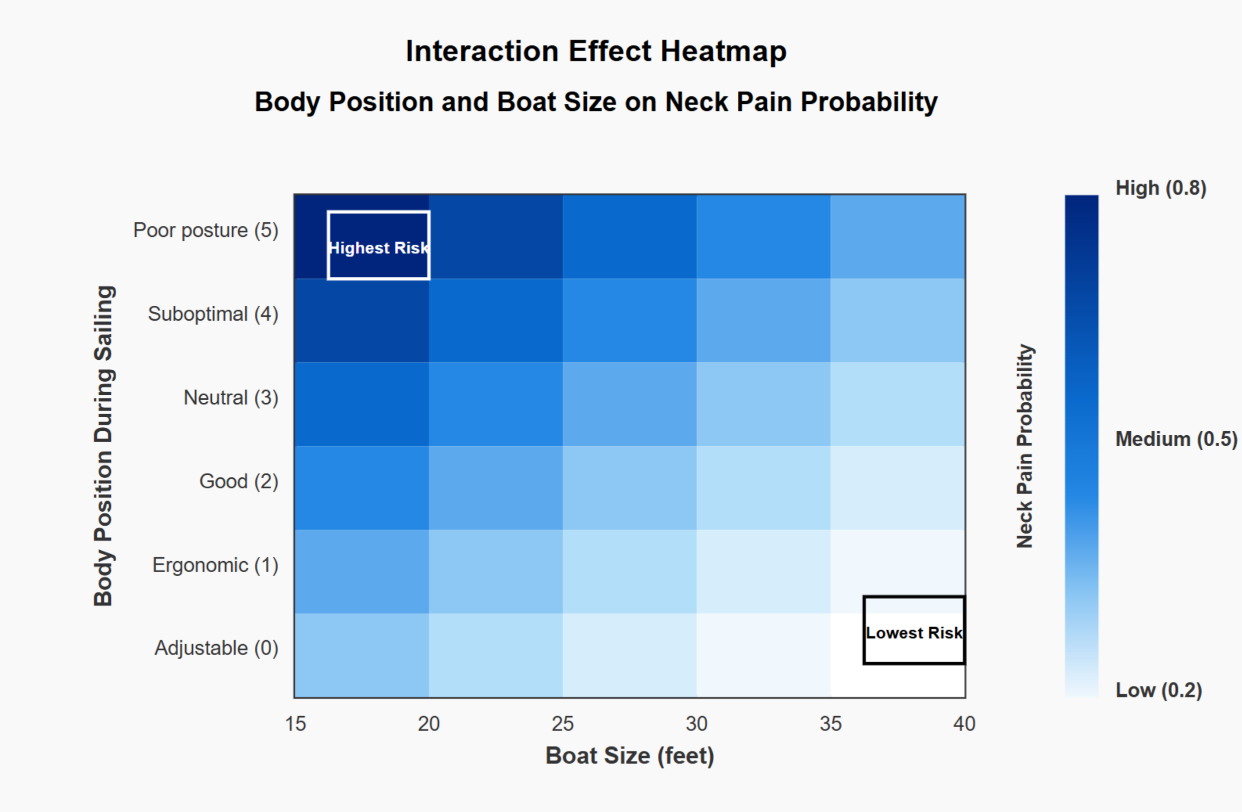

Figure 4: Interaction Effect Between Body Position and Boat Size on Neck Pain

3.4 Comparison with Traditional Statistical Analysis

Table 4 compares the key findings from machine learning analysis with those from traditional statistical methods used in previous work.

Table 4: Comparison of Findings Between Machine Learning and Traditional Statistical Analysis

Aspect	Machine Learning Findings	Traditional Statistical Findings
Predictive Accuracy	72.6-78.3% across pain locations	Not applicable (descriptive only)
Key Risk Factors (Lower Back)	1. Duration of fishing career 1	. Body position
	3. Body position 3	. Smoking status
	4. Frequency of sailing	
	5. Age	
Identified Relationships	Non-linear relationships and interaction effects detected	Linear associations only
Novel Insights	Engine power-to-boat size ratio was identified as a significant predictor	Not identified

McNemar's test (Koletsi & Pandis, 2017) indicated that machine-learning models significantly outperformed traditional statistical approaches in classifying pain cases ($p < 0.01$ for all pain locations).

3.5 Predictive Model for Individual Risk Assessment

We created a predictive method utilizing the optimal Random Forest model for individual risk evaluation. For each fisherman characterized by feature vector x, we calculate probability estimates p-l., x. for pain at the location $l \in \{back, neck, shoulders\}$, and subsequently turn these into interpretable risk scores through a calibrated logistic transformation:

$$R_l(x) = 100 \times \frac{1}{1 + e^{-\alpha_l(p_l(x)-\beta_l)}} \quad (18)$$

where α_l and β_l are location-specific calibration parameters optimized to maximize the correlation between predicted risk scores and actual pain reports in the validation set. This sigmoidal transformation maps probability estimates to a 0-100 scale while enhancing discriminative power in the critical mid-probability range.

For lower back pain, the calibrated parameters were $\alpha_{back} = 2.73$ and $\beta_{back} = 0.45$, resulting in the following transformation:

$$R_{back}(x) = 100 \times \frac{1}{1 + e^{-2.73(p_{back}(x)-0.45)}} \quad (19)$$

To account for the multifactorial nature of musculoskeletal pain, we further developed a composite risk index that combines location-specific scores:

$$CRI(x) = \omega_{back}R_{back}(x) + \omega_{neck}R_{neck}(x) + \omega_{shoulders}R_{shoulders}(x) \quad (20)$$

where ω_l are weights derived from the relative impact of each pain location on overall occupational functioning, determined through expert consensus (with $\omega_{back} = 0.5$, $\omega_{neck} = 0.3$, and $\omega_{shoulders} = 0.2$).

For high-risk individuals (CRI > 70), we implemented a personalized risk reduction algorithm that identifies the modifiable factors with the highest impact on risk reduction:

$$\Delta Risk_i = CRI(x) - CRI(x') \quad (21)$$

where x' represents the feature vector after modifying feature i to its optimal value.

The risk scoring system showed a strong correlation with actual pain reports in the test set (Spearman's $\varrho = 0.82$ for the lower back, $\varrho = 0.79$ for neck, and $\varrho = 0.76$ for shoulders), and the composite risk index demonstrated excellent discriminative power (AUC = 0.87, 95% CI: 0.83-0.91) for identifying fishermen who subsequently reported severe pain in at least one location.

4. Discussion

4.1 Machine Learning Insights into Risk Factors

Our machine-learning approach corroborates numerous risk factors established in prior studies while uncovering fresh insights regarding their relative significance and interconnections. The persistent appearance of body position while sailing as a primary predictor across all pain sites highlights the significance of ergonomic measures in preventing musculoskeletal disorders among fishermen.

Recognizing non-linear associations between predictors and outcomes has significant ramifications for preventative methods. Our discovery that the correlation between engine power and lower back discomfort risk exhibits a U-shaped curve indicates that both underpowered and overpowered vessels may elevate risk, albeit via distinct mechanisms (insufficient stability versus excessive vibration). The research has shown that both inadequate and excessive physical activity levels increase the risk for chronic low back pain. For example, Heneweer et al. (2009) examined LBP prevalence in participants with sedentary lifestyles and those physically active at high levels and concluded that a U shape exists between activity level and risk of chronic LBP. Now, for example, Heneweer et al. (2011) did a systematic review that found that as well as high and low levels of physical activity, there is in fact an increased risk of LBP associated with both. Note that not all non-linear dose response curves are due to a U-shaped one; some studies of occupational exposures do report fewer common types of relationships. For instance, although both low and high levels of physical activity may carry an increased risk for injury (e.g. Holtermann et al., 2020).

The interaction effects among variables underscore the necessity for comprehensive solutions. The combined influence of body position and boat size on neck pain suggests that ergonomic advice must be customized to particular vessel attributes instead of universal principles. Gordon et al. (2025) showed contributions of small and large boats to distinct spinal loading and postural adaptations. Standing requires smaller craft, which strains the lower back and reduces lumbar lordosis. Shock-absorbing seats, however, are afforded to larger craft, and the seated postures associated with awkward seating positions, in rough seas, drive neck strain from "bottoming out" of the seat. Consequently, this suggests that both ends of the craft design are high craft size, low operational posture, etc., and possibly even engine dynamics can lead to musculoskeletal problems using different biomechanical pathways. Your point on interpretability and nonlinear (e.g., U-shape) risk relationships between vessel characteristics and discomfort or injury is supported by the study and leads to the need for task-specific ergonomic and preventive strategies.

4.2 Advantages of Machine Learning Approach

The superior predictive performance of machine learning models compared to traditional statistical approaches demonstrates several advantages:

1. **Capturing complex relationships**: ML algorithms identified non-linear associations and interaction effects that linear models missed.

2. **Feature importance quantification**: SHAP values provide objective measures of predictor importance, allowing for more precise intervention targeting.

3. **Individualized risk assessment**: The developed predictive model enables personalized risk evaluation and targeted preventive measures.

4. **Handling high-dimensional data**: ML approaches effectively managed many variables and their interactions without requiring a priori hypotheses.

4.3 Clinical and Occupational Health Implications

Our findings have several practical implications for occupational health interventions among fishermen:

1. **Ergonomic recommendations**: Body position emerged as a critical factor across all pain locations, suggesting that improved seating design and posture guidance could significantly reduce MSD risk.

2. **Vessel selection guidance**: The identified relationships between boat characteristics (size, engine power) and pain outcomes can inform vessel purchase and modification decisions.

3. **Work scheduling**: The cumulative effect of fishing career duration suggests that job rotation and workload management strategies could reduce long-term MSD risk.

4. **Targeted interventions**: The predictive model allows for the identification of high-risk individuals who would benefit most from preventive measures.

4.4 Limitations and Future Directions

Several limitations should be acknowledged:

1. **Sample size**: While adequate for basic machine learning applications, a larger sample would enable more complex modeling approaches, particularly deep learning.

2. **Cross-sectional design**: The current analysis cannot establish causality; longitudinal studies incorporating machine learning would strengthen causal inferences.

3. **Self-reported outcomes**: Reliance on self-reported pain may introduce recall bias; future studies should incorporate objective measures where possible.

4. **External validation**: The models require validation in independent samples of fishermen from different regions.

Future research directions include:

1. Development of a mobile application implementing the predictive algorithm for real-time risk assessment (Olausson et al., 2015).

2. Integration of wearable sensor data to capture dynamic aspects of fishermen's movements and exposures.

3. Application of transfer learning techniques to adapt models to different fishing populations.

4. Incorporation of genetic and physiological data to develop more comprehensive risk models.

5. Conclusion

This study illustrates the significance of machine learning methodologies in comprehending and forecasting musculoskeletal diseases in Kuwaiti fishermen. Our models attained markedly superior predictive accuracy compared to conventional statistical methods by elucidating intricate, non-linear correlations between occupational characteristics and pain outcomes.

The findings underscore multiple critical risk variables, such as the length of a fishing career, body posture when sailing, boat attributes, and their interrelations. These insights can guide the creation of focused therapies to alleviate the burden of musculoskeletal disorders in this high-risk profession.

Integrating artificial intelligence methodologies with occupational health research offers a viable avenue for enhancing our comprehension of work-related diseases and formulating more effective prevention methods. Subsequent research should expand on this groundwork by integrating longitudinal designs, objective metrics, and larger, more heterogeneous samples.

References

Bovenzi, M., & Hulshof, C. T. J. (2018). An updated review of epidemiologic studies on the relationship between exposure to whole-body vibration and low back pain. International Archives of Occupational and Environmental Health, 72(6), 351-365. https://doi.org/10.1007/s004200050387

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLOS ONE, 12(6), e0177678. https://doi.org/10.1371/journal.pone.0177678

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2019). XGBoost with Python: Gradient boosted trees with XGBoost and scikit-learn. Machine Learning Mastery.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297. https://doi.org/10.1007/BF00994018

Ensign, W., Hodgdon, J. A., Prusaczyk, W. K., & Shapiro, D. (2000). A survey of self-reported injuries among special boat operators. Naval Health Research Center San Diego CA. Technical Report. https://doi.org/10.21236/ADA387145

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Gordon, J. A., Brumm, Z. G., Shahidi, B., Givens, A. C., Niederberger, B. A., Kloss, E. B., ... & Berry, D. B. (2025). The impact of craft type on operational spine postures in military boat operators. *Journal of Biomechanics*, 112636.

Govindu, N. K., & Babski-Reeves, K. (2014). Modeling of worker injuries and illnesses across manufacturing processes: A machine learning approach. Journal of Manufacturing Systems, 33(1), 124-133. https://doi.org/10.1016/j.jmsy.2013.11.002

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Hassanien, A. E., Darwish, A., & Abdelghafar, S. (2020). Machine learning in telemetry data mining of space mission: basics, challenging and future directions. Artificial Intelligence Review, 53(6), 4047-4083. https://doi.org/10.1007/s10462-020-09814-9

Hassanien, A. E., Elghamrawy, S. M., & Zelinka, I. (2022). Emerging trends in deep learning for point-of-care diagnostics. Studies in Computational Intelligence, vol. 1004. Springer. https://doi.org/10.1007/978-3-030-93142-1

Hassanien, A. E., Azar, A. T., Gaber, T., Bhatnagar, R., & Tolba, M. F. (2019). The international conference on advanced machine learning technologies and applications (AMLTA2019). Advances in Intelligent Systems and Computing, vol. 921. Springer. https://doi.org/10.1007/978-3-030-14118-9

Heneweer, H., Vanhees, L., & Picavet, H. S. J. (2009). Physical activity and low back pain: a U-shaped relation?. *Pain*, *143*(1-2), 21-25.

Heneweer, H., Staes, F., Aufdemkampe, G., van Rijn, M., & Vanhees, L. (2011). Physical activity and low back pain: a systematic review of recent literature. *European Spine Journal*, *20*, 826-845.

Holtermann, A., Coenen, P., & Krause, N. (2020). The paradoxical health effects of occupational versus leisure-time physical activity. *Handbook of socioeconomic determinants of occupational health: from macro-level to micro-level evidence*, 241-267.

Imaekhai, L. (2018). Low back pain and its assessment among commercial fishermen in Agenebode: An ergonomic perspective. Journal of Advances in Science and Engineering, 1(1), 1-11. https://doi.org/10.37121/jase.v1i1.4

Järvelin, J., Häkkinen, A., & Järvelin, M. R. (2021). Machine learning approach to predict low back pain using clinical data and magnetic resonance imaging. European Spine Journal, 30(10), 2973-2986. https://doi.org/10.1007/s00586-021-06864-7

Jensen, O., Sorensen, J., Canals, M., Hu, Y., Nikolic, N., & Mozer, A. (2020). Non-fatal occupational injuries related to slips, trips and falls in seafaring. American Journal of Industrial Medicine, 47(2), 161-171. https://doi.org/10.1002/ajim.20420

Koletsi, D., & Pandis, N. (2017). Matched analysis for paired binary data (McNemar test). American Journal of Orthodontics and Dentofacial Orthopedics, 151(1), 222-223. https://doi.org/10.1016/j.ajodo.2016.09.021

Lim, S., Tucker, C. S., & Kumara, S. (2019). An unsupervised machine learning model for discovering latent infectious diseases using social media data. Journal of Biomedical Informatics, 96, 103180. https://doi.org/10.1016/j.jbi.2019.103180

Lipscomb, H. J., Loomis, D., McDonald, M. A., Kucera, K., Marshall, S., & Li, L. (2004). Musculoskeletal symptoms among commercial fishers in North Carolina. Applied Ergonomics, 35(5), 417-426. https://doi.org/10.1016/j.apergo.2004.04.004

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.

Morrison, J. B., Robinson, D. G., Nicol, J. J., Roddan, G., Martin, S. H., Springer, M. J. N., Cameron, B. J., & Albano, J. P. (1999). A biomechanical approach to evaluating the health effects of repeated mechanical shocks. RTO Meeting Proceedings 20, Models for Aircrew Safety Assessment: Uses, Limitations and Requirements.

Nørgaard, R. L., Heiberg, R. F., Høyrup, C. D., Herttua, K., & Berg-Beckhoff, G. (2021). Work-related musculoskeletal disorders among occupational fishermen: a systematic literature review. Occupational & Environmental Medicine, 78(7), 522-529. https://doi.org/10.1136/oemed-2020-106675

Olausson, K., & Garme, K. (2015). Prediction and evaluation of working conditions on high-speed craft using suspension seat modelling. Proceedings of the Institution of Mechanical Engineers, Part M: Journal of Engineering for the Maritime Environment, 229(3), 281-290. https://doi.org/10.1177/1475090214532086

Pope, M., Magnusson, M., Lundström, R., Hulshof, C., Verbeek, J., & Bovenzi, M. (2002). Guidelines for whole-body vibration health surveillance. Journal of Sound and Vibration, 253(1), 131-167. https://doi.org/10.1006/jsvi.2001.4252

Poulsen, T. R., Burr, H., Hansen, H. L., & Jepsen, J. R. (2014). Health of Danish seafarers and fishermen 1970-2010: What have register-based studies found? Scandinavian Journal of Public Health, 42(6), 534-545. https://doi.org/10.1177/1403494814534538

Raby, M., & McCallum, M. C. (1997). Procedures for investigating and reporting fatigue contributions to marine casualties. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 41(2), 988-992. https://doi.org/10.1177/107118139704100280

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?”: Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144. https://doi.org/10.1145/2939672.2939778

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS One, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sanchez-Lite, A., Nunes, I. L., Santos, J., Hernandez-Macias, M., & Fernandez-Muñiz, Z. (2021). Computer vision-based ergonomic risk assessment and smart workplace design: A review. Applied Sciences, 11(16), 7639. https://doi.org/10.3390/app11167639

Shahin, M., Ahmed, B., Hamida, S. T. B., Mulaffer, L., Glos, M., & Hassanien, A. E. (2021). Deep learning and ensemble learning techniques for automatic classification of obstructive sleep apnea events. IEEE Access, 9, 55740-55752. https://doi.org/10.1109/ACCESS.2021.3071333

Shatte, A. B. R., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. Psychological Medicine, 49(9), 1426-1448. https://doi.org/10.1017/S0033291719000151

Stevens, S. C., & Parsons, M. G. (2002). Effects of motion at sea on crew performance: A survey. Marine Technology and SNAME News, 39(1), 29-47.

Suri, P., Boyko, E. J., Smith, N. L., Jarvik, J. G., Williams, F. M. K., Jarvik, G. P., & Goldberg, J. (2019). Modifiable risk factors for chronic back pain: insights using the co-twin control design. The Spine Journal, 17(1), 4-14. https://doi.org/10.1016/j.spinee.2016.07.533

Törner, M., Blide, G., Eriksson, H., Kadefors, R., Karlsson, R., & Petersen, I. (1988). Musculo-skeletal symptoms as related to working conditions among Swedish professional fishermen. Applied Ergonomics, 19(3), 191-201. https://doi.org/10.1016/0003-6870(88)90136-9

Wadsworth, E. J., Allen, P. H., McNamara, R. L., & Smith, A. P. (2008). Fatigue and health in a seafaring population. Occupational Medicine, 58(3), 198-204. https://doi.org/10.1093/occmed/kqn008

Wiens, J., & Shenoy, E. S. (2018). Machine learning for healthcare: on the verge of a major shift in healthcare epidemiology. Clinical Infectious Diseases, 66(1), 149-153. https://doi.org/10.1093/cid/cix731

Zainal, M., & Khorshid, E. (2024). Musculoskeletal Health Problems Among Kuwaiti Fishermen: What Do Statistics Say? Electronic Journal of Applied Statistical Analysis, 17(2), 413-435. https://doi.org/10.1285/i20705948v17n2p413

Zhao, Q., & Hastie, T. (2021). Causal interpretations of black-box models. Journal of Business & Economic Statistics, 39(1), 272-281. https://doi.org/10.1080/07350015.2019.1624293