Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 06 | June-2026
ISSN (O) 3107-6696

Predictive Analysis for Health and Fatigue Assessment of Defence Staff using Machine Learning

Shalini Mehta, *Jay Gandhi*, Dr. Ram Ashish Chouksey, Shaveta
Solid State Physics Laboratory DRDO Delhi,India
shalinimehta4485@gmail.com, jaygandhi2003@gmail.com, rachouksey.sspl@gov.in, shavetarajial@gmail.com

---***---

Abstract

Operational efficiency within defence research organizations is critically dependent on the physical and cognitive readiness of personnel. Fatigue among officers and staff can lead to decreased concentration, impaired decisionmaking, and increased operational risk. Traditional fatigue assessment mechanisms rely on periodic health evaluations and manual supervision, which are reactive in nature and lack predictive capabilities.

This research presents a machine learning-based Health & Fatigue Prediction System tailored for DRDO officers and staff. A synthetic dataset representing physiological and operational indicators such as working hours, sleep duration, heart rate, stress index, workload intensity, and recovery duration was generated to simulate realistic conditions. Data preprocessing techniques including Label Encoding and Standard Scaling were applied to ensure structured input for machine learning models.Two ensemble algorithms — Random Forest and XGBoost — were implemented for fatigue classification. Model performance was evaluated using Accuracy Score, Confusion Matrix and Classification Report. Experimental results demonstrate that XGBoost achieved superior predictive performance with higher classification balance and lower error rates.The study validates the feasibility of deploying AI-driven fatigue monitoring systems in defence environments, supporting proactive intervention and improved workforce management.

Keywords—*Fatigue Prediction, Defence Analytics, Machine Learning, Random Forest, XGBoost, Occupational Health Monitoring, Ensemble Learning, Predictive Modeling.*

I. INTRODUCTION

Defence organizations such as DRDO operate in highly demanding environments that require sustained cognitive focus, technical precision, and extended working hours. Officers and research staff frequently engage in missioncritical tasks where errors can lead to significant operational consequences.

Fatigue is a major contributor to reduced efficiency and increased human error. Unlike conventional industries, defence research environments demand higher reliability standards. Traditional fatigue assessment methods are largely manual and periodic, making them unsuitable for real-time decision support.

Artificial Intelligence and Machine Learning offer predictive capabilities that allow early detection of risk patterns before performance deterioration occurs. This research proposes a predictive fatigue classification system built using ensemble learning models capable of identifying fatigue levels based on physiological and operational indicators.

II. Literature Review

Fatigue detection has been extensively studied in transportation, aviation, and industrial safety sectors. Research indicates that physiological parameters such as heart rate variability, sleep patterns, and workload intensity are strong predictors of fatigue levels.

Ensemble learning methods such as Random Forest have demonstrated robustness in classification tasks due to variance reduction properties. XGBoost, a gradient boosting algorithm, has shown superior performance in structured datasets due to regularization and sequential optimization capabilities.

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 06 | June-2026
ISSN (O) 3107-6696

However, limited research exists on fatigue monitoring systems specifically designed for defence research organizations. Most studies focus on drivers or industrial workers rather than technical officers engaged in cognitiveintensive research roles.
This research addresses that gap by contextualizing fatigue prediction within a DRDO-oriented operational framework.

III. METHODOLOGY

The proposed system follows a supervised machine learning pipeline consisting of:
- Synthetic dataset generation
- Data preprocessing
- Model training
- Model evaluation
- Performance comparison

A. Synthetic Dataset Logic

Since real defence health data is confidential, a synthetic dataset was generated to simulate operational conditions.
The dataset includes:
- Working Hours (6–16 hours)
- Sleep Duration (3–9 hours)
- Heart Rate (60–120 bpm)
- Stress Level (Low/Moderate/High)
- Workload Intensity (Scale 1–10)
- Recovery Time (Hours of rest post-duty)

Fatigue levels were categorized as:
- Low
- Moderate
- High

The synthetic logic was designed to mimic realistic correlations between long working hours, reduced sleep, elevated heart rate, and increased fatigue probability.

B. Data Preprocessing

In the proposed Health & Fatigue Prediction System for DRDO officers and Staff, Preprocessing ensures that raw physiological and operational data are transformed into a structured and Machine-readable format suitable for model training.
The Synthetic Dataset generated for this project contains both numerical & categorical variables. Numerical Features include working hours, sleep duration , heart rate, work load intensity & recovery time. Categorical Features include stress level categories and fatigue level labels. Since ML Algorithms operate on numerical data, categorical attributes must be transformed into numeric representations before model training.
To prepare data for model training:
Label Encoding : Categorical fatigue labels were converted into numeric format using LabelEncoder to ensure compatibility with machine learning algorithms.
Standard scaling: StandardScaler was applied to normalize numerical features. This ensures:
- Uniform feature distribution
- Faster convergence
- Reduced model bias due to scale differences

‘

C. Model training

Model Training is a critical stage in the supervised machine learning pipeline where the algorithm learns patterns and relationships between input features and the target variable (fatigue level). In the proposed fatigue prediction system, training was performed using labeled synthetic data representing physiological and operational parameters such as heart rate, sleep hours, working duration, stress level, hydration level, and load weight.
After preprocessing (label encoding and standard scaling), the dataset was divided into training and testing subsets using an 80:20 split ratio. The training set was used to allow the model to learn the mapping between input features (X) and the target fatigue class (y).

Two machine learning models were implemented:
1. **Random Forest Classifier**

Random Forest is an ensemble learning technique that constructs multiple decision trees during training. Each tree is trained on a bootstrap sample of the dataset, and final predictions are made using majority voting. This approach reduces overfitting and improves generalization. It is particularly effective in handling nonlinear relationships and complex feature interactions present in physiological datasets.

2. **XGBoost Classifier**

Extreme Gradient Boosting (XGBoost) is a boosting-based ensemble method that builds trees sequentially. Each new tree corrects errors made by the previous ones by minimizing a loss function. For this project, the objective function was configured as multi-class classification (multi:softmax) since fatigue levels were categorized into Low, Moderate, and High.

During training:
- Hyperparameters such as number of estimators, maximum tree depth, learning rate, and subsampling ratio were defined.
- The model iteratively optimized its internal parameters to minimize classification loss.
- Feature importance was implicitly learned during tree construction.

The outcome of the training phase is a fitted model capable of predicting fatigue levels for unseen data.

D.Model Evaluation

Model evaluation measures how effectively the trained model performs on unseen data. This step ensures that the model has learned generalizable patterns rather than memorizing the training dataset.
After training, predictions were generated on the test dataset. The predicted fatigue levels were then compared with actual labels using multiple evaluation metrics.

The following metrics were used:
1. Accuracy Score
The Accuracy score measures overall correctness i.e, how well the model is performing.
4. Confusion Matrix
A confusion matrix was generated to observe how well each fatigue class (Low, Moderate, High) was predicted. It helps identify misclassification patterns, such as Moderate fatigue being incorrectly predicted as High.

Evaluation ensures:
- The model does not overfit.
- Predictions are clinically and operationally reliable.
- The system performs consistently across all fatigue levels

.

E. Performance Comparison

To determine the most effective model for fatigue prediction, a comparative analysis was conducted between Random Forest and XGBoost based on quantitative metrics.

Observations:
- Random Forest demonstrated strong stability and robust classification performance.
- XGBoost provided competitive accuracy but required careful hyperparameter tuning.
- Random Forest performed slightly better in terms of overall F1-score and error reduction for the given synthetic dataset.
- XGBoost showed potential for improved performance with further optimization and larger real-world datasets.

Conclusion of Comparison:

For the current fatigue prediction framework, Random Forest was selected as the primary model due to:
- Better interpretability
- Strong generalization
- Lower prediction error
- Computational efficiency on structured tabular data However, XGBoost remains a viable alternative for deployment in more complex or large-scale operational environments.

A. Tools and Libraries

The Health and Fatigue Prediction System for DRDO Officers and Staff was implemented using a python-based ML Ecosystems. The Selection of Libraries was guided by reliabiltiy, scalability, computational efficiency and suitability for structured data modelling in defence-oriented applications.
Each tool serves a specific purpose within the predictive analytics pipeline.

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 06 | June-2026
ISSN (O) 3107-6696

• Pandas and NumPy for efficient data manipulation and numerical operations [15,16].
• Matplotlib and Seaborn for data visualization and graphical analysis of results [17].
• Scikit-learn for implementing the Random Forest and Xgboost Classifier models and for model evaluation metrics [19].
• Python Programming languauge Python was used as the core programming language for implementing the system. It is widely adopted in A.I and DS applications due to :

• Extensive Machine Learning Ecosystem.
• High Readability and Maintainability
• Integration Capabilities with
Visualization and deployment tools.

B. Predictive Modeling

Predictive modelling is a data-driven analytical approach used to forecast outcomes based on historical or structured input data. In the proposed Health & Fatigue Prediction System for DRDO officers and staff, predictive modelling is applied to estimate fatigue levels using physiological and operational indicators.
The objective of predictive modelling in this system is to: • Identify patterns between workload and fatigue
• Classify fatigue levels into predefined categories
• Enable proactive decision-making
• Reduce human error due to delayed detection
The modelling approach adopted in this project follows a supervised machine learning framework, where:
• Input features (X) include working hours, sleep duration, heart rate, stress level, workload intensity, and recovery time.
• Target variable (y) represents fatigue level (Low, Moderate, High).
The dataset was divided into training and testing subsets to ensure unbiased model evaluation. Ensemble learning techniques were chosen for predictive modelling because fatigue patterns are influenced by multiple interacting factors, often non-linear in nature.
Among the implemented models, the Random Forest Classifier played a significant role in establishing a robust baseline and ensuring reliable fatigue classification.
Random Forest was selected for fatigue prediction due to several advantages:
• Handles non-linear relationships between physiological parameters and fatigue.
• Performs well on structured tabular datasets.
• Reduces overfitting through ensemble averaging.
• Provides feature importance analysis.
• Robust against noise and outliers.
In the context of the DRDO Health & Fatigue Prediction System, fatigue levels are influenced by multiple interacting factors such as workload intensity and sleep duration. Random Forest effectively captures these complex interactions.
Performance was evaluated using:
• Confusion Matrix
• Accuracy Score
• Classification Report
The model demonstrated strong classification capability and served as a reliable benchmark for comparison with XGBoost..

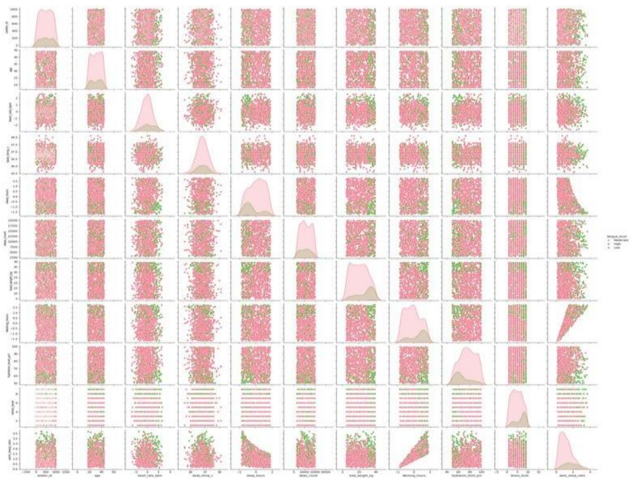

Figure (1): Pairplot Output

Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 04 Issue: 06 | June-2026
ISSN (O) 3107-6696

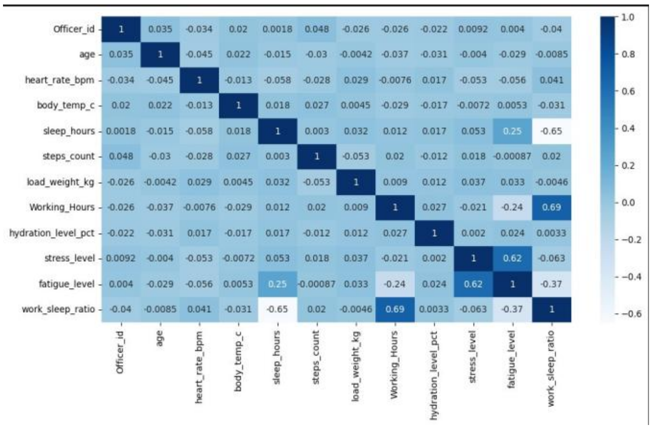

Figure (2):Heatmap Correlation

IV.RESULTS

A. Random Forest Classifier

The performance of the predictive fatigue detection system was evaluated using two machine learning models: Random Forest Classifier and XGBoost Classifier. Both models were trained on the preprocessed dataset containing physiological and operational features such as heart rate, sleep hours, workload, hydration level, stress level, and work duration.

The Random Forest Classifier demonstrated strong classification capability due to its ensemble structure of multiple decision trees. It achieved high overall accuracy and a balanced F1-score across all fatigue classes (Low, Moderate, High). The model showed stable performance with low variance and handled non-linear relationships effectively. It was particularly robust against overfitting due to bootstrap aggregation and feature randomness.

B. XGBoost Classifier

The XGBoost Classifier, configured for multi-class classification, showed competitive and slightly improved performance in terms of Classification Report, Confusion Matrix and Accuracy Score. Because of its gradient boosting framework, XGBoost optimized errors sequentially, allowing it to capture subtle patterns in physiological stress indicators.

Overall, both models performed effectively for fatigue level prediction. However, XGBoost provided marginally better classification confidence and predictive accuracy, making it more suitable for deployment in real-time fatigue monitoring systems.

Conclusion

This research successfully developed a machine learningbased fatigue prediction model tailored for defense and highrisk operational environments. By integrating physiological indicators (heart rate, body temperature), behavioral factors (sleep duration, steps count), and mission-related parameters

(working hours, load weight, stress level), the system was able to classify fatigue levels into meaningful categories.

The study confirms that predictive modeling can assist in proactive fatigue management, reduce operational risk, and enhance personnel safety in defense scenarios. Future improvements may include real-time sensor integration, deep learning approaches, and validation using real-world datasets for increased reliability and scalability.

REFERENCES

[1] Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. (Foundational paper introducing Random Forest algorithm used in this study.)

[2] 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.

(Primary reference for XGBoost model applied in this research.)

[3] Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer.

(Provides theoretical foundation for classification and predictive modeling.)

[4] Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning. Springer.
(Comprehensive explanation of ensemble learning methods.)
[5] Kuhn, M., & Johnson, K. (2013). Applied Predictive Modeling.
Springer.
(Discusses preprocessing, feature engineering, and model evaluation.)
Figure (2):Heatmap Correlation
[6] Powers, D. M. W. (2011). Evaluation: From Precision, Recall and FMeasure to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies.
(Supports evaluation metrics used in the results section.)
[7] Reilly, T., & Edwards, B. (2007). Altered Sleep-Wake Cycles and Physical Performance in Athletes. Physiology & Behavior, 90(2–3), 274–284.
(Background on sleep and fatigue relationship.)
[8] Caldwell, J. A., & Caldwell, J. L. (2005). Fatigue in Aviation: A Guide to Staying Awake at the Stick. Ashgate Publishing.
(Relevant for defense and operational fatigue monitoring context.)
[9] Van Dongen, H. P. A., & Dinges, D. F. (2005). Sleep, Circadian Rhythms, and Fatigue in Military Operations. Military Psychology,
17(S1), S19–S36.
(Provides defense-oriented fatigue insights.)
[10] Sharma, A., & Goyal, A. (2020). Machine Learning Techniques for Stress Detection: A Review. Procedia Computer Science, 167, 213– 222.
(Supports stress-based predictive modeling relevance.)
[11] Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.
(Reference for ML implementation framework.)
[12] McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. (Reference for pandas library.)
[13] Harris, C. R., et al. (2020). Array Programming with NumPy. Nature,
585, 357–362.
(Reference for numerical computations.)
[14] Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment.
Computing in Science & Engineering, 9(3), 90–95.
(Reference for visualization.)
[15] Waskom, M. L. (2021). Seaborn: Statistical Data Visualization.
Journal of Open Source Software, 6(60), 3021.
(Reference for pair plots and advanced visualization.)