Optimizing Athlete Workload Monitoring with Supervised Machine Learning for Running Surface Classification Using Inertial Sensors

WenBin Zhu[1], QianWei Zhang[2]*, SongYan Ni[3]
Chengdu Sport University, Chengdu, Sichuan, 610041, China[1, 2]
Sichuan University High School, No. 12 High School of Chengdu, Sichuan, 610061, China[3]

***Abstract*—Monitoring athlete movement is important to improve performance, reduce fatigue, and decrease the likelihood of injury. Advanced technologies, including computer vision and inertial sensors, have been widely explored in classifying sport-specific movements. Combining automated sports action labeling with athlete-monitoring data provides an effective approach to enhance workload analysis. Recent studies on categorizing sport-specific movements show a trend toward training and evaluation methods based on individual athletes, allowing models to capture unique features peculiar to each athlete. This is particularly beneficial for movements that exhibit large variations in technique between athletes. The current study uses supervised machine learning models, including Neural Networks and Support Vector Machines (SVM), to distinguish between running surfaces, namely, athletics track, hard sand, and soft sand, using features extracted from an upper-back inertial measurement unit (IMU) sensor. Principal Component Analysis (PCA) is applied for feature selection and dimensionality reduction, enhancing model efficiency and interpretability. Our results show that athlete-dependent training approaches considerably enhance the classification performance compared to athlete-independent approaches, achieving higher weighted average precision, recall, F1-score, and accuracy ($p < 0.05$).**

Keywords*—Athlete monitoring; machine learning models; running surface classification; Inertial Measurement Units (IMU); neural networks; Support Vector Machines (SVM); Principal Component Analysis (PCA)

I. Introduction

Supervised machine learning algorithms have emerged as flexible statistical tools capable of modeling both classification and regression problems. These algorithms use mathematical frameworks for model optimization that map input features to output targets of a training dataset to make accurate predictions on unseen data. Sports science has of late embraced the power of data-driven methodologies to track and improve the performance of athletes [1]. Of these, supervised machine learning and artificial neural networks have been at the forefront of automating sport-specific movement classification, quantifying athlete workload, and predicting physiological states-for example, fatigue-to optimize training regimens and reduce injury risks [2].

The introduction of MEMS has subsequently miniaturized IMUs into wearable devices that have revolutionized performance monitoring in elite sports by providing real-time, high-resolution data on athlete movements [3]. These normally have IMUs positioned at key areas around the body for the extraction of critical features that are important in the study of athletic performance in various sports. These feed into input-supervised machine learning models aimed at classifying specific sport movements or environmental contexts, such as the running surface [4]. However, these models are as good as their methodologies applied during the training and validation process. Data partitioning, whether athlete-dependent or athlete-independent, has a huge impact on model performance and generalizability [5].

Despite the progress made in applying machine learning to sports, recent systematic reviews have identified important shortcomings in the validation strategies. Studies tend to rely on non-independent data splits, such as cross-validation or simple train-test partitions, which often inflate classification performance. Conversely, leave-one-subject-out (LOSO) validation-a method that ensures athlete-independent evaluation-remains underutilized [6], [7]. This is a particularly important gap, as non-independent methods do not consider the inter-athlete variability that may result in models performing well on known data but poorly on new, unseen athletes. In sports applications, where the ability to adapt to new athletes is crucial, reliance on athlete-dependent validation methods risks compromising the broader applicability and reliability of machine learning solutions.

The aim of this work is to classify the running surfaces into athletics track, hard sand, and soft sand using the features computed from upper-back IMUs. To compare different approaches, six supervised machine learning models were trained and then evaluated using both athlete-dependent and athlete-independent methodologies. Although the former is generally more accurate because the model may learn features specific to a given individual, they do not generalize well for other unknown athletes. This work is, to the best of the authors' knowledge, one of the first sport-specific case studies directly comparing these methodologies in running surface classification. Its results are particularly relevant to sports organizations seeking to implement robust machine learning models in monitoring athletes: it conveys actionable insight regarding their training, validation, and deployment strategy. The current study has attempted to guide the development of more reliable and scalable solutions for sport-specific movement classification by underlining some of the trade-offs between accuracy and generalizability.

*Corresponding Author

The remainder of this paper is structured as follows: Section II presents a literature review, discussing relevant studies on athlete workload monitoring, running surface classification, and machine learning applications in sports analytics. Section III describes the methodology, detailing the data collection process, feature extraction from the inertial sensor, and the supervised machine learning models, including Neural Networks and Support Vector Machines (SVM), utilized for classification. Section IV presents the results, evaluating the classification performance of athlete-dependent and athlete-independent models based on multiple metrics. Finally, Section V concludes the study, summarizing key findings and highlighting potential future directions for improving athlete workload monitoring using advanced machine learning techniques.

II. Literature Review

Recently, a significant boost of machine learning applications in sports analytics was seen; thus, researchers started reaping full benefits of this technique while solving a number of problems connected with monitoring performance of athletes and classification of movements. Wearables, together with IMUs, have become an essential tool for motion data acquisition and further analysis of several features of athletic performances. Unlike traditional observation-based methods, IMUs are accurate and scalable; they can capture multidimensional data about an athlete's movements in real time. Works such as that by Umer and Riaz [8] show the capability of IMUs within gait analysis to identify ground contact events with high accuracy on different surfaces. It goes without saying that this type of research places increasing dependence on the use of IMU sensors within sports and rehabilitation applications [9].

Subsequently, machine learning models based on IMU data had very good performance, especially in the classification of environmental and movement context, such as running surfaces. Buckley et al. [10] presented a road surface type classification approach using IMU data, incorporating traditional machine learning algorithms like SVM and KNN with deep learning approaches. The results showed that machine learning could identify surface types with a high degree of accuracy, thus opening the way for possible applications in sports [11]. This study concerned transportation, but its implications reach to athletic performance, where running surface classification can improve workload monitoring and injury prevention.

One of the most critical issues when developing machine learning models to monitor athletes is the classification based on an athlete-dependent or independent approach. The athlete-dependent model is specific for particular features of a certain athlete, and this results in a higher accuracy if the test on the same subject is done. Most of these models fail to generalize when applied on different athletes. On the other hand, the athlete-independent model is general and permits variations in individual features. Koul et al. [10] discussed surface recognition regarding electric scooters using deep neural networks based on smartphone IMU sensors. While not directly related to running, their findings emphasize the importance of designing models that balance specificity and generalizability—principles that are highly relevant to sports performance monitoring.

The broader literature also underlines the increasing role of machine learning in sports injury prediction and prevention. Surveys such as Diss et al. [11] review various algorithms ranging from Random Forests to neural networks, using data derived from athletes in order to predict injury risks. Such studies indicate the potential of machine learning to analyze complex data sets and determine patterns associated with injury-prone conditions [12]. Though different from running surface classification, all these applications are unified in the aim of bettering athlete safety and optimizing performance using data-driven insights. The literature underscores the transformative potential of machine learning in sports analytics [13]. Although recent advancements in IMU-based models and validation methodologies have achieved higher classification performance, challenges still remain regarding how to balance accuracy and generalizability. The study contributes to the field by comparing the athlete-dependent and independent approaches to classify running surfaces, filling key gaps in existing research and informing future model development and deployment.

III. Methodology

A. Participants

Seven healthy subjects, four males and three females, participated voluntarily in this study and gave their informed consent. The group's mean age was 32.4 years, with a standard deviation of 17.89 years, which shows the very high variability in age among the group members. Their mean height was 171.9 cm, with a standard deviation of 8.91 cm, and their mean weight was 70.3 kg with a standard deviation of 16.87 kg. Ethical approval for this study was granted under protocol number GU 2017/587. The population of the subjects was heterogeneous regarding their fitness level and running experiences; consequently, it constituted a rich sample for the study's objectives. Their training routines varied, with some individuals reportedly spending up to nine hours a week training.

B. Experimental Design

This experiment consisted of running 400 meters at a light to moderate pace on three different surfaces: first, on soft, dry sand; then, on hard, water-saturated sand; and finally, the same on a synthetic tartan running track. This completed the trials for all surface conditions. Each run was designed to maintain consistency in pace and effort across the surfaces. Data of the motion were captured with one IMU per participant. The IMU was positioned near the third thoracic vertebra, T3, and fixed with a specifically developed sports harness not allowing any displacement during the runs. This setup made the sensor stable, and there was no interference with the data collection process [14], [15]. Fig. 1 depicts the orientation of sensor axes, which is important for accurate motion analysis. The figure shows that data acquisition was uniform across all participants and conditions.

C. IMU Sensor Technology

The device used was a custom-made, 9DOF IMU, designed at Chengdu Sport University. For this present research, it was based on the unit known as SABELSense (Sichuan, China) with a weight of 23 g and specified as +16 g accelerometer, +2000 deg/s gyroscope, and +7 Gauss magnetometer; all data was captured at a sampling frequency of 250 Hz. Each IMU output was then comprehensively calibrated before the trial to capture proper data. The data were logged locally onto a 4GB microSD card that enabled continuous, reliable logging of the experiment. 3D orientation of the sensor is obtained using Euler angles: roll, pitch, and yaw through the Madgwick AHRS algorithm. This has an accuracy characterized with a root mean square error below 0.8° for static, and below 1.7° in dynamics. This setup ensured that the motion data was valid and reliable during the study.

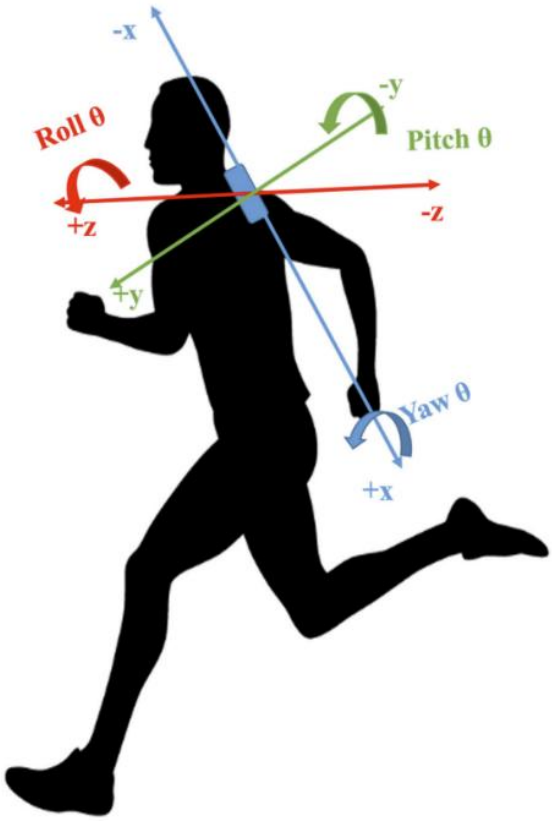

Fig. 1. The x-axis is the superior-inferior direction, the y-axis is the medial-lateral direction, and the z-axis is the anterior-posterior direction. The rotation around the z, y, and x axes corresponds to the roll, pitch, and yaw, respectively.

D. Designed Algorithm

1) Feature extraction: When running on a 400-metre athletics track, due to the curviness of the path, the Euler angles recorded drift progressively. In order to overcome this problem, a feature extraction method had been developed and was very robust, being inspired by different previous methodologies in which [16] is highly remarkable. By using a sliding window technique, the whole process was executed with the assistance of MATLAB from MathWorks, Natick, MA, USA:. This window was set to a duration of 4 seconds, while the overlap between two successive windows was set to 0.5 seconds. This would give approximately 10 to 11 strides on each surface, providing a reliable dataset and at the same time, reduce the effects of directional drift [17]. Normalization of Euler angle data and transforming into absolute values was done to nullify the effect of heading drift within each window. Too much spurious data was removed, which would have had an adverse effect on the classification. The above window-sliding procedure was applied for 11 data channels recorded by the IMU: acceleration components, gyroscope outputs, and orientation angles in three dimensions. For every one of these data windows, various features were extracted in time and frequency domains. These included the mean values, standard deviation, skewness, kurtosis, and dominant frequency components. These features together gave a full representation of the pattern of movement-a basis on which effective classification and analysis could be done.

2) Training-validation of feature data: The feature data was divided into training and validation sets using two different strategies. First, an athlete-independent LOSON was used: one participant was randomly selected for model evaluation, and the remaining six participants' data was used for training. In the second strategy, the data was divided in an athlete-dependent way, where 75% of the data was used for training and 25% for testing. These methods were applied to evaluate how individual participant features influenced the classification performance of the models. In the athlete-independent approach, Method 1, the number of training observations for soft sand, hard sand, and the athletics track were 1537, 1237, and 944, respectively. The respective test observations for the considered surfaces were 183, 153, and 196. By contrary, the approach dependent on athletes-Method 2 resulted in 1720, 1390, and 1140 training observations for soft sand, hard sand, and athletics track surface, respectively, leaving 413, 341, and 309 observations for testing. These two partitioning strategies have allowed a more complete assessment of the model for its ability both to generalize across individuals and to perform when fit to specific athletes.

3) Feature engineering: Feature engineering and model training were performed on Python, Python Software Foundation, https://www.python.org/ using popular libraries like scikit-learn and pandas [18, 19]. All the features were scaled into a uniform range from 0 to 1 using the mean and standard deviation of the training dataset before modeling. This way, the features were normalized, and no single feature biased the model due to its magnitude. The challenge in high dimensionality was approached by Principal Component Analysis (PCA) [20]. In this process, PCA transforms the original features into a new orthogonal set of variables known as principal components. Every principal component carries part of the dataset variance, and only those components needed to describe 95% of the total variance were retained for this study. This reduced the number of features from 132 to 45, thus

significantly reducing computational complexity and enhancing model efficiency. By removing noisy and redundant features, PCA also helped reduce overfitting and enhanced the generalization capability of the model.

Among the unsupervised dimensionality reduction techniques, PCA was preferred over supervised ones such as LDA because of its advantages in cases with limited training samples. Unlike these supervised methods, PCA does not depend on class labels and hence avoids the bias toward a particular subset of data. Thus, PCA is particularly suitable for the comparisons among the athlete-independent and the dependent methods. Even though LDA is originally designed to maximize class separability, it could amplify overfitting in the case of limited training data, an important concern in this study. Moreover, previous studies have demonstrated that PCA performs better than LDA when sample sizes per class are small, which further supports the appropriateness of the choice for this study [21] [22]. By applying PCA, the research was able to balance the computational efficiency with feature relevance; hence, the model can process meaningful information without getting overwhelmed by noise or irrelevant data. This embedding had not only reduced the computational burden and reduced training time but also made it a just comparison between the athlete-independent and athlete-dependent methodologies during the conduct of the research, making PCA an essential part of the feature engineering pipeline.

4) Model training and evaluation: Six different machine learning models were developed and tested to classify sport-specific movements using data from inertial sensors. These include some of the most commonly used movement classification models: logistic regression (LR), support vector machines (SVM) with linear (LSVM) and Gaussian radial basis function (GSVM) kernels, multilayer perceptron neural networks (MLP-NN), random forests (RF), and gradient boosting machines (XGB) [23] [24]. Model configurations were selected without hyperparameter tuning in order to provide a baseline for comparisons. Logistic Regression models relied on an L2 penalty while using the lbfgs solver. Support vector machines consisted of a linear kernel, with C = 1 and a Gaussian kernel, with C = 1 and gamma = scale. The neural network, MLP, consisted of three layers of 8 nodes, ReLU activation, constant learning rate, and Adam optimizer. The random forest model was set with the Gini criterion for impurity, number of features as the maximum feature parameter, and included 20 estimators. The gradient boosting model used the Friedman mean squared error (mse) criterion, deviance loss, a maximum depth of 3, and 100 estimators.

The models were then evaluated in their classification of running surfaces using both athlete-dependent and athlete-independent training and validation segmentation methods. The performance metrics for each classification technique included weighted averages of precision, recall, and F1-score and the overall accuracy for classification. The statistical comparisons between models that have used two segmentation methods employed a paired t-test, $\alpha = 0.05$ as shown in Fig. 2.

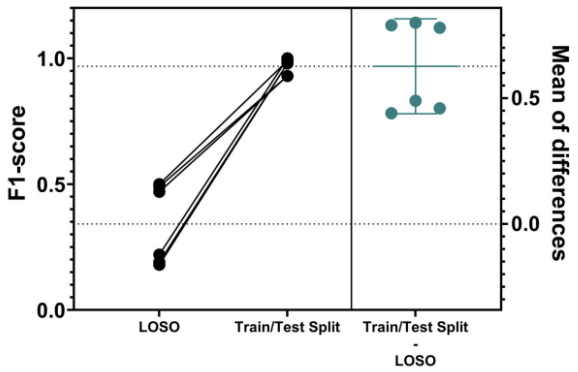

Fig. 2. Estimation plot showing the significant difference in F1-scores across all models comparing the train/test split to the LOSO validation.

IV. Results

The statistical results of the test on the train/test split provided significantly higher values for all model types with respect to the predictive performance measure. More concretely, these are huge increases in the weighted averages for precision, recall, F1-score, and accuracy, with p-values 0.0002, 0.0004, 0.0004, and 0.0004, respectively. This result points once more to the importance of letting the models see all participant features during training and hence letting them capture the variabilities in individual movement patterns. Results indicate large differences between F1-scores from the two validation methods; Fig. 2 provides further visualization, whereas a detailed comparison of various evaluation metrics for models considering both two validation methods can be seen in Fig. 3(a) – Fig. 3(d).

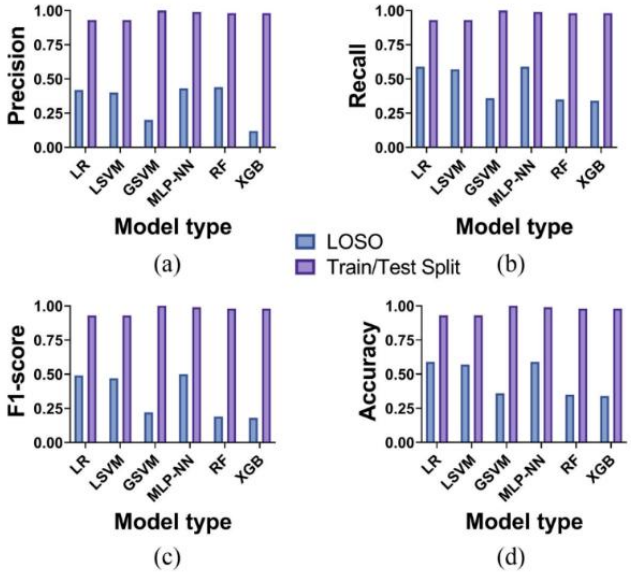

Fig. 3. Comparison of evaluation metrics of all models, considering both the train / test split and LOSO validation methods: (a) weighted precision, (b) weighted recall, (c) weighted F1-score, and (d) overall classification accuracy.

As observed by the method of train/test split, using training based on participant-specific features pays off significantly in

the case of all the models. Such exposure enables them to learn complex patterns that are unique in different running styles and surface interactions, thus giving enhanced accuracy of classification. However, to generalize this application into generalized surface classification-for instance, across diverse populations or settings-the participants in the dataset would have to be increased in number for broader applicability. Among the models tested, the best performances were from the MLP-NN under the LOSO and the GSVM under the train/test split. Their classification capabilities are described by confusion matrices shown in Fig. 4 and Fig. 5. The MLP-NN model was in a fairly medium range when segmenting soft sand from the remaining two surface classes using the LOSO method. The precision by classification is 0.71, recall 0.89, and F1-score 0.79 for the soft sand class; hence, one can say it identified the features of running on the soft sand rather successfully. This agrees with highly observable changes in gait mechanics when walking on soft sand compared to harder surfaces. This is expected since the closer physical properties result in more misclassifications between these two surface types of hard sand and athletics track surfaces. On the other hand, the GSVM model, when tested by the method of train/test split, it gave a general accuracy of 0.99. Actually, the model's performance was really good on all surfaces, misclassifying only once between the athletics track and hard sand surface. This, therefore, ascertains how the GSVM model can handle such subtlety in running pattern variations across surface types. With a full presentation of all information during training that prepares this model to generalize best on all test datasets, even then, an optimum quotient from GSVM with a train/test split may be expected.

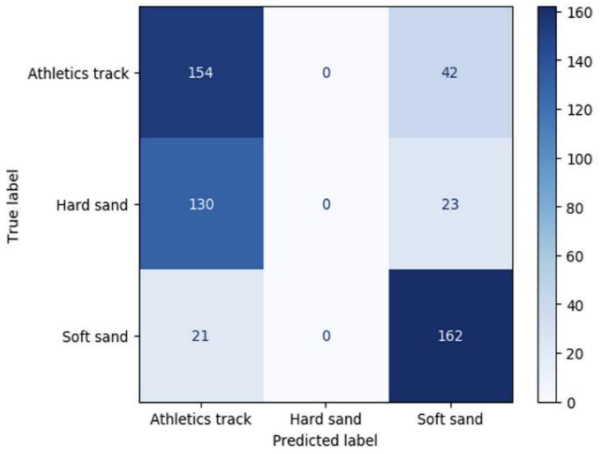

Fig. 4. Confusion matrix illustrating the classification performance of the MLP-NN model using the LOSO validation method.

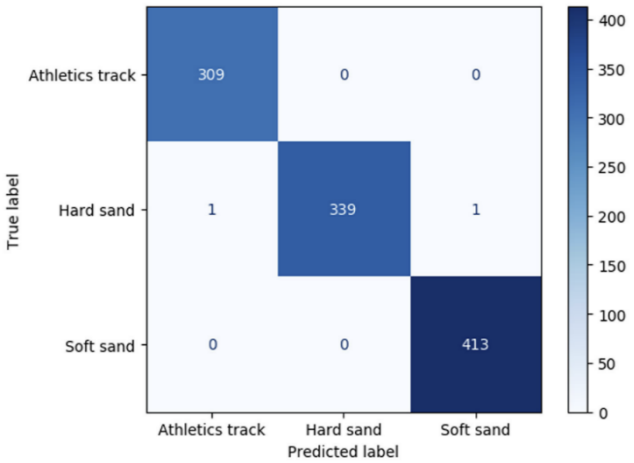

Fig. 5. Confusion matrix showcasing the classification performance of the GSVM model using the train / test split validation method.

These results have confirmed the intuition that a specific relationship exists between certain forms of validations pursued on classifications' various results. This was especially evident from the train/test split, representing strengths in model accuracy improvements using participant-specific features. In tasks where such information is available, this might prove very effective. However, the LOSO approach could be more fitting when, at application time, generalization to unseen individuals with one universal model is required. The findings from this study underpin some critical trade-offs between accuracy and generalizability for the classification of athlete movement and provide valuable insights into the development and implementation of machine learning models in sports analytics.

V. Conclusion

This study underlines, with greater significance, the improved classification performance that is attained with athlete-dependent train/test split methods ($p < 0.05$). Individual differences in the execution of movements are key to monitoring athletes, and allowing models to learn such specific features significantly enhances the accuracy of classification. A generalized sport-action classification model should have high performance on completely independent athletes; hence, it requires training data from a diverse group of individuals. This would include the type, body of the cyclists (height and weight), standard, physical fitness of the cyclists. However, due to issues of privacy, hardly any athlete performance data can be provided, making the creation of a fully athlete-independent model very challenging. An issue that gets very crucial when there is high individual variation in the styles of movement is, for instance, while running on different surfaces. Given these limitations, any sporting organization looking to utilize

automated tagging of sport-specific actions as a way of supplementing current approaches to athlete-load monitoring would have to, at the very least, retrain the classification models on data from all participating athletes. This can also include a calibration session when new athletes join in, allowing the model to learn features from the new person. This proposed approach using an upper-back IMU sensor for running surface classification may, therefore, inherently be an athlete-dependent one. The proposal would still be very useful. Besides, it also holds prospect for adequate adjustment of an athlete's session work rate estimate, particularly on occasions when some direct physiological monitoring implements, like heart rate monitors, cannot be used.

In future work, the authors aim to develop privacy-preserving methodologies, such as federated learning, to facilitate athlete-independent classification models while ensuring data security. Additionally, they intend to integrate multi-modal sensor fusion techniques to enhance the robustness and generalizability of movement classification across various sports activities.

REFERENCES

[1] Cust, E.E., Sweeting, A.J., Ball, K. & Robertson, S., 2019. Machine and deep learning for sport-specific movement recognition: A systematic review of model development and performance. Journal of Sports Sciences, 37(5), pp.568–600.

[2] McGrath, J., Neville, J., Stewart, T. & Cronin, J., 2020. Upper body activity classification using an inertial measurement unit in court and field-based sports: A systematic review. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*. [Online] Available at: https://doi.org/10.1177/1754337120959754.

[3] Eyobu, O.S. & Han, D., 2018. Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network. *Sensors*, 18(9), pp.1–36.

[4] Wan, S., Qi, L., Xu, X., Tong, C. & Gu, Z., 2020. Deep learning models for real-time human activity recognition with smartphones. *Mobile Networks and Applications*, 25(2), pp.743–755.

[5] Gao, Z., Xuan, H.Z., Zhang, H., Wan, S. & Choo, K.K.R., 2019. Adaptive fusion and category-level dictionary learning model for multiview human action recognition. *IEEE Internet of Things Journal*, 6(6), pp.9280–9293.

[6] Dixon, P.C., 2019. Machine learning algorithms can classify outdoor terrain types during running using accelerometry data. *Gait and Posture*, 74, pp.176–181.

[7] Einicke, G.A., Sabti, H.A., Thiel, D.V. & Fernandez, M., 2018. Maximum-entropy-age selection of features for classifying changes in knee and ankle dynamics during running. *IEEE Journal of Biomedical and Health Informatics*, 22(4), pp.1097–1103.

[8] Buckley, C. et al., 2017. Binary classification of running fatigue using a single inertial measurement unit. *Proceedings of IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks*, pp.197–201.

[9] Khan, N.A., Hussain, S., Spratford, W., Goecke, R., Kotecha, K. and Jamwal, P.K., 2025. Deep learning-driven analysis of a six-bar mechanism for personalized gait rehabilitation. *Journal of Computing and Information Science in Engineering*, *25*(1).

[10] Koul, A., Becchio, C. & Cavallo, A., 2018. Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*. [Online] Available at: https://doi.org/10.3389/fpsyg.2018.01117.

[11] Diss, C.E., 2001. The reliability of kinetic and kinematic variables used to analyse normal running gait. *Gait and Posture*, 14(2), pp.98–103.

[12] Yam, C.Y., Nixon, M.S. & Carter, J.N., 2002. On the relationship of human walking and running: Automatic person identification by gait. *Proceedings of Object Recognition Supported by User Interaction for Service Robots*, pp.287–290.

[13] Khandelwal, S. & Wickström, N., 2017. Evaluation of the performance of accelerometer-based gait event detection algorithms in different real-world scenarios using the MAREA gait database. *Gait and Posture*, 51, pp.84–90.

[14] Espinosa, H.G., Shepherd, J.B., Thiel, D.V. & Worsey, M.T.O., 2019. Anytime, anywhere! Inertial sensors monitor sports performance. *IEEE Potentials*, 38(3), pp.11–16.

[15] Khan, N.A., Goyal, T., Hussain, F., Jamwal, P.K. and Hussain, S., 2024. Transformer-Based Approach for Predicting Transactive Energy in Neurorehabilitation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

[16] Camomilla, V. et al., 2018. Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: A systematic review. *Sensors*, 18(3), p.873.

[17] Shepherd, J.B., Thiel, D.V. & Espinosa, H.G., 2017. Evaluating the use of inertial-magnetic sensors to assess fatigue in boxing during intensive training. *IEEE Sensors Letters*, 1(2), p.6000104.

[18] Thiel, D.V., 2020. Predicting ground reaction forces in sprint running using a shank mounted inertial measurement unit. *Proceedings of MDPI Sensors*.

[19] Lai, A.D.A., James, D.P., Hayes, P. & Harvey, E.C., 2004. Semi-automatic calibration technique using six inertial frames of reference. *Proceedings of SPIE Microelectronics Design, Technology and Packaging*, 5274, pp.531–542.

[20] Madgwick, S.O.H., Harrison, A.J.L. & Vaidyanathan, R., 2011. Estimation of IMU and MARG orientation using a gradient descent algorithm. *Proceedings of IEEE International Conference on Rehabilitation Robotics*, pp.1–7.

[21] Khan, N.A., Sulaiman, M., Tavera Romero, C.A. and Alarfaj, F.K., 2021. Numerical analysis of electrohydrodynamic flow in a circular cylindrical conduit by using neuro evolutionary technique. *Energies*, *14*(22), p.7774.

[22] Worsey, M.T.O. et al., 2020. An evaluation of wearable inertial sensor configuration and supervised machine learning models for automatic punch classification in boxing. *IoT*, 1(2), pp.360–381.

[23] Pedregosa, F. et al., 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, pp.2825–2830.

[24] McKinney, W., 2010. Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, pp.56–61.