Feature Engineering Centric Approach for Knee Torque Estimation

R. Anitha
Scientist
DEBEL, DRDO
anitha.r.debel@gov.in

Vidharshana B
Intern
DEBEL, DRDO
vidharshanabalamurugan@gmail.com

Vikash Kumar
Scientist
DEBEL, DRDO
vikaskumar.debel@gov.in

doi: https://doi.org/10.37285/bsp.SACAD2025.70

***Abstract*—Accurate and real-time estimation of knee torque is crucial for advancing biomechanics, rehabilitation, and the development of assistive and augmentation robotic devices. This study introduces a robust methodology for predicting knee torque by employing advanced feature engineering techniques in conjunction with predictive modelling. Utilizing a comprehensive dataset from 40 subjects under various physiological loads, conducted a meticulous analysis of a wide range of biomechanical parameters. Our approach incorporates Mutual Information for initial feature relevance assessment, Pearson Correlation for redundancy analysis, and Principal Component Analysis for optimal dimensionality reduction and data transformation. A diverse array of regression models, including Gradient Boosted Trees, k-Nearest Neighbors (kNN), Random Forest, and XGBoost, were rigorously trained and evaluated. The Gradient Boosted Regressor and kNN models consistently demonstrated superior performance, achieving low Mean Squared Error (MSE < 0.052) and high scores across multiple optimized feature sets. This study underscores the critical role of effective feature engineering in capturing intricate biomechanical dynamics and significantly enhances the accuracy of knee torque prediction, paving the way for more sophisticated human movement analysis and device control.**

***Index Terms*—Knee torque, feature engineering, predictive modeling, gait analysis, IMU sensors, machine learning, Gradient Boosted Trees, kNN, biomechanics, and human movement.**

I. INTRODUCTION

Accurate estimation of knee joint torque is crucial for comprehending the complex biomechanics of human locomotion. Torque, defined as the rotational force generated around the knee joint, serves as a direct indicator of muscle exertion and joint loading. In comparison to kinematic parameters such as joint angles, torque provides a more comprehensive representation of musculoskeletal performance, particularly in the context of dynamic movement analysis.[1], [2].

Gait laboratory systems equipped with force plates and motion capture technologies are considered the gold standard for acquiring high-quality biomechanical data [3]. These setups facilitate precise measurements of joint kinetics and kinematics, offering reference data necessary for detailed analysis and model development. However, despite their accuracy, such systems are typically restricted to controlled laboratory environments and are not feasible for real-world applications.

Consequently, the development of computational models capable of estimating knee torque using data derived from gait lab measurements represent a valuable advancement toward efficient, scalable, and potentially real-time biomechanical analysis. This study introduces a data-driven framework that utilizes gait laboratory data to train and validate machine-learning models for accurate knee torque prediction. [4], [5]. The proposed approach integrates systematic feature engineering and algorithmic evaluation to identify the most informative biomechanical features and effective predictive models for this task. The primary objectives of this study were as follows:

1) **Feature Engineering:** To systematically extract and optimize meaningful features from a biomechanical dataset using techniques like Mutual Information, Pearson Correlation, and Principal Component Analysis to construct highly discriminative feature sets.
2) **Model Development:** To develop and train a suite of machine learning models capable of estimating knee torque from the engineered features.
3) **Performance Evaluation:** To evaluate the predictive accuracy and generalization capabilities of the developed models using established statistical metrics, identifying the most effective algorithms for this challenging biomechanical prediction task.

By combining the precision of gait laboratory data with the generalization capabilities of machine learning, this study aims to establish a robust framework for knee torque prediction, to support advancements in biomechanics research, augmentative exoskeletons, rehabilitation, and assistive device development.

II. CONTEXTUAL OVERVIEW AND RELATED WORK

A. Gait Analysis and Biomechanical Principles

Gait analysis provides a systematic approach to quantifying human locomotion, encompassing both kinematic and kinetic aspects of movement [6]. It evaluates spatial-temporal parameters, such as stride length, joint angles, and ground reaction forces, to assess walking performance. Kinematics describes motion in terms of joint angles, velocities, and accelerations, while kinetics examines the forces and moments that generate this motion. Through inverse dynamics, joint torques can be derived from measured motion and force data, offering critical insights into the mechanical demands placed on the musculoskeletal system [7].

B. Knee Torque and Gait Cycle Characteristics

Knee torque, defined as the net moment acting about the knee joint, is a crucial indicator of joint loading and muscle activity. A typical knee torque profile over a normalized gait cycle exhibits a pronounced flexion moment during the loading response (10–20%), followed by an extension moment during mid-stance (30–50%), which facilitates forward propulsion [8], [9]. These torque patterns reflect the coordinated contribution of muscle groups and external forces, serving as a benchmark for assessing normal and pathological gait, as well as for optimizing exoskeleton control algorithms [10].

C. Gait Laboratory Datasets and Computational Modeling

Gait laboratories integrate motion capture, force plates, and electromyography (EMG) to obtain high-precision, synchronized biomechanical datasets [11]. The gait laboratory setup used in this study during data collection is shown in Fig. 1. These datasets are crucial for inverse dynamics analysis, facilitating the estimation of joint torques and muscle forces. Due to their accuracy and comprehensiveness, gait laboratory data serve as the reference standard for the development and validation of computational models of human motion. Additionally, they provide the multimodal information necessary for advanced feature engineering in machine learning-based torque prediction tasks [12].

Fig. 1: Data collection at Gait Lab

D. Machine Learning in Biomechanical Prediction

Machine learning techniques have been increasingly employed in biomechanics to model the nonlinear and multidimensional relationships among movement variables [13], [14]. Regression-based approaches, such as Support Vector Machines, Random Forests, Gradient Boosting, and Neural Networks, have been investigated for predicting kinematic and kinetic parameters. However, dynamic knee torque prediction remains a complex challenge due to inter-subject variability and the coupling of muscle forces and joint mechanics. Building upon this foundation, the present work integrates advanced feature engineering with robust ML algorithms to achieve accurate torque estimation from gait laboratory data, contributing to improved biomechanical understanding and exoskeleton control applications.

III. **METHODOLOGY**

The proposed methodology for estimating knee torque adheres to a structured workflow that includes data acquisition, preprocessing, feature engineering, model development, and hyper parameter optimization. Each phase is designed to ensure data integrity, feature relevance, and optimal model generalization. The complete workflow is depicted in Fig. 2.

A. Dataset Description

The dataset utilized in this study was obtained from DEBEL Gait laboratory experiments involving 40 participants, each completing six walking trials under five distinct external load conditions (0 kg, 5 kg, 10 kg, 20 kg, and 30 kg). For each trial, synchronized kinematic and kinetic data were recorded, including knee and hip flexion/extension angles, as well as ground reaction force (GRF) components. This gait laboratory dataset provides a comprehensive biomechanical representation of lower-limb dynamics under varying load conditions. The knee torques of interest—tcmLKFE.M and tcmRKFE.M—were calculated using inverse dynamics and served as the primary prediction targets.

B. Data Preprocessing

The preprocessing phase was designed to ensure consistency, quality, and suitability for machine learning analysis.

1) **Data Concatenation:** Trials from individual subjects under each load condition were concatenated to form a unified dataset, thereby ensuring adequate representation across biomechanical variations.

2) **Missing Data Handling:** Given the continuous and temporally dependent nature of gait data, traditional imputation methods (such as mean or median filling) risk distorting physiological patterns. Consequently, rows containing missing values were removed to preserve temporal integrity and prevent artificial signal artifacts.

3) **Normalization:** Prior to further processing, all continuous variables were normalized using standardization or Min–Max Scaling, contingent upon downstream algorithmic requirements. Scaling ensures numerical stability and prevents features with larger magnitudes from dominating during dimensionality reduction or regression.

The final dataset used for feature engineering consisted of 11514 samples and 46 variables of 40 Subjects No load Condition. Out of the 48 variables, 46 were used as input features describing angular motion (acm), while the remaining two variables — tcmLKFE.M and tcmRKFE.M — served as target torque outputs for the left and right knees, respectively.

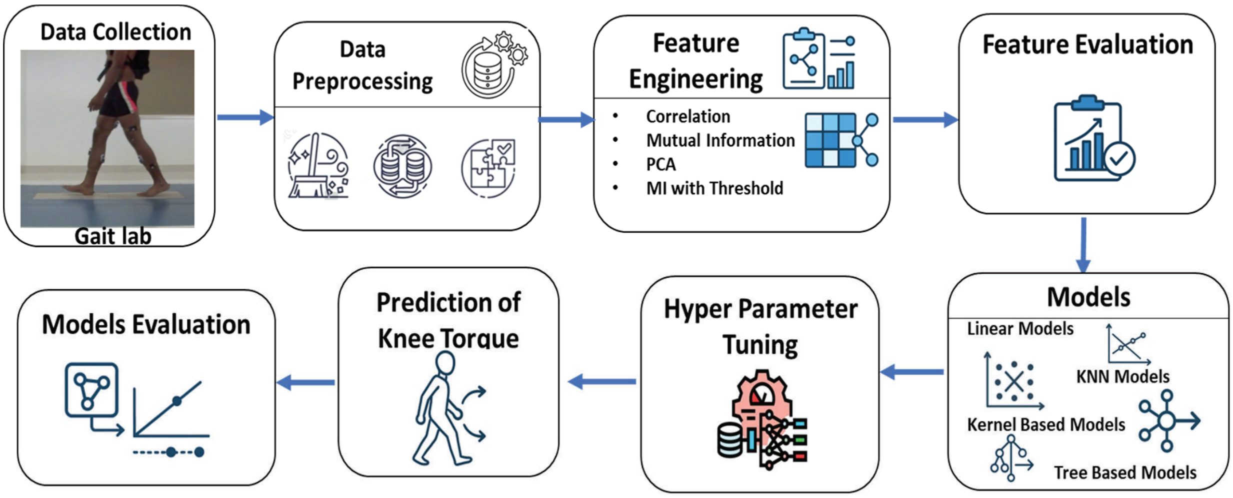

Fig. 2: Flow diagram of the proposed methodology.

C. The Approach Towards Feature Engineering

Feature engineering is pivotal in enhancing the predictive capabilities of machine learning models for knee torque estimation. This process entailed an iterative refinement of feature selection strategies to balance the trade-off between feature relevance, redundancy, and overall model generalization [20]. Following approaches were developed and tested to identify the most effective feature engineering methodology.

Approach *I*: **Dimension Reductionality:** In the initial approach, Pearson correlation coefficients were first computed pairwise among all the input features (excluding the target variables) as illustrated in Fig. 3. Highly correlated feature pairs were identified based on a predefined correlation threshold ($\rho > 0.7$), ensuring independence among predictors.

Subsequently, Mutual Information (MI) based feature selection was employed to quantify the nonlinear relationship between input variables and the target torque outputs (tcmLKFE.M, tcmRKFE.M) as shown in Fig.4, which depicts the ranked MI scores for the top predictive features. Finally, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the selected features. The PCA projection of the transformed features is presented in Fig. 5 ,while Fig. 6 illustrates the explained variance ratio for the principal components, capturing approximately 70% of the total variance. PCA, was effective at minimizing redundancy, transforms the original variables into orthogonal components that lack clear biomechanical interpretability. Since preserving the physical meaning of gait-related features was essential for this study, MI-based thresholding was preferred.

Approach *II*: **Threshold-Driven Mutual Information Selection:** The second approach adopted relied exclusively on Mutual Information (MI) based thresholding as the primary mechanism for feature selection. MI was calculated between each feature and the target knee torque across all load conditions combined as shown in Fig. 7. A series of MI thresholds was established to generate distinct feature subsets. For each threshold level, only features with MI values surpassing the threshold were retained in the corresponding feature set as shown in Fig. 8.

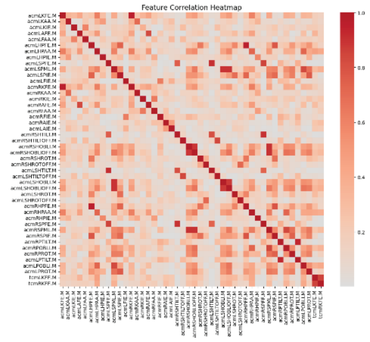

Fig. 3: Feature correlation heatmap showing pairwise Pearson correlation coefficients. Highly correlated pairs ($\rho > 0.7$) were filtered to ensure independence among predictors.

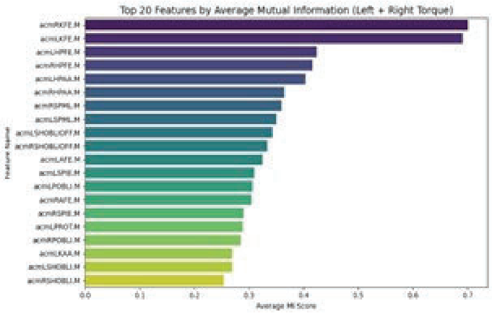

Fig. 4: Mutual information (MI) ranking of input features based on their non-linear relationships with target torque outputs.

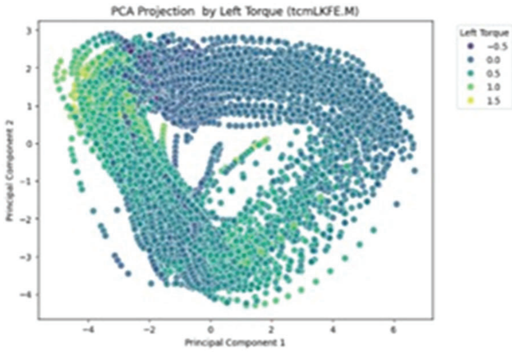

Fig. 5: Principal Component Analysis (PCA) projection of features, illustrating variance distribution across components.

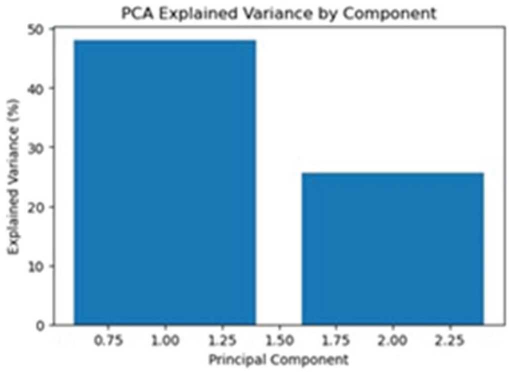

Fig. 6: Explained variance ratio showing cumulative variance captured by the first few principal components (≈70%).

This systematic threshold based filtering produced multiple feature subsets, facilitating a controlled exploration of the impact of feature relevance on model performance. This approach provided a more interpretable and reproducible framework for evaluating feature relevance and yielded consistent improvements in model performance compared to earlier correlation based methods.

D. Model Training and Selection

A wide array of supervised machine learning models was investigated to identify the most suitable algorithm for knee torque prediction [19], [22]. Given that the knee torque is often measured for both the left and right knees simultaneously, the MultiOutput Regressor wrapper from scikit-learn was employed, where appropriate, to handle multiple target variables efficiently. The models considered were as follows:

Linear Models: Linear Regression, RANSAC Regression, Lasso Regression, Ridge Regression. These models provide a foundational baseline for establishing linear relationships.

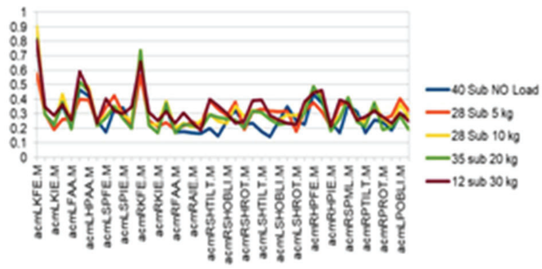

Fig. 7: Mutual Information (MI) values with respect to knee torques (tcmLKFE.M and tcmRKFE.M) for all load conditions combined.

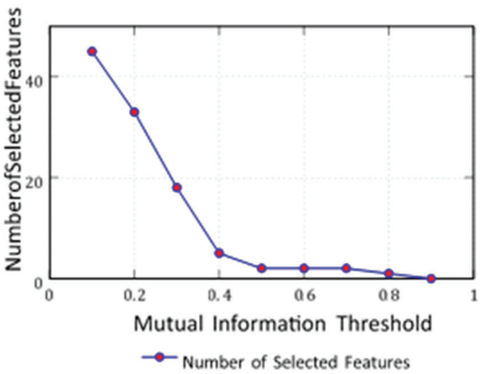

Fig. 8: Feature selection results showing the relationship between the mutual information threshold and the number of selected features using the Elbow Method.

Instance-Based Models: k-nearest neighbors (kNN) regressor, which predicts based on the average of its nearest neighbors.

Kernel Methods: Support Vector Regressor, capable of handling non-linear relationships through kernel tricks.

Tree-Based Models: Decision Tree Regressor, Random Forest Regressor (an ensemble of decision trees), AdaBoost (an ensemble boosting technique), XGBoost (eXtreme Gradient Boosting), and Gradient Boosted Regressor. These ensemble methods are known for their robustness and ability to capture complex, nonlinear patterns.

The dataset was split into training and testing subsets using a standard ratio (70% training, 30% testing) to ensure a rigorous evaluation of the generalization capabilities of the models and to prevent overfitting. [15]–[18].

E. Hyperparameter Tuning

The optimal performance of machine learning models relies heavily on appropriately tuned hyperparameters. GridSearchCV, a robust hyperparameter optimization technique, was systematically applied to each model and to each generated feature set. The key hyperparameters tuned included:

- n_neighbors for kNN (e.g., 3, 5, 7, etc.)
- max_depth and n_estimators for tree-based models (e.g., max_depth up to 10, n_estimators up to 500)
- alpha for regularized linear regressions (e.g., 0.001, 0.01, 0.1, 1.0)
- learning_rate for boosting models (e.g., 0.01, 0.1, 0.2)

This comprehensive tuning process ensured that each model operated at its peak performance for the given feature sets, thereby enabling a fair comparison of the predictive capabilities.

IV. RESULTS

Among the evaluated machine learning models, ensemble and nonlinear methods demonstrated superior predictive performance for knee torque estimation. The Gradient Boosted Regressor versus actual torque plots shown in Fig.9,achieved the highest accuracy with an R^2 of 0.979 and the lowest RMSE (0.052), followed closely by kNN and XGBoost models. In contrast, linear approaches such as Linear, Ridge, and Lasso Regression showed higher error values, indicating limited ability to capture non-linear torque dynamics shown in Fig. 10. Predicted versus actual torque plots further confirmed the effectiveness of ensemble-based models, as their outputs closely aligned with the actual reflecting high prediction reliability and generalization capability [21], [23]. Table I - summarizes the optimal hyperparameters identified and the corresponding MSE for one of the Feature Set. The Gradient Boosted Regressor achieved the lowest MSE, indicating its superior fit to the data for the selected feature set.

V. DISCUSSION

Two complementary feature-engineering approaches were employed to enhance model performance and interpretability. The first involved correlation analysis, where features exhibiting high pairwise correlation (($\rho > 0.7$)) were removed to reduce redundancy and multicollinearity. This ensured that the learning models relied on statistically independent predictors, improving model stability and convergence.

The second approach used Mutual Information (MI)–based ranking with an adaptive threshold to retain only those features exhibiting strong non-linear dependence with the target torque outputs. This thresholding method enabled the selective inclusion of biomechanical variables that contributed most significantly to torque prediction while discarding features with limited informational value. Together, these steps substantially reduced dimensionality and noise, allowing subsequent models to generalize more effectively.

Model evaluation revealed that ensemble and non-linear regressors benefited most from this refined feature space. The Gradient Boosted Regressor achieved the lowest error metrics and highest R^2 values, confirming its superior ability to capture complex feature–torque relationships. Predicted versus actual

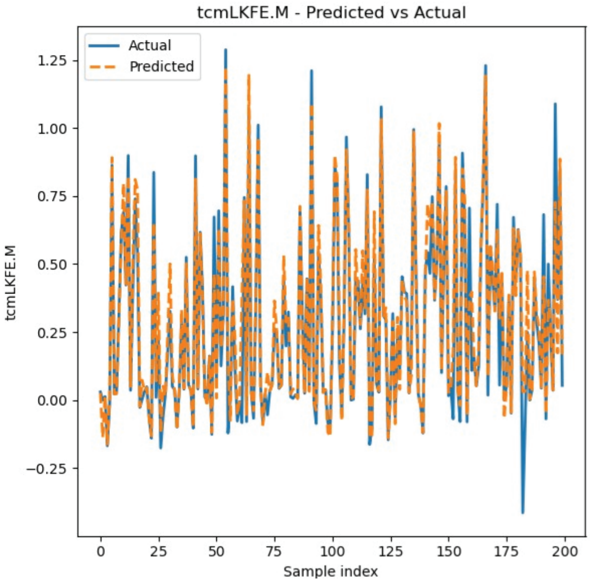

Fig. 9: Predicted Versus Actual Torque (Gradient Boosted Regressor)

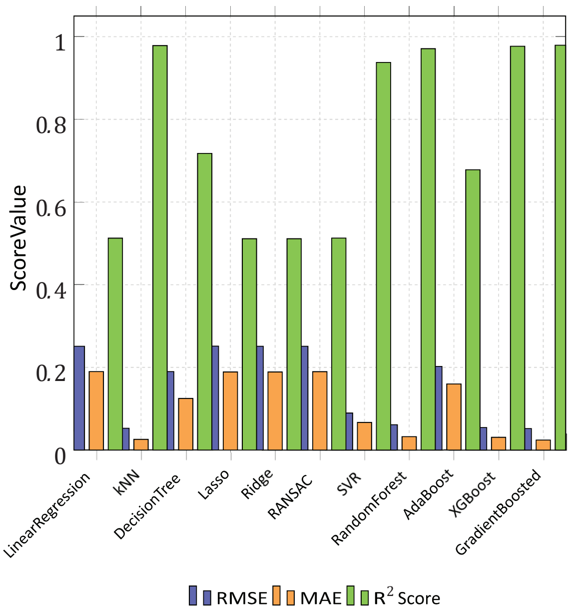

Fig. 10: Model evaluation results for Feature Set across different regression algorithms. Lower RMSE/MAE and higher R^2 indicate better model performance.

torque plots indicating high predictive fidelity and robustness across gait conditions. These outcomes demonstrate that integrating MI thresholding with correlation-based feature selection provides a powerful and efficient pipeline for torque estimation in biomechanical modeling.

TABLE I: Hyperparameter Tuning for Feature Set

Model	Hyperparameters	MSE
Linear Regression	-	0.0631
kNN	n_neighbors = 3	0.0028
Decision Tree Regressor	max_depth=10, max_features = 20, max_leaf_nodes = 40, random_state = 42	0.0360
Lasso regression	alpha = 0.01	0.0633
Ridge regression	alpha = 1.0	0.0631
RANSAC regression	min_samples=2, residual_threshold=3.0	0.0631
Support Vector Regressor	C = 10	0.0080
Random Forest Regressor	n_estimators=300, random_state=42	0.0038
adaboost	learning_rate = 0.1, n_estimators = 100, random_state = 1	0.0410
xgboost	learning_rate = 0.1, n_estimators = 500, max_depth = 10	0.0030
Gradient Boosted Regressor	learning_rate = 0.1, max_depth = 10, n_estimators = 500, random_state = 42	**0.0027**

These outcomes are consistent with recent research emphasizing the role of data-driven learning for biomechanical modeling and exoskeleton torque estimation. Kim and Karatsidis [4], [13] demonstrated that non-linear machine learning models outperform traditional regression methods when predicting joint torques from kinematic and sensor data. Similarly, Koelewijn and van den Bogert [14] reported that machine learning–based torque estimation offers improved accuracy compared to inverse dynamics methods, particularly when high-dimensional gait data are employed. The present study extends these findings by integrating MI thresholding and correlation filtering, achieving comparable or superior prediction accuracy with a reduced and optimized feature set.

Overall, the combination of correlation analysis, MI-based feature selection, and ensemble learning represents a robust and

ISBN: 978-93-47311-06-2

efficient pipeline for knee torque prediction. This approach enhances both model performance and computational efficiency, providing a promising framework for real-time applications in gait analysis, rehabilitation, and adaptive exoskeleton control systems.

VI. CONCLUSION

This study addressed the complex biomechanical challenge of knee torque estimation through the integration of advanced feature engineering and robust machine learning techniques. The proposed framework combined Mutual Information (MI) based feature relevance analysis and Principal Component Analysis (PCA) for dimensionality reduction, resulting in optimized and informative feature sets. Comprehensive model evaluation demonstrated that ensemble and nonlinear regression algorithms, particularly the Gradient Boosted Regressor, delivered superior performance with exceptionally low Mean Squared Error (MSE < 0.052) and high Coefficient of Determination ($R^2 > 0.97$).

VII. FUTURE WORK

Building upon the successful outcomes of this research, future investigations will focus on several key areas:

1) **Inter-Subject Variability Analysis:** A detailed investigation will be conducted across subjects and load conditions (e.g., 5 kg trials) to identify consistent biomechanical patterns and understand individual variations in knee torque dynamics. Insights from this analysis will facilitate the development of more adaptive and personalized predictive models.
2) **Temporal Modeling with Recurrent Architectures:** Future work will explore advanced Recurrent Neural Network (RNN) architectures—specifically Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks—to capture the inherent temporal dependencies in gait and load-bearing data, enabling more accurate and context-aware torque prediction.
3) **Model Optimization and Transfer Learning:** Further refinement through hyperparameter tuning, ensemble stacking, and transfer learning will be investigated to enhance the performance and generalization of the top performing models across diverse datasets and conditions.
4) **Edge Deployment and Real-Time Applications:** The next phase involves adapting the optimized models for real-time implementation on edge computing platforms. Such deployment will enable in-situ torque estimation for wearable robotics, exoskeletons, and continuous rehabilitation monitoring, paving the way for intelligent, responsive, and patient-specific assistance systems.

ACKNOWLEDGMENT

We would like to convey our heartfelt gratitude to the Dr. R Indu Shekhar, Scientist G, Director, DEBEL, and Dr. K Mohanavelu, Scientist G, Project Director – Exoskeleton, for providing the permission, guidance & infrastructure to carry out this research work.

REFERENCES

[1] D. A. Winter, *Biomechanics and Motor Control of Human Movement*, 4th ed., John Wiley & Sons, 2009.

[2] F. E. Zajac, "Muscle and tendon: properties, models, scaling, and application to biomechanics and motor control," *Critical Reviews in Biomedical Engineering*, vol. 17, no. 4, pp. 359–411, 1989.

[3] R. Baker, "Gait analysis methods in rehabilitation," *Journal of NeuroEngineering and Rehabilitation*, vol. 3, no. 1, pp. 1–10, 2006.

[4] S. Kim, J. Lee, H. Park, and Y. Choi, "Human–robot interaction for gait assistance: torque estimation and adaptive control," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, no. 7, pp. 1455–1464, 2021.

[5] K. Moustakas, G. I. Priniotakis, A. G. Papadopoulos, and P. D. Bamidis, "Data-driven biomechanical modeling using machine learning for human motion analysis," *Frontiers in Bioengineering and Biotechnology*, vol. 8, 2020.

[6] M. W. Whittle, *Gait Analysis: An Introduction*, 5th ed., Elsevier, 2014. [7] D. G. E. Robertson, G. E. Caldwell, J. Hamill, G. Kamen, and S. N. Whittlesey, *Research Methods in Biomechanics*, 2nd ed., Human Kinetics, 2013.

[8] A. G. Schache, R. P. Baker, and M. J. Lamontagne, "Knee joint kinetics during walking and running," *Medicine & Science in Sports & Exercise*, vol. 44, no. 4, pp. 653–661, 2012.

[9] J. Perry and J. M. Burnfield, *Gait Analysis: Normal and Pathological Function*, 2nd ed., SLACK Incorporated, 2010.

[10] T. Zhang, P. Fiers, K. A. Witte, R. W. Jackson, K. L. Poggensee, C. G. Atkeson, and S. H. Collins, "Human-in-the-loop optimization of exoskeleton assistance during walking," *Science*, vol. 356, no. 6344, pp. 1280–1284, 2017.

[11] R. B. Davis, S. Ounpuu, D. Tyburski, and J. R. Gage, "A gait analysis data collection and reduction technique," *Human Movement Science*, vol. 10, pp. 575–587, 1991.

[12] D. Kotiadis, J. Hermens, and P. Ketelaar, "Predicting joint kinetics from motion capture using statistical learning," *Journal of Biomechanics*, vol. 58, pp. 165–172, 2017.

[13] A. Karatsidis, G. Bellusci, H. Schepers, M. de Zee, M. Andersen, and P. H. Veltink, "Predicting kinetics and kinematics from wearable sensor data through machine learning," *Journal of Biomechanics*, vol. 95, 2020.

[14] A. D. Koelewijn and A. J. van den Bogert, "Joint torque estimation from kinematic data using machine learning," *Journal of Biomechanics*, vol. 126, 2021.

[15] R. V. Schulte, M. Zondag, J. H. Buurke, and E. C. Prinsen, "Multiday EMG-based knee joint torque estimation using hybrid neuromusculoskeletal modelling and convolutional neural networks," *Frontiers in Robotics and AI*, vol. 9, 2022. DOI: 10.3389/frobt.2022.869476.

[16] S. K. Lam, et al., "Joint torque prediction via hybrid neuromusculoskeletal modelling and statistical ground reaction estimates," *Sensors*, vol. 21, no. 15, 2021. DOI: 10.3390/s21155040.

[17] L. Zhang, X. Zhu, and E. M. Gutierrez-Farewik, "Neuromusculoskeletal model-informed machine learning for knee joint torque prediction with uncertainty," *Frontiers in Neuroscience*, vol. 17, p. 10498472, 2023.

[18] M. Wang, L. Shi, and Y. Zhao, "Lower limb joint torque prediction using long short-term memory networks," *Sensors*, vol. 23, no. 6, 2023. DOI: 10.3390/s23063001.

[19] Z. Wu, H. Huang, and X. Liu, "Deep learning-based estimation of lower limb joint torque using wearable sensors," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 30, pp. 1218–1228, 2022.

[20] P. B. Joseph, S. G. Jayaraman, and D. S. Kumar, "Feature extraction and selection for EMG-based torque prediction using mutual information and deep regression models," *Biomedical Signal Processing and Control*, vol. 74, p. 103493, 2022.

[21] Y. Huang, X. Chen, and Z. Zhang, "Comparative evaluation of machine learning algorithms for gait torque prediction using inertial measurement data," *Sensors*, vol. 22, no. 5, p. 1984, 2022.

[22] R. S. Jamwal, M. Hussain, and M. T. Alam, "Predictive modeling of human joint moments using hybrid machine learning approaches," *Frontiers in Neurorobotics*, vol. 16, pp. 1–12, 2022.

[23] S. K. Gupta, A. Singh, and K. R. Varma, "Real-time joint torque estimation in gait analysis using optimized ensemble regression," *IEEE Access*, vol. 11, pp. 7845–7858, 2023.