Research Article

Classifying Operator Experience from Electric Screwdriving Signals: A BiLSTM-Based Study with External Validation

Kader NİKBAY OYLUM[1*], Turgay Tugay BİLGİN[2]

[1] Trex Dijital Akıllı Üretim Sistemleri A.Ş., https://orcid.org/0000-0002-5218-9218, kaderoylum@trex.com.tr
[2] Bursa Teknik Üniversitesi, https://orcid.org/0000-0002-9245-5728, turgay.bilgin@btu.edu.tr

[*] Correspondence: kaderoylum@trex.com.tr; Tel.: 444 3 468

Received: 11 April 2025
Revised: 05 October 2025
2nd Revised: 12 November 2025
3rd Revised: 15 November 2025
Accepted: 22 November 2025
Published: 26 November 2025

Reference: Nikbay Oylum, K., & Bilgin, T. T. (2025). Classifying operator experience from electric screwdriving signals: A BiLSTM-based study with external validation. The European Journal of Research and Development, 5(1), 201–211.

Abstract

This study presents a deep learning–based approach to objectively classify operator experience levels (Novice–Intermediate–Expert) from multivariate signals and user interactions obtained during electric screwdriving operations. The dataset comprises 64 participant-specific files, each containing multiple tightening trials. Windowing was performed independently per file; short segments unsuitable for windowing were excluded, yielding 3,326 time windows (2,958 for training/testing; 368 for independent validation). A two-layer Bidirectional LSTM (BiLSTM) architecture was employed and evaluated on both the train–test split and an external validation set constructed from 12 previously unseen files. On the test set, the model achieved 76% overall accuracy with macro-averaged precision/recall/F1 of 77%/76%/76%. Class-wise analysis indicated stronger separability for the Expert class (recall ≈ 84%) and comparatively lower performance for Intermediate (recall ≈ 66%). On the hold-out validation set, accuracy was 75.00%, with a mean predicted probability of 85.0%, indicating moderate-to-high confidence. The findings

show that while BiLSTM provides a solid foundation for time-series classification, its effectiveness may be limited for complex patterns without a convolutional front end.

Keywords: BiLSTM, Time-series classification, Multimodal data fusion, Operator experience level classification

1. Introduction

In industrial production, the correct, efficient, and ergonomic use of electric screwdriving tools has a critical impact on product quality, workforce productivity, and occupational health and safety. Nevertheless, on-site assessments often rely on isolated observations or fragmented data, which makes it difficult to interpret operator behavior, process variability, and user experience in a holistic manner.

This study proposes a practical analysis framework for the objective assessment of operator experience levels in electric screwdriving processes. The approach is designed with real shop-floor conditions in mind and aims to make discriminative patterns in user behavior observable. In this way, not only instantaneous performance deviations but also differences related to learning and expertise—as well as opportunities for ergonomic improvement—can be systematically revealed.

The contributions of this work are distinguished by their shop-floor applicability and decision-support value: it renders operator proficiency measurable and reproducible using data captured from routine screwdriving operations; it produces more robust and consistent conclusions by evaluating multiple streams of evidence (process outcomes, behavioral patterns, and user-experience indicators) rather than relying on single metrics; and it translates these insights into directly actionable recommendations for improvement teams by linking them to key performance levers such as quality, cycle time, and ergonomic load.

The remainder of this paper is organized as follows: Section 2 summarizes the relevant literature and positions the study. Section 3 describes materials and methods Section 4 presents the experimental findings and implementation observations. Section 5 concludes and future work with overall remarks and directions for future research.

2. Literature

The literature indicates that Bidirectional Long Short-Term Memory (BiLSTM) models effectively capture discriminative temporal dependencies in time-series problems by simultaneously exploiting past and future context. In this vein, Jungpil Shin et al. propose a hybrid approach tailored to real-time, resource-constrained (edge) settings, wherein spatial and temporal modeling are decoupled: videos are decomposed into frames, frame-level features are extracted using a pretrained CNN/ConvNeXt feature encoder,

and temporal dependencies are modeled with a Temporal Convolutional Network (TCN) (with BiLSTM included for comparison). Evaluations on UCF11, UCF50, UCF101, and JHMDB demonstrate that the ConvNeXt + TCN combination attains higher accuracy with shorter inference time and lower memory consumption than existing methods—achieving approximately ~98% accuracy on the UCF series and >83% on JHMDB. Collectively, these findings suggest that pairing a strong visual feature extractor with an efficient temporal module yields notable advantages in both accuracy and computational efficiency, particularly for practical scenarios such as industrial automation and surveillance[1]. In another study, Guo Huafeng, Xiang Changcheng, and Shiqiang Chen propose a self-attention CNN–BiLSTM architecture for wearable-sensor HAR that integrates sliding-window preprocessing, convolutional feature extraction, bidirectional LSTM for long-range temporal dependencies, and a self-attention module to weight salient time points. Evaluated under subject-wise protocols (50% window overlap) on UCI-HAR, WISDM, PAMAP2, DaLiAc, RealWorld, and USC-HAD, the approach reports consistent accuracy improvements over state-of-the-art baselines, underscoring the advantage of coupling local (CNN) and contextual (BiLSTM) representations with attention in a compact, deployment-oriented model. Taken together with the hybrid, edge-focused design of Shin et al., these results further motivate the use of convolutional front ends plus context-aware sequence modeling as a foundation for robust, real-world activity recognition[2]. A further contribution by Yong Li and Luping Wang presents a sensor-based HAR architecture that pairs a Residual Network (ResNet) front end for local spatial feature extraction with a Bidirectional LSTM (BiLSTM) back end to model forward–backward temporal dependencies. Using lower-limb IMU signals (accelerometer/gyroscope) segmented via sliding windows, the model attains 96.95% accuracy on a self-collected dataset and 97.32% and 97.15% on WISDM and PAMAP2, respectively, while retaining a relatively small parameter budget compared with several CNN/RNN and attention-based baselines. These findings further underscore the value of coupling a convolutional front end with context-aware sequence modeling for robust, resource-efficient activity recognition and provide a methodological bridge to the summaries that follow[3]. In a recent contribution, Pooja Lalwani and R. Ganeshan present a multi-branch CNN–BiLSTM–GRU hybrid for sensor-based HAR that operates on (near) raw smartphone/wearable signals with minimal pre-processing, using sliding-window segmentation (e.g., fixed-length windows with overlap) and parallel convolutional branches to capture local patterns at multiple receptive fields before bidirectional and gated recurrent modeling of long-range temporal dependencies. On WISDM, the model attains up to 99.7% accuracy, with analyses highlighting the role of architectural hyperparameters (e.g., branch widths/batch settings) in convergence and generalization. Taken together with the edge-oriented hybrid pipelines discussed above, these results further support the effectiveness of convolutional front ends coupled with

context-aware sequence models as a robust foundation for real-world activity recognition[4]. In a separate contribution, Amir A. Aljarrah and Ali H. Ali investigate a PCA-augmented BiLSTM pipeline for wearable-sensor HAR on the mHealth dataset (10 subjects, 12 activities), using PCA to reduce dimensionality while preserving most variance before training a BiLSTM RNN; under a 60/20/20 train–validation–test split, they report 97.64% test accuracy, outperforming classical baselines such as kNN, decision trees, Naïve Bayes, and SVM. Taken together with the hybrid, edge-oriented CNN/TCN–BiLSTM literature summarized earlier, these results further substantiate the value of convolutional or statistical front ends coupled with context-aware sequence models as an effective foundation for robust, real-world activity recognition[5]. In another study, Ameer Ali Ridha, Ihab Almaameri, László Blázovics, and Haider Mohammed Abbas introduce a hybrid BiLSTM–SVM framework for smartphone-based HAR on the UCI HAR dataset, where raw IMU signals are preprocessed, segmented into fixed-size blocks, and passed through a BiLSTM to learn sequence representations that are subsequently classified by an SVM; under a 70/30 train–test split, the approach attains 98.74% accuracy—surpassing several CNN/LSTM and traditional baselines reported on the same benchmark—thereby reinforcing the value of context-aware recurrent encoding coupled with a lightweight discriminative classifier for resource-conscious deployments[6]. In a related study, Sai Vyshnavi Modukuri and friends propose a BiLSTM-based HAR pipeline that integrates Linear Discriminant Analysis (LDA) for feature extraction and a univariate filter for feature selection prior to bidirectional sequence modeling, using the UCI HAR dataset (train–test split reported at 70/30). The approach attains 97% accuracy and demonstrates that coupling discriminative feature engineering with context-aware recurrent encoding yields competitive performance under realistic settings. Taken together with the CNN/TCN–BiLSTM literature summarized above, these results further support the use of front-end feature refinement plus bidirectional temporal modeling as a practical baseline for robust activity recognition—providing a smooth methodological bridge to the subsequent summaries[7]. A parallel line of work by Junjie Zhang, Yuanhao Liu, and Hua Yuan proposes an attention-based residual BiLSTM architecture with a 1D-CNN front end (1DCNN-ResBLSTM-Attention) to better discriminate similar activities in wearable-sensor HAR. The model augments bidirectional LSTM with residual connections and layer normalization, and applies an attention mechanism to reweight salient temporal features; extensive tests on UCI-HAR, WISDM, and KU-HAR report overall accuracies of 98.37%, 99.01%, and 97.89%, respectively, alongside gains in stability for look-alike behaviors. These results further underscore the efficacy of combining convolutional front ends with context-aware sequence modeling and attention to improve robustness under practical conditions[8]. Additionally, Pooja Lalwani and R. Ganeshan report a multi-branched CNN–BiLSTM–BiGRU hybrid that operates directly on (near) raw wearable-sensor streams using sliding-window segmentation, multi-scale

convolutional filters for local pattern extraction, and bidirectional recurrent units to capture long-range temporal dependencies; across WISDM, PAMAP2, and UniMiB-SHAR, the best configuration attains ≈99.3% accuracy on WISDM and strong results on the remaining benchmarks, outperforming several CNN/RNN baselines while maintaining competitive efficiency, thereby reinforcing the effectiveness of convolutional front ends coupled with context-aware sequence modeling for sensor-based HAR[9]. Complementing these studies, Sakorn Mekruksavanich, and friends propose a deep multi-task learning framework that jointly tackles activity recognition and user identification from smartphone sensors by sharing a CNN–BiLSTM feature backbone with task-specific heads and a composite loss. On UCI-HAR, the model reports 97.64% accuracy for activity recognition and 82.59% for user identification, outperforming single-task CNN/RNN baselines and underscoring the benefits of shared representations for generalization. Taken together with BiLSTM- and TCN-based pipelines, this evidence highlights a converging design patternthat will frame the subsequent summaries[10].

3. Materials and Methods

The dataset used in this study comprises 64 separate files from electric screwdriving operations; each file corresponds to an individual participant record and contains multiple tightening trials (i.e., multiple tasks/samples) performed by the same person. Participants are represented across three experience levels: Novice (n = 24), Intermediate (n = 21), and Expert (n = 19). Windowing was conducted independently for each file; during the segmentation of multiple trials within a file, short segments not suitable for windowing were excluded from the analysis. This process yielded a total of 3,326 time windows, of which 2,958 were used for training and testing, while 368 were reserved as a separate hold-out set for model validation.

3.1 Bidirectional LSTM Model Implementation

After completion of data augmentation, deep learning experiments were conducted on the expanded dataset with emphasis on the Bidirectional Long Short-Term Memory (BiLSTM) architecture, a widely adopted and empirically validated approach for time-series classification. In contrast to standard LSTM, BiLSTM processes sequences in both forward and backward directions, thereby accessing past and future context simultaneously. This bidirectional modeling is particularly suitable for sensor-based analyses of human behavior and motion patterns. In the present study, 30-second windows of sensor and positional signals were used to capture discriminative temporal dependencies and to improve classification accuracy. Alternative architectures such as GRU and Transformer were not pursued: although GRU is more parameter-efficient, it does not provide the same level of contextual richness as BiLSTM for the stated objectives,

whereas Transformer-based models typically require substantially larger datasets and computational resources than warranted in this setting.

In this study, a two-layer BiLSTM-based architecture was employed to capture temporal patterns. The first BiLSTM layer produces a 128-dimensional output at each time step and, owing to its bidirectional structure, comprises 42,496 trainable parameters; this is followed by batch normalization (512 parameters) and dropout. The second BiLSTM layer contains 41,216 parameters and summarizes the sequence by using only the final time step output; batch normalization (256 parameters) and dropout are subsequently applied to support generalization. A dense layer with 64 neurons (4,160 parameters) and an additional dropout layer follow, and multi-class classification is ultimately performed via a three-unit dense layer with softmax activation (195 parameters). The total number of parameters is 88,835, of which 88,451 are trainable and 384 are non-trainable. For training, the dataset was split into 80% training and 20% testing, with 20% of the training set further reserved for validation. This protocol enabled a reliable assessment of the model's generalization both during training and on the independent test set

Table 1: Bidirectional LSTM Model Results

Class	Precision	Recall	F1-Score	Support
0	0.75	0.79	0.77	245
1	0.74	0.66	0.70	191
2	0.81	0.84	0.83	156

According to the results reported in Table 1, the Bidirectional LSTM model achieved an overall accuracy of 76%, which is a strong outcome for a multi-class time-series classification task. In addition:

Macro Average: Precision, recall, and F1 across classes are 77%, 76%, and 76%, respectively, indicating balanced performance over all classes.

Weighted Average: Accounting for class distribution, precision, recall, and F1 are each 76%.

Class 0 (Novice):

- Recall: 79%, F1: 77%.
- Instances in this class are largely classified correctly.
- Precision (75%) is lower than recall, suggesting a higher rate of false positives for this class.

Class 1 (Intermediate):

- Lowest performance among the classes.
- Recall: 66%, implying that roughly one third of true instances are confused with other classes.
- The confusion appears most pronounced between Class 1 and Class 0.
- F1: 70%, indicating partially balanced yet comparatively weaker performance.

Class 2 (Expert):

- Highest performance: Recall 84%, Precision 81%.
- Patterns in this class are more easily distinguishable and are typically predicted correctly.

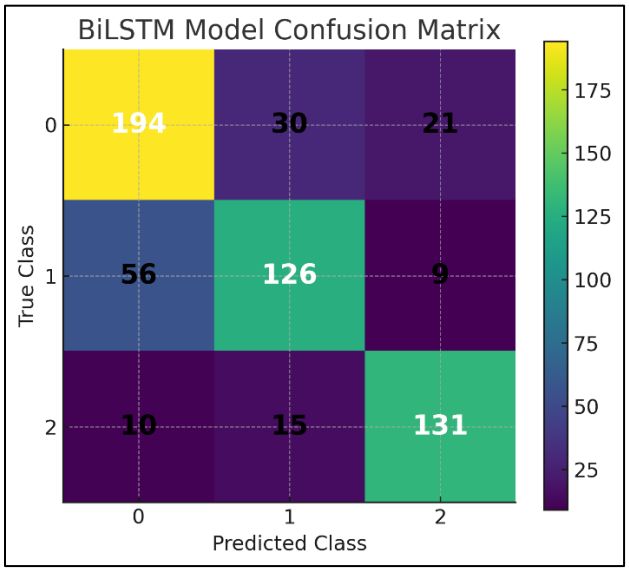

Figure 1 BiLSTM Confusion Matrix

Figure 1 presents the model's classification performance based on the confusion matrix computed on the test set.

Class 0 (Novice): Of 245 instances, 194 were correctly predicted as Class 0, while 30 were misclassified as Class 1 and 21 as Class 2. This pattern suggests similarity between Class 0 and the other classes (particularly Class 1). The recall (per-class accuracy) for this class is approximately 79%.

Class 1 (Intermediate): This is the most challenging class for the model. Of 191 instances, only 126 were correctly classified; 56 were predicted as Class 0 and 9 as Class 2. The error distribution indicates substantial overlap between Class 1 and Class 0, leading to comparatively lower discriminability. The recall for this class is 66%, indicating a clear need for improvement.

Class 2 (Expert): Of 156 instances, 131 were correctly classified, reflecting strong performance. 10 instances were misclassified as Class 0 and 15 as Class 1. These results imply more distinctive patterns for Class 2. The recall for this class is approximately 84%.

Overall, the BiLSTM model attains an overall accuracy of 76% on the test set, indicating satisfactory generalization. The high performance for Class 2 suggests more salient class-specific patterns, whereas the lower performance for Class 1 reflects its overlap with other classes and the model's limited separability for this group.

3.2 Model Validation

An independent validation procedure was conducted to evaluate the generalizability of the BiLSTM deep learning model developed in this study. For this purpose, 12 CSV files—excluded from both training and testing and selected to reflect real-world operating conditions—were utilized. Each file was segmented into fixed-length windows of 600 rows using a windowing procedure, yielding a total of 368 validation windows. The resulting accuracy for the BiLSTM model on this hold-out set was 75.00%.

Table 2 BiLSTM Model Accuracy (Per Class)

Class	BiLSTM Accuracy
Class 0 (Novice)	0.50
Class 1 (Intermediate)	1.00
Class 2 (Expert)	0.75

According to the class-wise accuracy analysis in Table 2, the model exhibits varying performance across experience levels. The BiLSTM attains 100% accuracy for the Intermediate class, 50% for the Novice class, and 75% for the Expert class. These findings suggest that the BiLSTM architecture's sensitivity to specific pattern types varies by class. This suggests that the sensitivity of the BiLSTM architecture to specific pattern types varies across classes.

Table 3 Detailed Prediction Outcomes

File Name	True Class	BiLSTM Prediction	BiLSTM Confidence	BiLSTM Correct (✓/✗)
C24	Novice	Intermediate	0.75	✗
C11	Novice	Novice	0.93	✓
C32	Intermediate	Intermediate	0.81	✓
C20	Expert	Expert	0.95	✓
C22	Novice	Novice	0.93	✓
S24	Novice	Intermediate	0.75	✗
C2	Expert	Expert	0.95	✓
C26	Intermediate	Intermediate	0.98	✓
C21	Expert	Expert	0.79	✓
C5	Intermediate	Intermediate	0.81	✓
S3	Expert	Novice	0.57	✗
C18	Intermediate	Intermediate	0.98	✓

As shown in Table 3, detailed prediction analyses are reported for each validation sample. This evaluation, conducted on 12 independent validation files, reveals both each model's classification correctness and its confidence (predicted probability) in those decisions.
For the BiLSTM, 9 predictions were correct and 3 were incorrect. The mean predicted probability of 85.0% indicates that the model generally issues decisions with relatively high confidence. However, the presence of high confidence even for certain misclassifications suggests potential mislearning of specific classes or insufficient class separability in the learned representations. These observations point to the need for retraining and architectural refinements to improve robustness.

4. Results

A deep learning–based classification system was developed to infer individual experience levels from sensor data and user interactions associated with the use of electric hand tools. The Bidirectional Long Short-Term Memory (BiLSTM) architecture was evaluated, and the model was analyzed comprehensively on both train–test data and an independent validation set. Classification performance was considered holistically—not only in terms of overall accuracy, but also with respect to class-wise performance, prediction confidence (probability scores), and generalizability.
Although model achieved a certain level of success on the training and test splits, more reliable conclusions for real-world deployment were sought through re-evaluation on an independent validation set. In this validation protocol, 12 previously unused data files were segmented into 600-row windows, yielding 368 samples that were used to assess generalization. The BiLSTM attained an accuracy of 75.00% on this hold-out set. This

suggests that the sensitivity of the BiLSTM architecture to specific pattern types varies across classes.

Analyses of detailed prediction outputs considered not only the number of correct classifications but also the models' confidence levels in their decisions. In this context, the BiLSTM demonstrated moderate confidence, with a mean predicted probability of 85.0%. These findings provide important insights into model stability, the capacity to generalize learned patterns, and the prospects for practical application.

5. Conclusion and Future Work

This study demonstrates that operator experience levels can be objectively classified using multivariate signals obtained from electric screwdriving processes. According to the findings, although the BiLSTM model is a strong alternative for time-series analysis, its effectiveness may be limited when not supported by convolutional pre-processing. Accordingly, it is recommended to first capture local patterns with convolutional layers and then model temporal context with BiLSTM (e.g., CNN→BiLSTM), while evaluating performance on an independent validation set. In sum, the BiLSTM-based approach provides a viable foundation; with convolutional pre-processing and appropriate validation protocols, both performance and generalizability can be further strengthened. Future work will focus on;

(i) introducing a convolutional front end (1D-CNN/TCN) before BiLSTM to better capture time–frequency/local patterns,

(ii) calibration of confidence scores and deployment-oriented thresholding to manage application risk,

(iii) class-aware augmentation and domain adaptation/transfer learning to enhance robustness across tools and workstations,

(iv) explainability and error analysis (attention maps, feature importance) to identify weak class boundaries and provide actionable feedback for process improvement.

References

[1] Shin, J., Al, M., Maniruzzaman, M., Nishimura, S., & Alfarhood, S. (2025). Video-Based Human Activity Recognition Using Hybrid Deep Learning Model. Computer Modeling in Engineering & Sciences, 143(3), 3615.

[2] Huafeng, G., Changcheng, X., & Shiqiang, C. (2023). Wearable sensors for human activity recognition based on a self-attention CNN-BiLSTM model. Sensor Review, 43(5/6), 347-358.

[3] Li, Y., & Wang, L. (2022). Human activity recognition based on residual network and BiLSTM. Sensors, 22(2), 635.

[4] Lalwani, P., & Ganeshan, R. (2024). A novel CNN-BiLSTM-GRU hybrid deep learning model for human activity recognition. International Journal of Computational Intelligence Systems, 17(1), 278.

[5] Aljarrah, A. A., & Ali, A. H. (2019, August). Human activity recognition using PCA and BiLSTM recurrent neural networks. In 2019 2nd International Conference on Engineering Technology and its Applications (IICETA) (pp. 156-160). IEEE.

[6] Ridha, A. A., Almaameri, I., Blázovics, L., & Abbas, H. M. (2023, July). Human activity recognition by BiLSTM recurrent neural networks and support vector machine. In 2023 6th International Conference on Engineering Technology and its Applications (IICETA) (pp. 459-465). IEEE.

[7] Modukuri, S. V., Mogaparthi, N., Burri, S., & Kalangi, R. K. (2024, September). Bi-LSTM based real-time human activity recognition from smartphone sensor data. In 2024 International Conference on Artificial Intelligence and Emerging Technology (Global AI Summit) (pp. 474-479). IEEE.

[8] Zhang, J., Liu, Y., & Yuan, H. (2023). Attention-based residual BiLSTM networks for human activity recognition. IEEE Access, 11, 94173-94187.

[9] Lalwani, P., & Ramasamy, G. (2024). Human activity recognition using a multi-branched CNN-BiLSTM-BiGRU model. Applied Soft Computing, 154, 111344.

[10] Mekruksavanich, S., Phaphan, W., & Jitpattanakul, A. (2025). A Deep Multi-Task Learning Network for Activity Recognition and User Identification Using Smartphone Sensors. Procedia Computer Science, 256, 1350-1357.