Dynamic Hand-Signal Recognition and Prediction Framework for Crane Operations Using Deep Learning

Zhihao Wei[1,2], Ghulam Muhammad Ali[2], and Xinming Li[1]

[1]School of Mechanical Engineering, University of Alberta, Canada
[2]SensiImage Technologies Ltd. Canada
wzhihao@ualberta.ca, gmail@sensiimage.com, xinming.li@ualberta.ca

Abstract –

The study proposed a real-time sensor-based hand gesture signal classification framework to address communication challenges between signalmen and crane operators on construction sites. The system collects data from a 3-axis accelerator, 3-axis gyroscope, and three flex sensors installed on gloves to recognize and predict dynamic hand signals. A dedicated data structure is developed to capture temporal characteristics of gestures by combining five consecutive time step sensor data, making sure that the deep learning model can differentiate between complex, sequential gestures. A hybrid CNN-LSTM deep learning model is employed for gesture recognition, and an LSTM-based model is used for predicting sensor data for subsequent signals, helping crane operators improve operation accuracy. The proposed framework successfully classifies 18 predefined crane operation hand signals with 98% overall accuracy, and the prediction process achieves a R^2 0.872. For the real situation test, the framework achieves a correct classification rate of 95.72%. These results demonstrate the effectiveness and robustness of deep learning models with dynamic data structure in addressing communication challenges and improving work safety in crane operations.

Keywords –

Crane operations; Deep learning; Sensor-based systems; Safety in construction.

1 Introduction

Cranes have become indispensable tools on construction sites, especially with the growing adoption of prefabricated components and modular construction techniques. In crane operations, effective communication is vital to ensuring safety and efficiency. Traditionally, hand signals and two-way radios are used for communication, with crane operations being entirely dependent on the signalman's instruction.

However, the complexity of construction sites makes signal communication challenging. The main barriers include low visibility caused by obstructions of large size prefabricated modular components and weather conditions, as well as low clarity of radio communication caused by excessive construction site noise, low signal stability, and diverse workers accents. These difficulties increase the likelihood of miscommunication and the failure to execute instructions correctly. For crane operations, even the slightest error can result in fatal consequences. In fact, miscommunication between crane operator and signalman is a major factor in the high rate of fatalities in the construction industry [1]. According to the US Bureau of Labor Statistics 2017, the average number of fatalities in crane operations is 42 per year in the past 7 years [2]. Furthermore, a study [3] found that 34% of crane-related fatalities are caused by signalman and operator errors, with miscommunication accounting for 11% of these cases.

The alarming statistics highlights the need for a smart system to capture real-time signalman hand-signal to enhance existing communication methods and improve worker safety. In the presented study, we address this gap by proposing a real-time sensor-based hand signal classification framework for crane operations. A detailed deep learning algorithm to improve the classification accuracy of dynamic gesture recognition is developed and validated, while data with comprehensive temporal information, collected from HandCoMM glove set, is used to train the model. Additionally, a prediction process is also implemented for providing crane operators with more reaction time, enabling them to make the correct decisions and operations.

The proposed system can recognize signals, as illustrated in Figure 1. It is worth noting that an additional gesture, "*standby/activate*," is used to initialize the system proposed in this study. This gesture is performed with the forearm vertical and the index and middle fingers pointing upward. However, it is not shown in Figure 1.

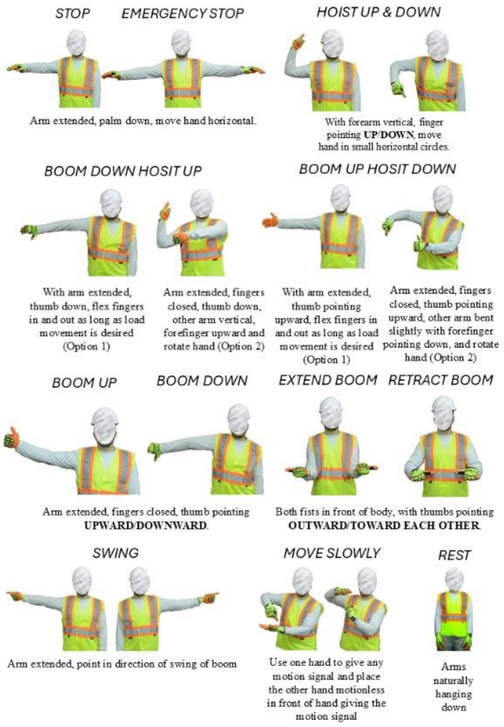


Figure 1. Hand signal for hoist and crane operations

2 Literature Review

2.1 Internet of Things (IoT) in crane operation

Internet of Things technology has been developed, tested, and used in crane operation in construction sites. For instance, IoT technologies such as RFID [4] and GPS [5] positioning are used to integrate crane resources, which can reduce latency and improve efficiency in construction sites. Automation in crane operations is also a hotpot of IoT applications, X. Li et al. [6] used laser scanner and 4G communication to realize automatic crane operation for a prefabricated part. With AI-decision implementation, an automatic crane operation model has been achieved in a digital simulation model [7].

Sensor-based wearable devices represent another promising development direction. By integrating sensors into workers' personal protective equipment (PPE) rather than mounting them on crane arms or hooks, IoT systems can enable faster responses and provide more accurate hazard warnings [8], signal transmission [9], and other functionalities.

2.2 Deep Learning (DL) in crane operation

In crane operations, computer vision is one of the most widely applied deep learning technologies. Plenty of DL models, like YOLO (You Only Look Once), CNN (Convolutional Neural Network), are applied to achieve automation and intelligentization. CNNs are particularly effective in computer vision tasks, as they excel at capturing spatial features and local dependencies within image data. Some gesture recognition tasks rely on CNN-based architectures [15], while others use CNNs to extract combined feature representations from sensor data [14]. In addition, Long Short-term Memory (LSTM) model is crucial technology for processing time-series data. Many LSTM-based models are employed to achieve specific functions in construction sites. Wang et al. [10] used an optimized LSTM to achieve real-time and precise locating of the crane hook. LSTM is also applied to monitor and predict crane movement path [11], ensuring efficient crane operation planning and real-time adjustment.

2.3 Hand gesture recognition

With the widespread development of IoT technology and deep learning, more advanced systems for hand gesture recognition have further integrated to enhance the safety and efficiency of crane operations. Among these studies, vision-based sensors are the one of the most widely applied, with many efforts dedicated to combining visual signal with deep learning networks to achieve efficient hand signal recognition. Wang et al. [12] conducted a comparative study on different DL architectures, including ResNeXt-101 and Res3D + ConvLSTM + MobileNet, to evaluate their performance in image-based hand gesture recognition tasks in construction site. In addition to visual sensors, IMU sensors have been widely used to capture dynamic hand motion information. Mansoor et al. [1] integrated IMU sensors with traditional machine-learning algorithms and LSTM-based deep learning models for crane hand signal classification. Their study successfully demonstrated the feasibility of using wearable glove-based sensor layouts for reliable gesture recognition. Many other types of sensors have been selected to adopt to gesture recognition on under different conditions. EMG [13] helps to reduce restrictions on finger movement, flex sensors [9] improve the recognition of complex finger movements. LiDAR [14] minimizes the influence of lighting condition. Another effective approach is multimodal hand gesture recognition, by combining eye-tracking and IMU sensor [15], and adding context-aware sensing [16], researchers achieved more efficient the hand signal communication.

Building on these prior studies, this paper optimizes the use of accelerometers, gyroscopes, and flex sensors, while introducing a custom time-series data structure to better capture the temporal dynamics of hand gestures. Furthermore, we develop a CNN-LSTM hybrid model for precise gesture classification and incorporate an LSTM-based prediction module to anticipate future hand signals. This predictive capability extends the crane operator's reaction time, enabling more accurate and timely execution of commands, ultimately enhancing operational safety and efficiency.

3 Research Methodology

The methodology for this study has been structured into three main parts: system design, data structure and preprocess, and deep learning model and implementation.

The overall workflow in the study is illustrated in Figure 2:

Step 1: Data is collected by sensors installed in the glove and transmitted to the central device.

Step 2: The data undergoes auto labeling, preprocessing and data structure management.

Step 3: The hybrid CNN-LSTM recognition model and LSTM prediction model are trained and tuned using processed data.

Step 4: The trained and fine-tuned models are deployed on the central device to enable real-time gesture signal recognition and prediction.

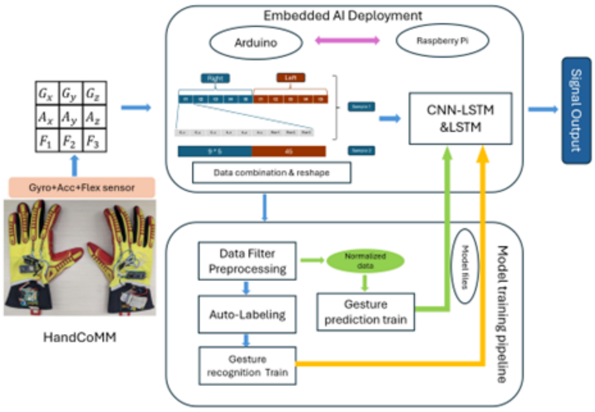


Figure 2. Dynamic crane signal recognition and prediction workflow.

3.1 System Design

The system consists of two main hardware components: the HandCoMM glove set, and a central device mounted on a belt. The glove set is responsible for collecting gesture signal data, while the central device handles data preprocessing for deep learning model and serves as the host for the deployed deep learning model using the preloaded weighted model file.

As shown in Figure 3, three flex sensors are strategically positioned on the thumb, index finger, and middle finger to precisely detect finger bending. As the ring and pinky fingers tend to move together with the middle finger, their bending pattens can be inferred from the flex sensor on the middle finger. This simplification not only reduces sensor complexity while maintaining high accuracy in gesture recognition but also lowers the communication burden and minimizes transmission delay. An Arduino development board, integrated with an accelerometer and gyroscope, is securely mounted on the back of the glove near the wrist. This configuration enables the system to monitor a wide range of hand movements, including hand pitch, wrist rotation, arm rotation, and the overall direction, speed, and acceleration of the hand. The battery and related accessories are securely fixed at the base of the glove.

The central device comprises an Arduino development board, a Raspberry Pi, and related supporting accessories etc. battery and power management unit. The gloves communicate with the central device via the Bluetooth protocol, while the Raspberry Pi and Arduino exchange data through a serial connection. These components are integrated and mounted on a belt to ensure the signalman's comfort during wear.

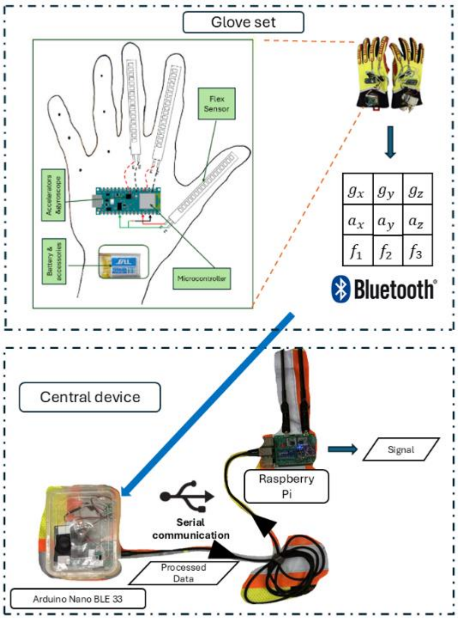


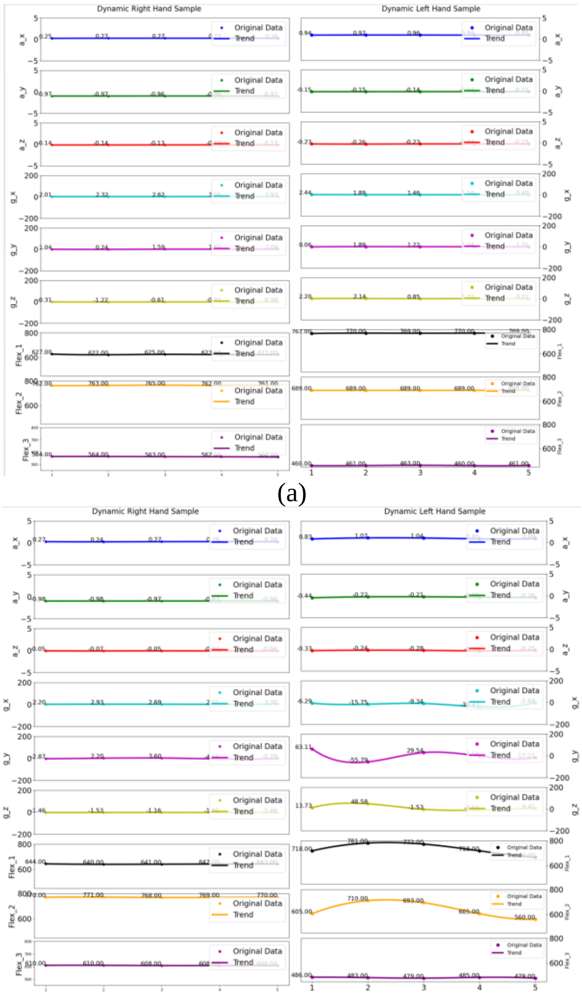
Figure 3. Layout of HandCoMM glove [17], [18] and system setup.

In this study, all devices are worn by the signalman. The gloves collect data from the accelerometer, gyroscope, and flex sensors, which are then transmitted via Bluetooth 5.0 Low Energy (BLE) to an Arduino development board for integration and preprocessing. This data transmission method ensures real-time performance and stability over short distances. The maximum Bluetooth packet size per transmission is 247 bytes, while the study's data size is approximately 200 bytes per packet, resulting in a transmission delay of just 1-2ms, depending on the communication mode. With a continuous communication interval of 7.5ms, this delay is negligible compared to the 4-6Hz frequency of hand movements (150ms-250ms per motion cycle), ensuring smooth and uninterrupted communication. Although Bluetooth communication has a typical range of 10 meters, the actual distance between the gloves and the central device during operation is less than 1 meter, further enhancing the stability and reliability of the connection. The prepared data is subsequently sent to the central device (Raspberry Pi) via serial communication for deep learning inference, where gesture signals are generated and transmitted.

3.2 Data Structure and Preprocess

The dataset is designed to capture the temporal dynamics of hand gestures, essential for distinguishing complex dynamic gestures. Unlike static gestures, dynamic gesture signals require temporal context to represent transitions and movements. For example, the "*Boom up*" signal involves making a fist with the thumb raised, while "*Boom up hoist down (Option 1)*" requires sequentially bending four fingers while maintaining the thumb's position. Since these gestures share certain hand positions, single time-point data cannot differentiate them effectively. To address this, each sample integrates data from five consecutive time points, encoding the gestures' temporal characteristics. In Figure 4, the structural differences in sensor data between the two types of hand gesture signals are illustrated.

Each glove records nine types of data per time point: gyroscope readings (G_x, G_y, G_z), accelerometer readings (A_x, A_y, A_z), and flex sensor readings from the index finger, middle finger, and thumb (flex_1, flex_2, flex_3). Combining five consecutive time points results in a 45-dimensional feature vector per glove, and merging data from both gloves produces a 90-dimensional dataset per sample. With a sampling frequency of 30 Hz, the system captures samples at 6 Hz after combining five time points, ensuring efficient data processing and remaining gesture nuances. The selection of five-time steps in this study is determined by multiple factors. Research on human motion sensor sampling frequencies has shown that a sampling rate between 20Hz and 40Hz [19] provides a more accurate representation of human movements. Based on this, a baseline sampling frequency of 30Hz was adopted in this study. Additionally, studies on hand movement frequencies indicate a range of 1Hz to 4.5Hz [20], [21]. Considering the time required for data transmission, the effective sampling frequency was set to 6Hz, leading to the choice of five consecutive time steps. Furthermore, selecting a longer time window would increase communication time, as there are limitations on the data packet size per transmission. Exceeding this limit would require data segmentation, significantly increasing communication latency. As shown in Figure 4(b), five-time steps sufficiently capture the transition from finger bending to extension, including the two noticeable gyroscope fluctuations caused by the full bending and straightening of the fingers, as well as the rise and return of flex sensor values for the index and middle fingers.



(a)

(b)

Figure 4. Dynamic feature of sensor data comparison. (a) Gesture "*Boom up*". (b) Gesture "*Boom up Hoist down (Option 1)*"

For prediction tasks, due to the significant differences in absolute values among different sensor data, additional scaling is required to enhance the accuracy of the model's predictions. The gyroscope and accelerometer data possess directional properties, meaning their values can be both positive and negative; therefore, they are normalized within the range of (-1, 1). In contrast, the flex sensor measures resistance values, which are always positive, so it is normalized within (0, 1). The normalization process was conducted using the following formula:

$$x' = a + \frac{(x - x_{min})}{(x_{max} - x_{min})} \cdot (b - a)$$

Where x represents the original sensor value, x_{min} and x_{max} are the minimum and maximum values of the data, a and b are the lower and upper bounds of the target range, x' is the normalized value. Then generated data by the model could be reversed to its true value for analysis and practical applications.

Each dataset sample is labeled with one of 18 predefined crane signaling gestures. Stringent quality control measures were applied to ensure the data's integrity, removing incomplete or corrupted samples. The labeled data is stored in CSV format for accessibility and structured analysis.

The dataset includes contributions from six volunteers, reflecting diverse crane signaling experiences. This diversity ensures the dataset represents realistic variations in hand gestures. The final dataset comprises 13,514 samples, providing a robust foundation for developing deep learning models capable of recognizing dynamic gestures accurately in real-world scenarios.

3.3 Deep learning Model and Implementation

In the present study, deep learning is employed to achieve two critical objectives: gesture recognition and gesture prediction. Based on the data from flex sensors and IMU mounted on gloves, deep learning model will extract the features and time sequence characteristics to recognize different hand signal and generate sensor data at time step t+1. The section provides the details of deep learning models and implementation used in the proposed paper.

3.3.1 Model architecture and implementation: gesture recognition

To achieve robust gesture recognition, a combined CNN-LSTM model is employed. This architecture takes advantage of the strengths of CNNs in spatial feature extraction and LSTM in modeling temporal dependencies. Below, the model components and their functionalities are outlined:

- 1D-CNN Module

The 1D-CNN layers are designed to extract spatial features from the input sensor data, capturing correlations among the mounted sensors at each time step. This means CNN can identify static patterns. For example, if G_x and G_y increase together, it may signal a clockwise rotation. Through convolutional and pooling operations, the module reduces the dimensionality of the data while retaining its essential features. The first convolutional layer employs 64 filters with a kernel size of 3, followed by batch normalization to stabilize and accelerate training, max-pooling for dimensionality reduction, and dropout for regularization. A second convolutional layer, consisting of 128 filters, further enhances the model's ability to extract complex spatial features.

- LSTM Module

The LSTM module processes the sequence of features extracted by the CNN module to capture temporal dependencies and dynamic patterns in the sensor data. A single LSTM layer with 100 units is implemented, allowing the model to learn both short-term and long-term dependencies across time steps. To mitigate overfitting and improve generalization, a dropout rate of 0.5 is applied.

- Implementation

A Global Max Pooling layer was used to retain the most critical temporal features, followed by a dense layer with 256 units and ReLU activation to capture high-level representations. The final output layer employed SoftMax activation to classify the data into 18 predefined gesture categories.

The model was trained using the Adam optimizer with a learning rate of 1×10^{-3} and categorical cross-entropy loss. An early stopping mechanism with a patience of 10 epochs was employed to prevent overfitting, and the best weights were restored after training. The dataset was split into 80% training and 20% testing subsets, with a batch size of 64 and a maximum of 300 epochs.

3.3.2 Model architecture and implementation: Gesture Prediction

The predictive model employs an LSTM-based architecture designed to forecast sensor data for the next time step. The model consists of a single LSTM layer with 50 units, followed by a dense layer to generate predictions. Optimized using the Adam optimizer with a learning rate of 1×10^{-5}, the model minimizes Mean Squared Error (MSE) during training. Early stopping and model checkpointing are implemented to prevent overfitting and ensure optimal performance, with the best model weights saved for future use.

The hyperparameters of both models are summarized in Table 1, including architecture, training, and regularization settings selected for optimal performance.

Table 1. Summary of parameters of model structure and training details for task recognition and task prediction.

Aspect	Gesture Recognition	Gesture Prediction
Model type	CNN-LSTM hybrid	LSTM
LSTM Units	100	50

LSTM layers	1 LSTM Layer	1 LSTM Layer
Dropout Rate	0.5	Not Applicable
Optimizer	Adam	Adam
Learning Rate	1e-3	1e-5
Loss Function	Categorial Cross-Entropy	Mean Squared Error (MSE)
Activation Function	ReLU (Dense Layer), SoftMax (Output Layer)	Linear (Output Layer)
Batch Size	64	32
Epochs	Maximum of 300	Maximum of 20000
Early stopping Patience	10	10
Metrics	Accuracy	MSE
Output	Gesture Classification (18 classes)	Predicted Sensor Data

4 Result

The results of this study demonstrate the effectiveness of the proposed deep learning models and dedicatedly designed data structure for gesture recognition and prediction.

4.1 Gesture recognition

The CNN-LSTM model achieved a high accuracy of 98% on the testing dataset, which was derived from a split of 20% (2703 samples) of the entire dataset, in classifying 18 predefined crane signal gestures.

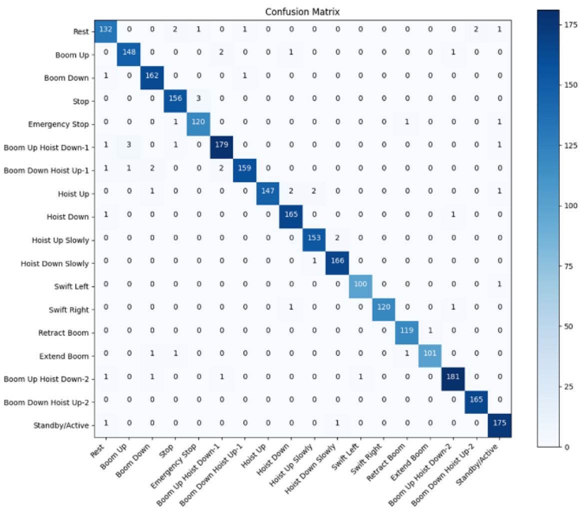


Figure 5. Confusion matrix of CNN-LSTM model for task gesture recognition.

In real situation test, a volunteer was instructed to wear the gloves and perform all 18 predefined gestures. The model successfully detected the correct signals with an accuracy of 95.72% (179 out of 187 samples), demonstrating its strong generalization ability and effectiveness in recognizing both dynamic and static gesture signals. In Figure 5, the confusion matrix provides a comprehensive view of the model's classification performance across all gestures. Figure 6 displays the detailed classification report. In the report, precision measures the proportion of correctly predicted positive observations out of all predicted positive observations, recall (or sensitivity) measures the proportion of correctly predicted positive observations out of all actual positive observations, F1-score is the harmonic mean of precision and recall, and support represents the total number of true instances for each class.

Classification Report:

	precision	recall	f1-score	support
Rest	0.96	0.95	0.95	139
Boom Up	0.97	0.97	0.97	152
Boom Down	0.97	0.99	0.98	164
Stop	0.97	0.98	0.97	159
Emergency Stop	0.97	0.98	0.97	123
Boom Up Hoist Down-1	0.97	0.97	0.97	185
Boom Down Hoist Up-1	0.99	0.96	0.98	165
Hoist Up	1.00	0.96	0.98	153
Hoist Down	0.98	0.99	0.98	167
Hoist Up Slowly	0.98	0.99	0.98	155
Hoist Down Slowly	0.98	0.99	0.99	167
Swift Left	0.99	0.99	0.99	101
Swift Right	1.00	0.98	0.99	122
Retract Boom	0.98	0.99	0.99	120
Extend Boom	0.99	0.97	0.98	104
Boom Up Hoist Down-2	0.98	0.98	0.98	185
Boom Down Hoist Up-2	0.99	1.00	0.99	165
Standby/Active	0.97	0.99	0.98	177
accuracy			0.98	2703
macro avg	0.98	0.98	0.98	2703
weighted avg	0.98	0.98	0.98	2703

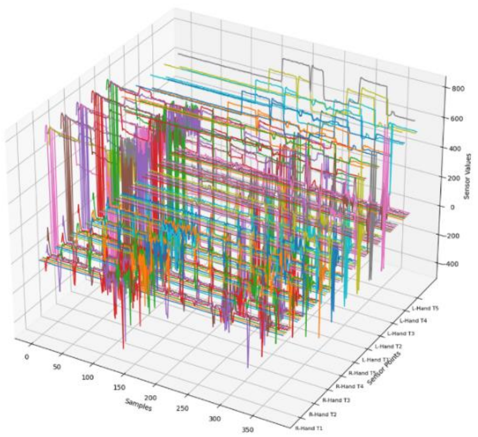
Figure 6. Classification report for gesture recognition

4.2 Gesture prediction

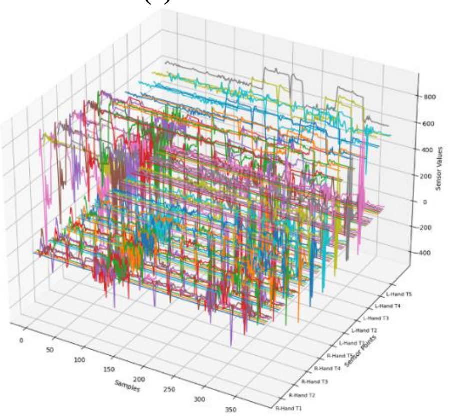
The LSTM model effectively predicted sensor data with an overall R^2 of 0.872, demonstrating reliable accuracy. Here, R^2 measures the proportion of variance in the data explained by the model. With values closer to 1 indicating better predictions. The mean squared error (MSE) quantifies the average squared difference between predicted and true values. The root mean squared error (RMSE) represents the average prediction error in the same unit as the data.

Table 2. Summary of Sensor Prediction Performance

	R^2	MSE	RMSE	Scope
Accelerator	0.924	0.030	0.172	-4.0 - 4.0
Gyroscope	0.789	759.019	27.550	-481.81 - 618.84
Flex sensor	0.903	2557.124	50.568	0 - 843



(a) True sensor values



(b) Predicted sensor values

Figure 7. Comparison between the true values and predicted values

For individual sensors, the accelerator achieved R^2 = 0.924, MSE = 0.030, and RMSE = 0.172 within a range of -4.0 to 4.0, indicating precise predictions. The gyroscope data showed R^2 = 0.789, MSE = 759.019, and RMSE = 27.550 over a broader range of -481.81 to 618.84. The flex sensor achieved R^2 = 0.903, MSE = 2557.124, and RMSE = 50.568, spanning 0.0 to 843.0. The detailed prediction performance indicators are shown in table 2. These results confirm the model's robust performance across diverse sensor types, reliably supporting crane operation safety and decision-making.

From the overall perspective, the prediction results essentially reflect the data characteristics, including the temporal dynamics and distribution patterns of the real data. The comparison between the true value and the predicted value is shown in Figure 7, (a) is the true value, (b) is the predicted value.

The predicted value will be reinput the CNN-LSTM model, the gesture recognition model, to get a predicted signal, so that the operator has an expectation of the next operation and more reaction time to perform correctly.

5 Conclusion and Future work

In the proposed system, the gloves, equipped with multiple sensors (accelerator, gyroscope, and flex sensors), are used as both data collection devices and personal protective equipment (PPE). A dedicated dynamic data structure, combining five consecutive time steps, ensures the preservation of temporal features in signal recognition. This is particularly important for differentiating dynamics signals like "*Boom up*" and "*Boom up hoist down (Option 1)*", where sequential motion is essential for correct classification.

The hybrid CNN-LSTM deep learning model successfully recognizes 18 predefined signals with 98% accuracy on test dataset and 95.72% accuracy in real situation test. Considering the high accuracy, the model performance is effective and robust. In addition, an independent LSTM model is employed for forecasting sensor data for the next time step. The predicted sensor data will be re-input the recognition model to provide crane operators with the oncoming signals. This enables operators to have an expectation for the oncoming instructions, which can help them have longer reaction time and increase the possibility of making correct operations. In turn, operation losses caused by miscommunication and accidents will be minimized.

Future work will focus on continuous data collection and online learning, allowing the system to adapt new data from different users and continuous improvement through real-world usage; enhancing the signal transmission process, which will ensure that correct signals are reliably passed between signalmen and crane operators; improving hardware reliability and standardization, which will prevent equipment damage during use and ensure consistent system performance. Finally, integrating the system with crane control systems, which will enable semi-autonomous or fully automated crane operations, further improving efficiency and safety.

Acknowledgement

This work was supported by China Scholarship Council (CSC), the author sincerely thanks CSC for their financial support.

References

[1] Q. Fang, H. Li, X. Luo, L. Ding, H. Luo, and C. Li. Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment. *Autom. Constr.*, 93:148–164, 2018, doi: 10.1016/j.autcon.2018.05.022.

[2] Crane-related work deaths trended down from 1992 to 2017: The Economics Daily: U.S. Bureau of Labor Statistics, Bureau of Labor Statistics.

2024. On-line: https://www.bls.gov/opub/ted/2019/crane-related-work-deaths-trended-down-from-1992-to-2017.htm, Accessed: 17/12/2024.

[3] G. Raviv and A. Shapira. Systematic approach to crane-related near-miss analysis in the construction industry. *Int. J. Constr. Manag.*, 18(4): 310–320, 2018, doi: 10.1080/15623599.2017.1382067.

[4] P. R. Zekavat, S. Moon, and L. E. Bernold. Performance of short and long range wireless communication technologies in construction. *Autom. Constr.*, 47: 50–61, 2014, doi: 10.1016/j.autcon.2014.07.008.

[5] B. Gao, Z. Zhu, J. Zhao, and B. Huang. A Wireless Swing Angle Measurement Scheme Using Attitude Heading Reference System Sensing Units Based on Microelectromechanical Devices. *SENSORS*, 14(12): 22595–22612, 2014, doi: 10.3390/s141222595.

[6] X. Li, G. Gao, J. Na, and X. Chen. Design of an Automatic T-beam Erection System Based on NB-IoT for Bridge-Erecting Cranes. In *PROCEEDINGS OF THE 33RD CHINESE CONTROL AND DECISION CONFERENCE (CCDC 2021)*, in Chinese Control and Decision Conference. New York: IEEE, 2021, 684–689. doi: 10.1109/CCDC52312.2021.9602504.

[7] F. B. Ter Haar, F. Ruis, and B. T. van Manen. A 3D World Interpreter System for Safe Autonomous Crane Operation. *Robotics*, 13(2): 23, 2024, doi: 10.3390/robotics13020023.

[8] N. Singh, V. Gunjan, G. Chaudhary, R. Kaluri, N. Victor, and K. Lakshmanna. IoT enabled HELMET to safeguard the health of mine workers. *Comput. Commun.*, 193: 1–9, 2022, doi: 10.1016/j.comcom.2022.06.032.

[9] A. Mansoor, S. Liu, A. Bouferguene, and M. Al-Hussein. Crane Signalman Hand-Signal Classification Framework Using Sensor-Based Smart Construction Glove and Machine-Learning Algorithms. *J. Constr. Eng. Manag.*, 150(8): 04024094, 2024, doi: 10.1061/JCEMD4.COENG-14458.

[10] L. Wang *et al.* Snake optimizer LSTM-based UWB positioning method for unmanned crane. *PLOS ONE*, 18(11): e0293618, 2023, doi: 10.1371/journal.pone.0293618.

[11] S.-T. Liong, F.-W. Kuo, Y. S. Gan, Y.-T. Sheng, and S.-Y. Wang. Predicting trajectory of crane-lifted load using LSTM network: A comparative study of simulated and real-world scenarios. *Expert Syst. Appl.*, 228: 120215, 2023, doi: 10.1016/j.eswa.2023.120215.

[12] X. Wang and Z. Zhu. Vision-based hand signal recognition in construction: A feasibility study. *Autom. Constr.*, 125: 103625, 2021, doi: 10.1016/j.autcon.2021.103625.

[13] A. A. Neacsu, G. Cioroiu, A. Radoi, and C. Burileanu. Automatic EMG-based Hand Gesture Recognition System using Time-Domain Descriptors and Fully-Connected Neural Networks. In *2019 42nd International Conference on Telecommunications and Signal Processing (TSP)*, Jul. 2019, 232–235. doi: 10.1109/TSP.2019.8768831.

[14] Y. Hu and L. Wang. Hand gesture recognition based on convolutional neural networks. In *LIDAR IMAGING DETECTION AND TARGET RECOGNITION 2017*, D. Lv, Y. Lv, and W. Bao, Eds., Bellingham: Spie-Int Soc Optical Engineering, 2017, UNSP 106051S. doi: 10.1117/12.2291737.

[15] X. Wang, D. Veeramani, F. Dai, and Z. Zhu. Eye Gaze and Hand Gesture-Driven Human-Robot Interaction in Construction. 991–1000, 2024, doi: 10.1061/9780784485262.101.

[16] X. Wang, D. Veeramani, F. Dai, and Z. Zhu. Context-aware hand gesture interaction for human–robot collaboration in construction. *Comput.-Aided Civ. Infrastruct. Eng.*, 39(22): 3489–3504, 2024, doi: 10.1111/mice.13202.

[17] Nano 33 BLE | Arduino Documentation. On-line: https://docs.arduino.cc/hardware/nano-33-ble/, Accessed: 24/09/2024.

[18] sea jump 4pcs 3.7v 150mAH Lipo Battery with 4in1 Charger for H36 E010 Furibee f36 REALACC H36 Drone, Battery Packs - Amazon Canada. On-line: https://www.amazon.ca/jump-Battery-Charger-Eachine-Furibee/dp/B075642WRF, Accessed: 24/12/2024.

[19] H. Junker, P. Lukowicz, and G. Troster. Sampling frequency, signal resolution and the accuracy of wearable context recognition systems. In *Eighth International Symposium on Wearable Computers*, Oct. 2004, 176–177. doi: 10.1109/ISWC.2004.38.

[20] Y. Xiong and F. Quek. Hand Motion Gesture Frequency Properties and Multimodal Discourse Analysis. *Int. J. Comput. Vis.*, 69(3): 353–371, 2006, doi: 10.1007/s11263-006-8112-5.

[21] H. L. Cornman, J. Stenum, and R. T. Roemmich. Video-based quantification of human movement frequency using pose estimation: A pilot study. *PLOS ONE*, 16(12): e0261450, 2021, doi: 10.1371/journal.pone.0261450.