

Multimodal Hand Signal and Speech Communication Classification Framework for the
Construction Industry: The Case of Communication between Crane Signalman and Crane
Operator.

by

Asif Mansoor

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Construction Engineering and Management

Department of Civil and Environmental Engineering

University of Alberta

© Asif Mansoor, 2023

ABSTRACT

On construction sites, communication is key to boosting teamwork and improving worker performance. It allows workers to coordinate their activities, share information, and respond to potential hazards quickly and efficiently. In current practice, workers on construction sites typically rely on verbal, hand signalling, and two-way radio communication systems. Verbal and hand signalling communication are used when the workers are in close proximity to one another, whereas two-way radio communication is used when the distance is long and verbal and hand signalling communication are not possible. However, construction sites today are increasingly busy, congested, and noisy, making traditional means of communication less effective. This, in turn, results in communication errors, which can lead to disastrous accidents on construction sites. Meanwhile, the introduction in recent years of technologies such as deep learning and sensor-based approaches has resulted in a number of applications to improve safety, productivity, and surveillance. In this regard, the present research proposes a multimodal construction site communication classification system that uses innovative technologies to improve the reliability of communication on construction sites. In this research, communication between crane operator and signaller on the construction site is used as a case. The proposed framework offers a reliable, real-time communication classification system for use on construction sites as a supplementary means of communication in crane operations. Firstly, by developing computer vision-based integrated deep learning model with the capability to detect and classify dynamic hand signals in real-time, even in the presence of complex and varying weather conditions at the construction site. Secondly, by developing sensor-based smart construction glove that uses machine learning models to classify crane signaller dynamic hand signals in real-time. Additionally, to enhance speech communication, this research proposed a one-dimensional convolutional neural network model.

This model is designed to identify the crane signalman speech commands in real-time by providing crane operators with the necessary keywords to understand the signalman's instructions and support their decision-making process. Finally, the proposed framework implement the concept of redundancy by utilizing ensemble models. These models combine the decisions from computer vision-based deep learning; sensor based smart construction glove; and keyword identification model using weighted average and majority voting techniques, resulting in a single, reliable decision output. Overall, this research improve the reliability of site communication by classifying the hand signals and speech commands (both individually and in the aggregate using ensemble models), classifying to a high level of specificity the hand signals and speech commands used in the communication between crane operators and signalmen.

PREFACE

This thesis is the original work of Asif Mansoor. Three journal papers and four conference papers related to this thesis have been submitted for review or published as listed below. This thesis is organized in a paper-based format in accordance with the guidelines specified by the Faculty of Graduate Studies and Research.

1. Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., and Al-Hussein, M. “Scientometric analysis and critical review on the application of deep learning in the construction industry.” *Canadian Journal of Civil Engineering*, 50(4), 253–269.
2. Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., and Al-Hussein, M. “A deep learning classification framework for reducing communication errors in crane dynamic hand signaling.” *Journal of Construction Engineering and Management*, 149(2), 04022167.
3. Mansoor, A., Liu, S., Bouferguene, A., and Al-Hussein, M. “Crane signalman hand signal classification framework using sensor-based smart construction glove (SCG) and machine learning algorithms.” Under review for publication in *Journal of Construction Engineering and Management*.
4. Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., Al-Hussein, M., and Soda (2022) “Mobile crane signalman static hand signals classification framework using deep convolution neural network.” *Proceedings of the 34th European Modeling & Simulation Symposium*, Rome, Italy, Sep. 19–21, 2022, ISSN 2724-0029.
5. Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., Al-Hussein, M., and Hassan, I. (2022). “Keyword identification framework for speech communication on construction sites.” *Proceedings of the 2022 Modular and Offsite Construction Summit*, Edmonton, AB, Canada, Jul. 27–29, pp. 106–113.

6. Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., Al-Hussein, M. "Effectiveness of data augmentation in construction site-related image classification." Accepted (Apr., 2022) for publication in *Proceedings of the Canadian Society for Civil Engineering Annual Conference*, Whistler, BC, Canada, May 25–28, 2022.
7. Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., and Al-Hussein, M. (2020). "Conceptual framework for safety improvement in mobile cranes." *Proceedings of the Construction Research Congress 2020: Computer Applications*, Tempe, AZ, USA, Mar. 8–10, pp. 964–971.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my supervisors, Dr. Mohamed Al-Hussein and Dr. Ahmed Bouferguene, for their unwavering support, guidance, and patience throughout my studies. Their invaluable feedback, suggestions, and constant encouragement have been instrumental in shaping this research and thesis. I would also like to extend my gratitude to my supervisory and examining committee, Dr. Yasser Mohamed, Dr. Abdulhakem Elezzabi, and Dr. Hyoungkwan Kim, for reviewing my thesis and providing valuable feedback.

I would like to express my thanks to Jonathan Tomalty and Kristin Berg for editing my thesis and for their valuable suggestions on my publications. I am also thankful to my friends and colleagues for their support and encouragement. Their insightful comments, suggestions, and discussions have been immensely helpful in shaping my ideas and improving my work.

I am also deeply grateful to my parents, my wife, and other family members, who have always been there for me, providing me with their love, support, and understanding. Without their constant encouragement, this achievement would not have been possible.

This thesis is dedicated to Kiran Baloch and Izyan Baloch, my motivation and strength to pursue my dream.

TABLE OF CONTENTS

ABSTRACT	ii
PREFACE	iv
ACKNOWLEDGEMENTS	vi
TABLE OF CONTENTS.....	vii
LIST OF TABLES	xii
LIST OF FIGURES	xiii
LIST OF ABBREVIATIONS.....	xvi
Chapter 1: INTRODUCTION	1
1.1 Background	1
1.2 Advancement of communication in the construction industry.....	2
1.2.1 Hand signals.....	3
1.2.2 Two-way radio speech communication	5
1.3 Problem Statement and Motivation.....	6
1.4 Research Objectives	9
1.5 Thesis organization	10
Chapter 2: SCIENTOMETRIC ANALYSIS AND CRITICAL REVIEW ON THE APPLICATION OF DEEP LEARNING IN THE CONSTRUCTION INDUSTRY.....	13
2.1 Introduction	14
2.2 Methodology	16

2.2.1	Scientometric analysis (Phase 1)	17
2.2.2	Results of scientometric analysis	20
2.2.3	Critical review of clusters (Phase2)	37
2.3	Challenges and Future directions	41
2.3.1	Challenges.....	41
2.3.2	Future directions	44
2.4	Conclusion.....	46
Chapter 3: A DEEP-LEARNING CLASSIFICATION FRAMEWORK FOR REDUCING COMMUNICATION ERRORS IN DYNAMIC HAND SIGNALLING FOR CRANE OPERATIONS.....		
3.1	Introduction	49
3.2	Related work on hand signal classification	51
3.3	Proposed intelligent framework	52
3.3.1	Input data collection and preprocessing.....	53
3.3.2	The main process of the intelligent framework	60
3.4	Deep-learning algorithms.....	60
3.4.1	YOLOv4 architecture.....	60
3.4.2	Modification in YOLOv4 model architecture.....	63
3.4.3	LSTM architecture	65
3.5	The architecture of the integrated model.....	67

3.6	Case Study.....	68
3.6.1	Model Training	68
3.6.2	Model Performance evaluation on the test dataset	71
3.6.3	Real-time implementation.....	75
3.7	Discussion	78
3.8	Conclusions and future work	80
Chapter 4: CRANE SIGNALMAN HAND SIGNAL CLASSIFICATION FRAMEWORK USING SENSOR-BASED SMART CONSTRUCTION GLOVE (SCG) AND MACHINE- LEARNING ALGORITHMS.....		
		82
4.1	Introduction	82
4.2	Related literature	84
4.2.1	Sensors for hand signal classification	84
4.2.2	Wearable Sensors in the construction industry.....	85
4.3	Research framework.....	87
4.3.1	Custom build SCG	88
4.3.2	Data collection and preprocessing	93
4.3.3	Machine-learning models.....	95
4.3.4	Models performance evaluation.....	98
4.4	Mobile application for real-time crane signalman hand signal classification	101
4.5	Conclusion.....	103

Chapter 5: KEYWORD IDENTIFICATION FRAMEWORK TO FACILITATE THE SPEECH COMMUNICATION ON CONSTRUCTION SITES	106
5.1 Introduction	107
5.2 Related studies.....	108
5.3 Research methods.....	109
5.4 Implementation and case study	110
5.4.1 Dataset collection and preprocessing.....	111
5.4.2 Model development	111
5.5 Results	114
5.6 Conclusion.....	115
Chapter 6: AN ENSEMBLE TECHNIQUE TO IMPROVE THE ACCURACY OF ONSITE COMMUNICATION CLASSIFICATION. A CASE OF COMMUNICATION BETWEEN THE CRANE OPERATOR AND CRANE SIGNALMAN.....	117
6.1 Introduction	118
6.2 Related work	119
6.3 Methodology	122
6.4 Ensemble models.....	123
6.4.1 Weighted average ensemble model	123
6.4.2 Majority voting ensemble model	125
6.5 Results	128

6.5.1	Results of weighted average ensemble model	129
6.5.2	Results of Soft Majority Voting ensemble model.....	130
6.5.3	Results of hard majority voting ensemble model	132
6.6	Discussion	133
6.7	Conclusion.....	135
Chapter 7:	CONCLUSIONS.....	137
7.1	Research Summary.....	137
7.2	Research Contributions	140
7.2.1	Industrial contributions	140
7.2.2	Academic contributions	141
7.3	Limitations and Future Research.....	142
REFERENCES	145
APPENDIX 1:	185
APPENDIX 2:	209

LIST OF TABLES

Table 2-1 21

Table 2-2 24

Table 2-3 27

Table 2-4 31

Table 2-5 33

Table 3-1 55

Table 3-2 59

Table 3-3 69

Table 3-4 75

Table 3-5 77

Table 4-1 99

Table 4-2 100

Table 5-1 112

LIST OF FIGURES

Figure 1-1. Timeline of key advancements in communication technologies in recent decades.....	1
Figure 1-2. Examples of hand signals.....	4
Figure 1-3. Number of fatalities in crane-related accidents each year in the United States.	8
Figure 2-1. Overview of research design.....	17
Figure 2-2. Historical trend of published studies on deep learning in construction, 2012–2021.	20
Figure 2-3. Network of co-occurring key words related to deep learning in the construction industry.	25
Figure 2-4. Network of co-authorship for publication related to deep learning in construction. .	26
Figure 2-5. Network of countries/regions.	29
Figure 2-6. Network of author co-citation for publication to deep learning in construction.....	30
Figure 2-7. Network of document co-citation.....	32
Figure 2-8. Network of document co-citation with clusters.	36
Figure 2-9. Timeline view of clusters.	37
Figure 3-1. Overview of proposed intelligent framework.	53
Figure 3-2. Crane signalman hand signals. (Reprinted from OSHA 2021).....	54
Figure 3-3. Camera placement for data collection and real-time crane signalman hand signal classification.	56
Figure 3-4. Example of data augmentation.....	58
Figure 3-5. Detail architecture of proposed YOLOv4 model.	61
Figure 3-6. Graphical representation of activation functions: (a) Mish; (b) ReLU; and (c) Leaky-ReLU.....	64
Figure 3-7. Architecture of LSTM model.....	66

Figure 3-8. Proposed YOLOv4+LSTM model architecture.	67
Figure 3-9. (a) Training accuracy; (b) validation accuracy; (c) validation loss; and (d) training loss of deep-learning models.	71
Figure 3-10. Confusion matrix of YOLOv4 (Mish)+LSTM on the test dataset.	72
Figure 3-11. Confusion matrix of YOLOv4 (ReLU)+LSTM on the test dataset.	72
Figure 3-12. Confusion matrix of YOLOv4(Leaky-ReLU)+LSTM on the test dataset.	73
Figure 3-13. Examples of real-time classification using improved YOLOv4+LSTM model.	79
Figure 4-1. Overview of the research framework.	87
Figure 4-2. Layout of smart construction glove.....	89
Figure 4-3. Example of sensor calibration.....	94
Figure 4-4. Example of data transmission via Bluetooth.....	95
Figure 4-5. Training and validation accuracies of machine-learning models.....	99
Figure 4-6. Confusion matrix of KNN model.....	101
Figure 4-7. Confusion matrix of CNN-LSTM model.....	101
Figure 4-8. Layout of developed mobile application.....	102
Figure 4-9. Real-time performance of KNN and CNN-LSTM model.....	103
Figure 5-1. Overview of keyword identification framework.....	110
Figure 5-2. Crane signalman speech command/keywords.	111
Figure 5-3. Waveform audio sample of keywords and corresponding MFCCs.	113
Figure 6-1. overview of the proposed ensemble framework.	122
Figure 6-2. Overview of the developed weighted average ensemble model.	124
Figure 6-3. Overview of soft majority voting	126
Figure 6-4. Overview of hard majority voting.....	127

Figure 6-5. Performance of weighted average ensemble model.....	129
Figure 6-6. Performance of Soft majority voting (2-medium) ensemble model.	131
Figure 6-7. Performance of Soft majority voting (3-medium) ensemble model.	131
Figure 6-8. Performance of hard majority voting ensemble model.	133
Figure 6-9. Performance of proposed ensemble model on test dataset.....	134
Figure 6-10. Comparison of the classification accuracy of the ensemble models with respect to each hand signal.....	135

LIST OF ABBREVIATIONS

ANN	Artificial Neural Network
AR	Augmented Reality
ASL	American Sign Language
BIM	Building Information modelling
CAC	Copyright Act of Canada
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
DT	Decision Tree
FPS	Frames per Second
GRU	Gated Recurrent unit
IMU	Inertial Measurement Unit
KNN	<i>k</i> -Nearest Neighbour
LSTM	Long Short-Term Memory
MFCCs	Mel-Frequency Cepstral Coefficients
MNIST	Modified National Institute of Standards and Technology database
OSHA	Occupational Safety and Health Administration
PIPEDA	Personal Information Protection and Electronic Documents Act
PPE	Personal Protective Equipment
RCNN	Region-Based Convolutional Neural Network
RNN	Recurrent Neural Network
SCG	Smart Construction Glove
SSD	Single-Shot Detector
SVM	Support Vector Machine
TBM	Tunnel Boring Machine
VGG	Visual Geometry Group
VR	Virtual Reality
YOLO	You Only Look Once

Chapter 1: INTRODUCTION

1.1 Background

Communication technologies play a significant role in establishing and maintaining quality working relationships in any industry. Means of communication have advanced over time. For instance, in the early 1800s, communication in the aviation industry, shipping industry, and postal services relied on Morse code and telegraph (OCN 2021), but with the rapid introduction of new technologies, the means of communication have changed dramatically in recent years. In the last 50 years, communication technologies have advanced rapidly, from the development of satellite radios in the 1970s to virtual reality and artificial intelligence more recently. Some of the most popular technological advancements to enhance communication in each decade are shown in Figure 1-1.

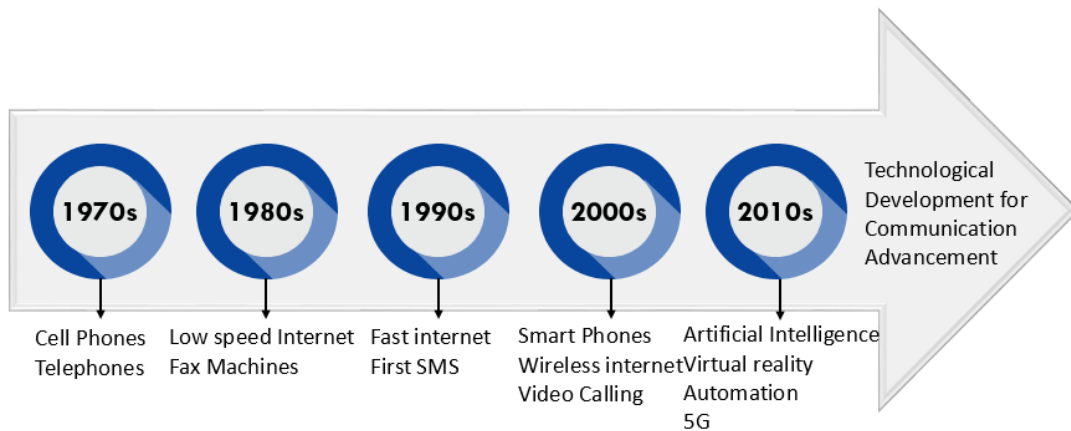


Figure 1-1. Timeline of key advancements in communication technologies in recent decades.

Various industries have adopted advanced communication technologies in different ways over the years. In the business sector, for example, corporations have used email, video conferencing, instant messaging, and other digital tools to facilitate communication among staff and with clients (Armírola Garcés et al. 2020; Scolari. 2009). Telemedicine and other forms of remote

communication have been implemented in the healthcare sector to enhance patient care and access to medical knowledge (Matusitz and Breen. 2007; Zagan et al. 2017; Ratta et al. 2021). Streaming services and social media platforms have been used to disseminate material and engage viewers in the entertainment sector (Ahuja. 2021). GPS and other types of digital communication have been used in the transportation industry to improve logistics and fleet management (Vivaldini et al. 2012; Hu et al. 2015). Overall, the introduction of communication technology has allowed various economic sectors to boost efficiency and production.

1.2 Advancement of communication in the construction industry

Historically the construction industry has exhibited a reticence to adopt advanced technologies due to the dynamic and unpredictable nature of construction work and the inconsistency of the results achieved when implementing advanced communication technologies in this environment (Hwang et al. 2022; Mansor 2010). Nevertheless, with these technologies become more well established, and with the high volume and quality of research on their implementation now available, these technologies are increasingly being adopted in the industry today (Hwang et al. 2022; Mansor 2010). For example, the introduction of Building Information modelling (BIM) technology has provided the foundation for a digital transformation in the construction industry in recent years (Azhar 2011). BIM allows for stakeholders to collaborate in a common digital platform on the planning, design, and construction in order to speed the construction process, reduce ambiguities, and improve the overall efficiency of building construction (Azhar 2011). Another example is “remote monitoring” of the construction process. Remote monitoring allows supervisors to monitor construction processes in real time from anywhere so that they can be apprised of updates and issues and take remedial action as required (Soltanmohammadlou et al. 2019).

Project management software tools such as Microsoft Project allow for task, deadline, and progress monitoring, as well as communication and collaboration among team members. Internet of Things (IOT) devices, meanwhile, are used in the construction industry to monitor construction site conditions such as humidity, temperature, and air quality, as well as to track the position and movement of construction equipment and supplies (Teizer et al. 2017; Patel et al. 2016; Mishra et al. 2022).

While many of these advancements have focused on the tools and methods used to complete construction tasks, worksite communication also plays a vital role, given its impact on worker safety, performance, and productivity (Kines et al. 2010; Neitzel et al. 2001). On construction sites, communication among workers can be categorized into short- and long-distance communication. When the distance between workers is short, face-to-face verbal communication is the most reliable means to convey the message, while hand signals are used when verbal communication is not possible, such as when the site is noisy (Hagan et al. 2015). For long-distance communication, meanwhile, two-way radio speech communication and hand signalling are the primary means of communication (Hagan et al. 2015).

1.2.1 Hand signals

Hand signals have been in use for millennia as a means of communication. In recent years, hand signalling has been used widely in many applications, such as for speech- and hearing-impaired people as a means of communication (Sriram and Nithiyandham 2013), for traffic direction on roadways (Guo et al. 2017), and in aviation for safe flight operations (IATA 2020). On construction sites, workers use standard hand signals for long-distance communication, as well as for short-distance communication in noisy areas where verbal communication would be impossible or severely hampered (Hagan et al. 2015). Hand signalling is also used extensively to send commands

to operators of heavy construction machinery (e.g., excavator, concrete truck, bulldozer) when the working area is not clearly visible to the operator. A trained signaller stands in a location from where the site is clearly visible and guides the operator using hand signals to ensure safe construction site operations. Examples of hand signals are shown in Figure 1-2.

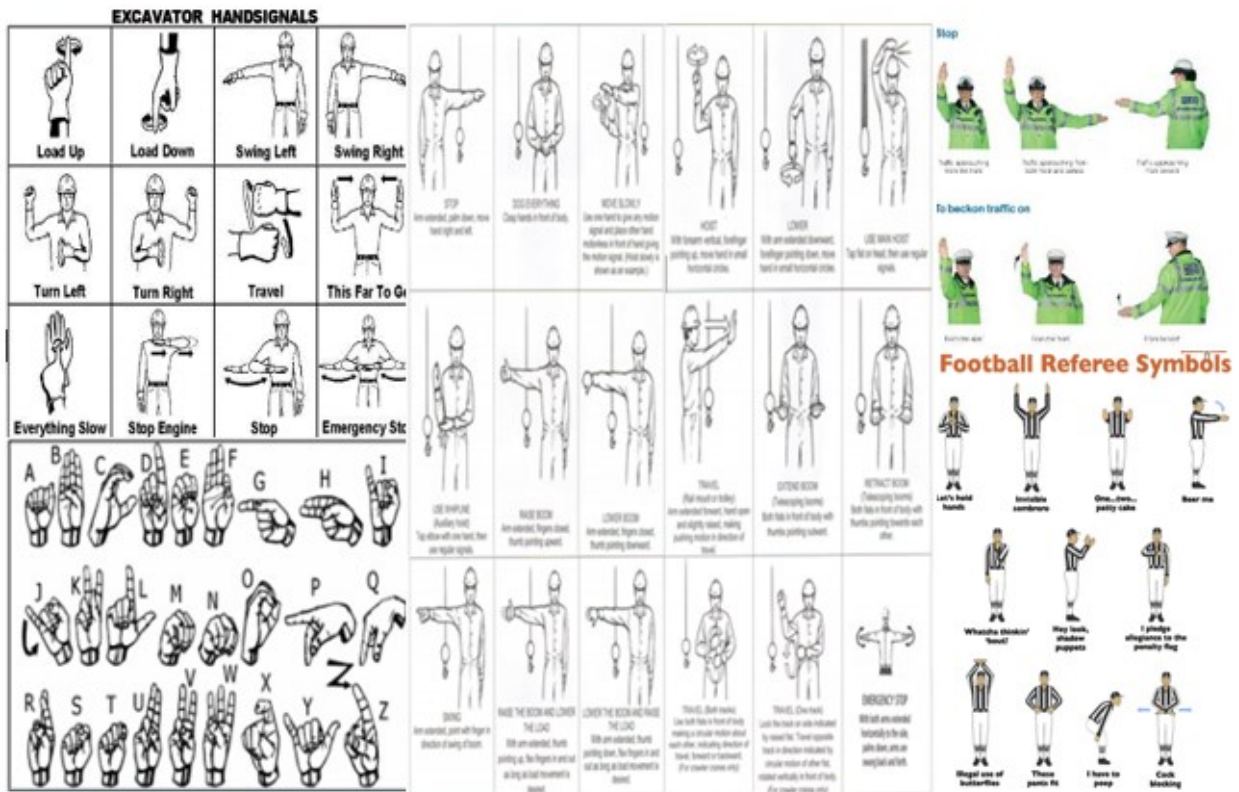


Figure 1-2. Examples of hand signals.

In crane operations, hand signalling is used to facilitate communication with the crane operator (All-West Crane and Rigging 2019). The advantage of using hand signals is that they are simple and effective, and they can be used in noisy areas where verbal communication is not possible. The operator receives direction from the signaller clearly and immediately, which makes the communication process efficient and effective (CCOHS 2019). On the other hand, site congestion and obstacles in the line of sight between signaller and operator can make the process less

efficient and effective, since a second signalman is often needed to relay the signals from the primary signalman to the operator, reducing the efficiency and increasing the risk of an accident due to transmission errors between signalmen (Fang and Cho 2016).

1.2.2 Two-way radio speech communication

Two-way radio speech communication is another solution for communication on construction sites. This approach is used when neither face-to-face verbal communication nor hand signalling is possible. On the construction site, this approach is mostly used for communication between workers on the ground and heavy construction machinery operators, such as in crane operations, where a two-way radio speech communication system is typically used when it is difficult for the operator to see the signalman due to an obstacle in their line of sight (Stevenson 2019). This communication approach requires a dedicated radio channel for the communication between the operator and signalman that must be maintained at all times. If there is any problem in the channel/communication line, such as radio interference from another radio, loss of signal due to radio blackspots, or excessive noise on the construction site, the operator may be unable to hear the signalman's voice clearly, in which case operations must be halted until the disruption to the line of communication is resolved (Maxim Crane Works 2019). These communication disruptions may also lead to misjudgments on the part of the operator, in turn resulting in safety and productivity issues. Furthermore, construction workers are often from different linguistic backgrounds and have different accents, meaning that it may be difficult for the listener to understand the speaker in some cases, leading to misjudgments in decision-making (Bust et al. 2008).

Given the inherent disadvantages of traditional means of communication such as verbal communication, hand signalling, and two-way radio speech communication, the construction

industry can benefit from recent developments in information technology. In particular, an intelligent and automated communication classification system can be introduced to improve communication between heavy construction machinery operators and workers on the ground and thereby enhance the safety and productivity of site operations. In this context, the present study develops a framework that provides an intelligent and advanced communication classification system to reduce the risk of miscommunication on construction sites.

1.3 Problem Statement and Motivation

One of the major concerns in the construction sector today is construction site safety. However, the inherent deficiencies of traditional means of communication on construction sites (i.e., hand signalling and two-way radio) mean that construction workers are still exposed to safety hazards, especially in heavy construction machinery operations. For example, in a recent case, a concrete truck struck an electrician who was replacing a traffic light due to a misread of a hand signal given by an officer (Reakes 2018). In another example, an accident occurred when a crew chief in an All-Terrain Vehicle (ATV) travelled into an active work area. Although he had been asked to clear the area by a foreman using a hand signal, the signal was not received by the crew chief. As a result, the ATV was struck by a bulldozer and the crew chief suffered an injury (ENFORM 2013). In crane operations in particular, communication is among the most significant challenges (Neitzel et al. 2001). With the emergence and continued evolution of modular construction in recent decades, heavy mobile cranes play an increasingly important role on construction sites, as cranes are required to load and assembly increasingly large, heavy, and complex modular components. This shift toward modular and offsite construction underscores the criticality of effective crane operations to project productivity and safety (Bernold et al. 1997, Ali et al. 2021). In practical terms, the rise in the fatality rate due to accidents involving cranes has meant that every activity

involving a crane on the construction site is now considered hazardous (King 2012). Indeed, a small error on the part of a crane operator or a rigging failure during a heavy lift can pose an enormous risk to both the operator and the workers in close proximity to the crane and load. According to statistics from the US Bureau of Labor Statistics (2017), from 2000 to 2017, the number of fatalities in crane-related accidents in the United States totalled 1,097 from all industry sectors, as shown in Figure 1-3. The construction industry was responsible for 632 (57%) of those fatalities. According to the U.S. Department of Labor as reported by Beavers et al. (2006), the construction industry had the third-highest fatality rate among the main economic sectors in 2001, with 13.3 fatalities per 100,000 construction workers. They estimated that about 84% of the fatalities in the construction sector were linked to crane operations on the site. Furthermore, according to OSHA's Integrated Management Information System (IMIS) database as reported by Zhao (2011), during the period 2000 to 2009, the 571 accidents involving crane work accounted for 587 fatalities, of which 105 were crane operators, 375 were riggers, and 8 were signalmen (Zhao 2011). Moreover, according to King (2012), during the period from 2004 to 2010, 43% of crane accidents were due to operator failure in their responsibilities.

OSHA has stated that improper communication is the leading cause of crane-related accidents, while studies by the Center for Construction Research and Training, similarly, has reported that improper communication is a contributing factor in over 40% of crane-related accidents (CPWR 2019). The National Commission for the Certification of Crane Operators, moreover, has stated that miscommunication between crane operator and signal person is one of the leading causes of crane-related incidents and accidents (NCCCO 2019), while the International Association of Tower Crane Owners (IATCO) has argued that signal-related errors are among the top three causes of tower crane incidents worldwide. A recent investigation of the risk factors in crane-related near

misses and accidents revealed that operator and signalman error is the predominant risk factor, accounting for 34% of the incidents, while failure in communication accounted for 11% of the 212 incidents investigated (Raviv and Shapira 2018). Indeed, the lack of proper communication between crane operators and signalmen can lead to disastrous accidents and is one of the major contributors to the high rate of fatalities in construction (Fang et al. 2018a; c).

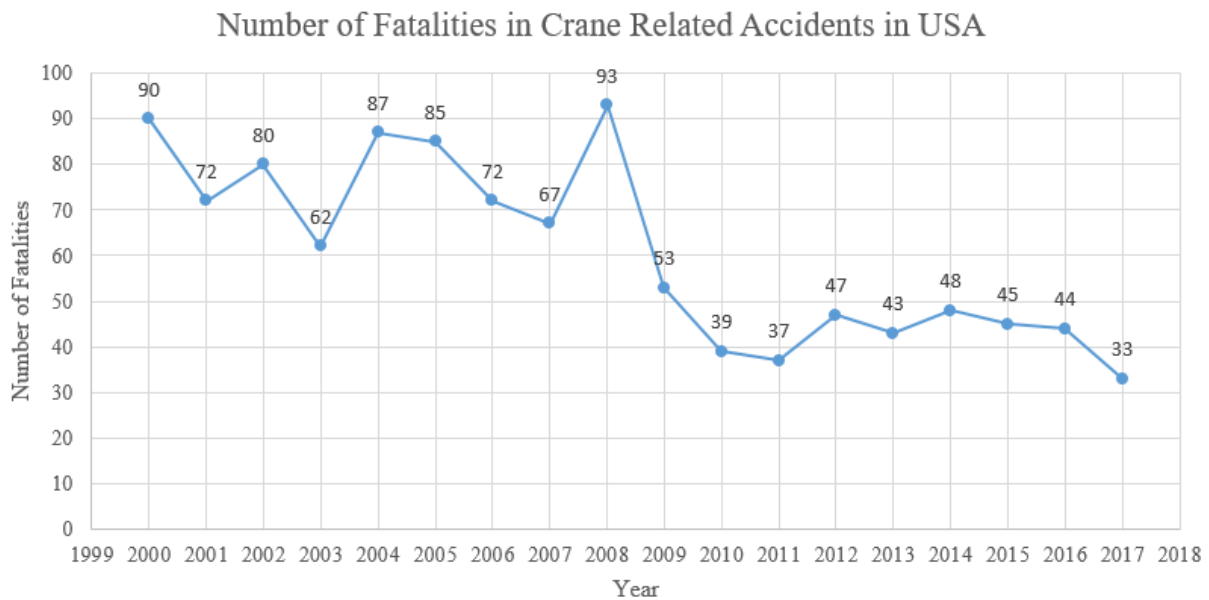


Figure 1-3. Number of fatalities in crane-related accidents each year in the United States.

In current practice, communication between heavy construction machinery operators and workers/signalmen on construction sites is still reliant on hand signalling and two-way radio speech communication (Fang et al. 2016). As demonstrated by the statistics discussed above, these methods are not always reliable, and their inadequacy can be a factor in disastrous accidents. With the ongoing development of information technologies, autonomous and intelligent communication classification systems can be implemented to improve communication between workers/signalmen and heavy machinery operators in order to reduce the probability of communication errors and minimize the number of miscommunication-related accidents on construction sites. Further

investigation is as to how the reliability of site communication might be improved through the application of such technologies (e.g., computer vision using deep-learning algorithms, sensor-based smart glove technology, speech recognition algorithms, and the concept of redundancy) is thus warranted.

1.4 Research Objectives

This research is built upon the following research hypothesis.

“The development of real-time multimodal communication classification on construction sites improves the reliability of communication and the accuracy of the information exchanged”

This research is rooted in the premise that the current practice of communication on construction sites (hand signalling and two-way radio) is unreliable, especially with respect to crane operations, contributing to the risk of disastrous incidents in modern construction. To address this gap, four-research questions are considered in this study.

1. How can we use computer vision-based deep-learning classification models to classify dynamic crane signalman hand signals on construction sites?
2. To what extent might a sensor-based smart construction glove (SCG) implemented in conjunction with machine-learning models help to classify crane signalman hand signals on construction sites?
3. How can we use deep learning models to classify crane signalman speech commands on construction sites?
4. How can we implement the concept of redundancy together with an ensemble model to combine the decisions from multiple models and achieve a single, more accurate and reliable decision output?

To answer the questions posed above, the following objectives are pursued in this research:

1. Conduct a detailed scientometric analysis and critical review of the literature on the application of deep learning in construction.
2. Develop a deep-learning classification framework for reducing communication errors in dynamic hand signalling for construction sites.
3. Develop a sensor-based SCG that uses machine-learning models to enhance hand-signalling communication on the construction site.
4. Develop a keyword identification framework to facilitate speech communication on construction sites.
5. Implement the concept of redundancy by developing ensemble models that improve the accuracy by combining the decision of multiple deep-learning models to classify crane signalman hand signals and speech commands on the construction site.

1.5 Thesis organization

This thesis is composed of seven chapters. Chapter 1 presents the background; problem statement, and motivation underlying this research and briefly introduces the current practice and recent advancements in communication on construction sites. This chapter also outlines the hypothesis, research questions, and objectives.

Chapter 2 presents a comprehensive scientometric analysis and critical review of the literature on the application of deep learning in construction. The science mapping method is used to quantitatively and systematically analyze the related bibliographic records retrieved from the Scopus database, and co-word, co-author, and co-citation analyses are then performed. A critical review of the themes identified in the relevant publications is also performed. Finally, challenges and future directions of research on deep learning in construction are presented. This chapter lays

the foundation for how deep-learning models are to be used in construction sites and what challenges construction practitioners face with respect to the adoption of deep-learning methods in construction sites.

Chapter 3 proposes a deep-learning classification framework to reduce communication errors in dynamic hand signalling in the construction industry, especially in crane operations. The framework uses two state-of-the-art deep-learning models (YOLOv4 and LSTM). Modifications are made in the YOLOv4 model by altering its activation function. The modified YOLOv4 and LSTM models are integrated to classify dynamic hand signalling in real time.

Chapter 4 is composed of two main parts. The first part describes the development of the SCG. The sensors used in the glove include an accelerometer, gyroscope, magnetometer, and flex sensors. These sensors are responsible for providing data related to the orientation of the hand and bend in the fingers. In the second part, four machine-learning models— k -nearest neighbour (KNN), Support vector machine (SVM), Decision Tree (DT), and Convolutional neural network-Long short-term memory (CNN-LSTM)—are selected and trained using the data collected by the SCG. The machine-learning models are further validated and tested using the collected dataset. Furthermore, a mobile application is developed, and the machine-learning models deemed to be most accurate in this particular application are deployed to recognize crane signalman hand signals in real time.

Chapter 5 presents a keyword identification framework for speech communication on construction sites. In this framework, 1D CNN deep-learning model is used to identify the crane signalman speech commands used by the crane signalman to guide the crane operator. The results of the presented model suggest that the model could be used effectively for real-time implementation in construction sites.

The research presented in Chapter 6 adopts the concept of redundancy and develops ensemble models to improve the accuracy of communication classification in crane operations by combining the decisions of multiple deep-learning models on the construction site. In this framework, the decision outputs of previously developed deep-learning and machine-learning models are combined in an ensemble model that yields a single decision output. The weighted average and majority voting ensemble models are implemented and evaluated. Finally, the results and limitations of the various models are compared.

Chapter 7 outlines the conclusions, research contributions, study limitations, and potential avenues of future research in this domain.

Chapter 2: SCIENTOMETRIC ANALYSIS AND CRITICAL REVIEW ON THE APPLICATION OF DEEP LEARNING IN THE CONSTRUCTION INDUSTRY¹

The construction industry is one of the most important economic sectors, contributing significantly towards the growth and development of the global economy. However, the industry has long been plagued by challenges such as communication issues, productivity issues, and safety concerns. There has been growing interest in the application of deep-learning techniques to address some of the challenges in the construction industry. This chapter presents a review and analysis of the literature on the application of deep learning in the construction industry. The main objective of the research described in this chapter is to define the body of knowledge, review and analyze the relevant literature, highlight some of the key research studies addressing various challenges in the construction sector, and further suggest some possible areas of improvement with respect to the application of deep learning in the construction industry. Initially, a scientometric analysis of the relevant literature is conducted using the Scopus database. The scientometric analysis encompasses both data acquisition and data analysis. In the data analysis phase, co-word, co-author, co-citation, and clustering analysis are performed to analyze the development and structure of the relevant scientific literature and identify the dominant researchers and countries in this particular research area. Cluster analysis is also performed to identify the semantic themes contained in the textual data. Furthermore, this chapter presents a critical review of the identified theme and adoptability challenges, and proposes some potential avenues of future work regarding the adoption of deep learning to improve construction processes. It is also observed that the deep-

¹ A version of this chapter has been published in the *Canadian Journal of Civil Engineering* as follows: Mansoor, A., Liu, S., Ali, G.M., Bouferguene, A., and Al-Hussein, M. (2023). “Scientometric analysis and critical review on the application of deep learning in the construction industry.” *Canadian Journal of Civil Engineering*, 50(4), 253–269.

learning model has the potential to improve communication on the construction site, a prospect that has received relatively little attention within the research community to date.

2.1 Introduction

As the construction industry moves towards adopting advanced computer technologies to improve the efficiency of its operations, artificial intelligence is becoming increasingly popular among researchers who are developing innovative solutions for this industry. Deep learning is a subset of machine learning, (which, in turn, is a subset of artificial intelligence), that was proposed by Hinton and Salakhutdinov (2006). It has since been applied in various fields, especially in connection with computer vision, image processing, and video analysis. Deep-learning methods focus on learning feature hierarchies with features from higher levels of the hierarchy formed by the composition of lower-level features (Hazrati Fard and Hashemi 2019). The learning of features at different levels of abstraction permits a system to learn complex functions, and map the output from the input data directly (Najafabadi et al. 2015). The capability to learn powerful features from data will become increasingly important as the amount of data and range of applications of deep learning continue to grow (Lecun et al. 2013).

The benefits associated with the application of deep learning have been reported extensively in the literature (Khan and Yairi 2018, Pan et al. 2021, Patel et al. 2021). For example, deep learning uses advanced algorithms to learn the features of objects and further detect, and classify objects of interest with high accuracy from images, and videos in real time. Moreover, deep learning provides vast opportunities for safety improvement and management, construction progress monitoring, building energy demand prediction, productivity monitoring of construction equipment, data sequencing, data extraction, and many more.

Due to these benefits, deep learning has attracted a substantial amount of attention in multiple industries, including the construction industry, and has captured the attention of construction engineering researchers. This has, in turn, led to a rise in research work and publications on the application of deep learning in construction in recent years. To grasp the current status of the body of knowledge and to decrease the risk of neglecting essential research questions and possible areas of improvement and laying the foundation of this research, a detailed review and analysis of this research area is needed.

The contributions of the previous review papers in this area are substantive. However, they tend to focus on a particular deep-learning method or the implementation of deep learning to a specific construction-related problem. For example, Sherafat et al. (2020) presented a state-of-the-art review on automated methods for construction worker and equipment activity recognition. Wang and Hong (2020) discussed the use of reinforced learning for building control in a review paper in which the authors also presented opportunities for and challenges of using reinforced learning. Darko et al. (2020) contributed a review article on the implementation of artificial intelligence in the AEC industry that provides both a scientometric analysis to find the gaps and a look at future trends and research directions. Martinez et al. (2019) provided a detailed review of the use of computer vision applications in construction by performing a scientometric analysis to determine the current and future directions of the research. However, the review studies discussed above are mainly focused on quantitative analysis, and the applications are limited to a certain deep-learning method or certain construction problem. Considering these facts, the previous review studies do not provide a broad view of the state-of-the-art of research on the application of deep learning in construction. A study that offers a comprehensive objective review, as well as a broad analysis of the literature on deep learning in construction, is missing. This review study aims to fill the gap by

comprehensively surveying the intellectual core and the landscape of the literature on deep learning in construction. This study assists construction researchers in numerous ways by finding the scope and evaluating the quality of the literature. A scientometric analysis (e.g., Co-Word, Co-Author, Co-Citation, Cluster analysis) was carried out to visualize the existing research by tracking, research keywords, influential journals, authors, and countries and explore the knowledge base of deep learning in construction research. Furthermore, a critical review of the identified clusters is also conducted. The fusion of critical review and scientometric analysis aids in the understanding of results, as well as helps to identify challenges in the application of deep learning in construction. In other terms, this review study helps by serving as an up-to-date reference point for expanding the knowledge of researchers in the construction domain.

The balance of this chapter is organized as follows: The next section illustrates the research methodology (i.e., scientometric analysis, data acquisition, keyword search, and data processing), which is followed by describing the results and findings of the scientometric analysis. Next is the discussion of the critical review of the identified cluster of the literature, followed by demonstrating the challenges and future direction of the application of deep learning in construction. The final section concludes by summarizing the research contributions.

2.2 Methodology

This study was conducted in two phases. Phase 1 includes scientometric analysis while Phase 2 is the critical review. The overview of the research design is shown in Figure. 2-1.

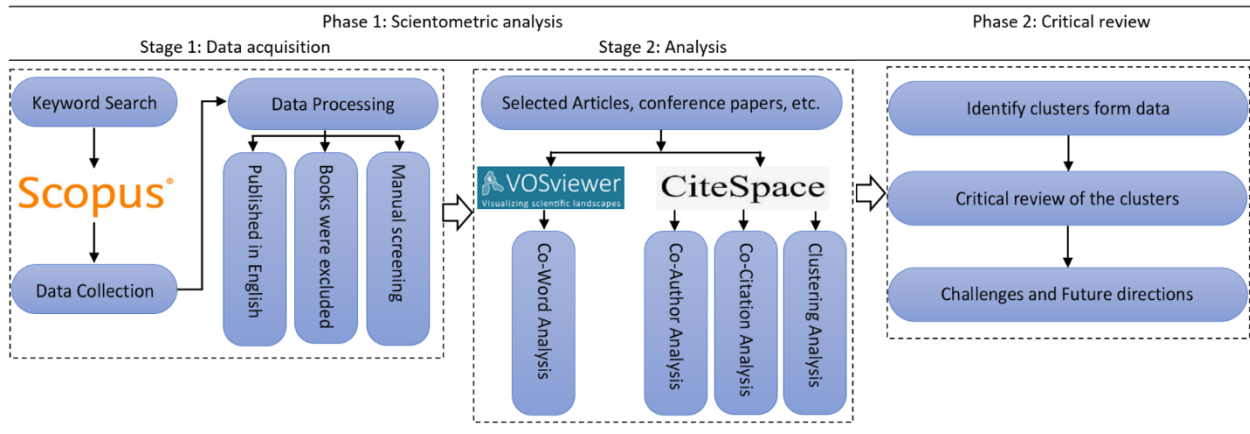


Figure 2-1. Overview of research design.

2.2.1 Scientometric analysis (Phase 1)

The scientometric analysis was conducted in two stages, namely, data acquisition and analysis of literature.

2.2.1.1 Data Acquisition (Stage 1)

In Stage 1, data acquisition of available literature (using keyword search and database selection) is of utmost importance, because it helps to determine the research studies available in the literature. There are several databases available, such as Web of Science, PubMed, Google Scholar, Scopus, etc., that can be used for data collection. However, the literature for the present study was derived from the Scopus database, because, in comparison to other previously mentioned databases, it has a wider range of construction industry-related research articles and provides a broader scope for inter-disciplinary research topics (Mok et al. 2015, Mongeon and Paul-Hus 2016).

2.2.1.1.1 Keyword search and data processing

The publications related to deep learning in construction were retrieved from the Scopus database by using the regular expression (“*deep learning*” OR “*YOLO**” OR “*convolutional neural*”).

*network" OR "CNN" OR "recurrent convolutional neural network" OR "RCNN" OR "perceptron" OR "SSD" OR "*detector*" OR "*neural network*" OR "deep convolutional neural network" OR "image classification" OR "object classification" OR "*classification*" OR "*detection*" OR "*recognition*") AND ("Construction industry" OR "Construction site" OR "construction project*" OR "*infrastructure*" OR "construction-site" OR "*construction *")*. Note that the variations in the keywords are captured using Asterisk (*) character. For example, “construction*” for “construction operation”, “YOLO*” for “YOLOv3”, and other related keywords. The publications having the selected keywords in their titles, abstracts, or selected keyword sections, are retrieved by choosing title/abstract/ keywords in the Scopus database keyword search engine. The search period of 10 years starting from 2012 to 2021 is taken into consideration, which is suitable considering the growth history of deep-learning technology in the construction industry. After the keyword search, a total of 531 references were found to satisfy the keyword search criteria. The data were further processed by applying the following filters to the search results.

- Only journal articles and conference papers were included.
- Manual screening was done to collect relevant journal and conference papers.

After applying all filters and manually screening the search results, the number of articles was reduced from 531 to 423. Of the 423 published papers, 298 were journal articles, and 125 were published in conference proceedings.

2.2.1.2 Analysis (Stage 2)

To process the data collected, scientometric analysis techniques were used to identify the systematized literature-related findings by linking concepts in the literature. Scientometrics was first defined by Mikołajczyk and Grochowski (2018) as “*a quantitative study of the research on the development of science*”. This technique measures the citation processes, impact of research,

and plots the existing knowledge and its development in the research area based on academic literature datasets (Siluo and Qingli 2017). The scientometric analysis facilitates the visualization and mapping of a knowledge domain that then helps researchers to analyze the intellectual landscape of a knowledge domain and find the research problems that may need to be solved, along with the techniques, the researchers have developed to solve similar research problems (Su and Lee 2010). There are several visualization and mapping tools available, such as “VOSviewer”, “Gephi”, “CiteSpace”, “Sci2”, and “HistCite” with each of the mentioned tools having its advantages (Chen 2017). In this research, VOSviewer was used for Co-Word analysis. VOSviewer is a suitable tool for visualizing large networks as it employs natural language processing algorithms and text-mining techniques (Eck and Waltman 2016). Therefore, VOSviewer was employed in the computing, visualizing, and analysis of data from journal articles, conference papers keywords in order to explore the research area. Visualization of the entire field of deep learning in construction allows the researcher to achieve a comprehensive perception of research trends and patterns in the field (Eck and Waltman 2010). CiteSpace was employed in the present study for Co-Citation, Co-Author and Cluster analysis due to its ability to provide visualizations that facilitate the analysis of scientific knowledge in the literature to identify the notion of the body of knowledge. These approaches have been well-known scientometric methods for discovering the hidden implications of a vast amount of information. CiteSpace is strong in mapping knowledge domains by systematically creating various accessible graphs (Kleinberg 2002, Chen 2016, Linton 2016); therefore, it was employed for the generation and analysis of co-citation networks, co-author networks and country co-occurrence, as well as for the generation of abstract clustering. The present study employs three scientometric techniques to analyze the selected publications. First, Co-Word analysis is performed, which includes keyword co-occurrence network. Second,

Co-Author analysis is performed, which includes co-authorship network and network of countries/regions. Finally, Co-Citation analysis is performed, which includes co-cited author, co-cited document and co-citation cluster analysis.

2.2.2 Results of scientometric analysis

2.2.2.1 Overview

In Stage 1 of the scientometric analysis, the keyword search strategy described in the previous section is employed to identify the number of publications (journal articles and conference papers) on the research topic under review for each year, as shown in Figure. 2-2. Deep learning in construction trended upward from 2018–2021 (i.e., 78% of the publications were published in the last 4 years) and is still gaining in popularity due to its easy adoption and the increasing availability of large datasets in construction, as shown in Figure. 2-2.

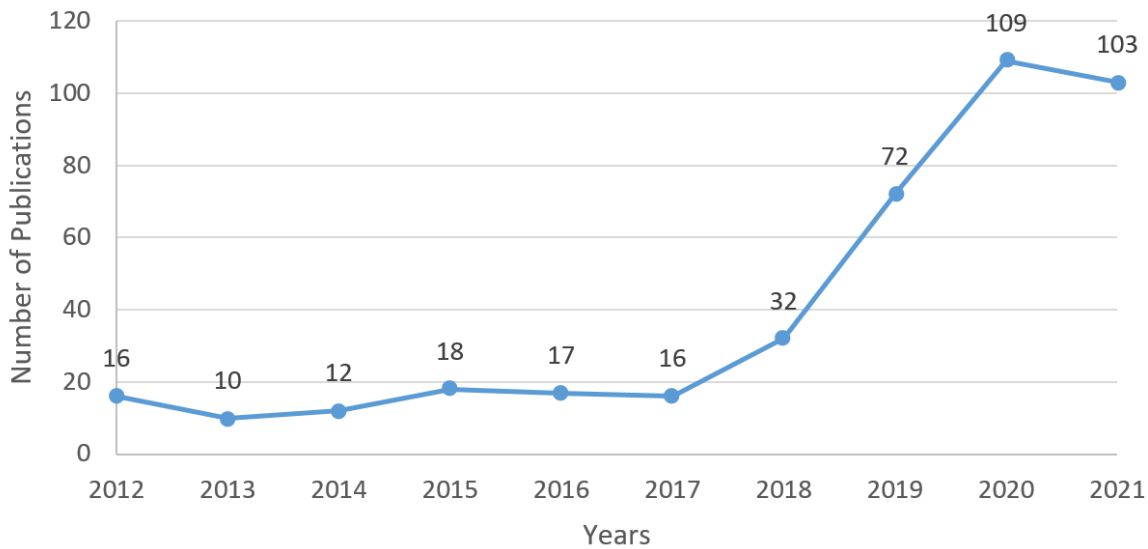


Figure 2-2. Historical trend of published studies on deep learning in construction, 2012–2021.

Table 2-1 summarizes the relevant academic journals and conference proceedings. The majority of academic publications on deep learning in construction were found in journals, such as

Automation in Construction, *Journal of Construction Engineering and Management*, and *Journal of Computing in Civil Engineering*. Among the mentioned journals, *Automation in Construction* represented the highest number of publications. Likewise, conference proceedings that make significant contributions to this topic include *Proceedings of the International Symposium on Automation and Robotics in Construction (ISARC)*, *American Society of Civil Engineers (ASCE) International Conference on Computing in Civil Engineering*, and *Proceedings of the Construction Research Congress (CRC)*. Journal articles accounted for 70% of the published papers reviewed for the present study, while the remaining 30% were papers published in conference proceedings.

Table 2-1.

List of top 10 academic journals and conference proceedings from 2012 to 2021 that published research related to deep learning in construction.

Journal/Conference name	Number of articles
<i>Journal of Automation in Construction</i>	52
<i>Journal of Construction Engineering and Management</i>	21
<i>Proceedings of ISARC</i>	18
<i>Journal of Computing in Civil Engineering</i>	17
<i>Proceedings of the American Society of Civil Engineers (ASCE)</i>	12
<i>International Conference on Computing in Civil Engineering</i>	
<i>Proceedings of the Construction Research Congress (CRC)</i>	12
<i>Journal of Applied science Switzerland</i>	8
<i>Journal of Expert Systems with Applications</i>	8
<i>Sensors Journal Switzerland</i>	7
<i>International Journal of Construction Management</i>	7

2.2.2.2 Co-word analysis

Co-word analysis was employed to analyze the structure and development of the scientific literature on deep learning in construction. In this study, keyword co-occurrence was used as the basic analysis unit.

2.2.2.2.1 Keyword co-occurrence analysis

Keywords provide a bullet-point summary of the main theme of the published research articles and highlight the range of researched areas within the boundaries of any domain (Kleinberg 2002, Eck and Waltman 2010, Chen 2016, Linton 2016, Daniels and Thistlethwaite 2017). Keyword co-occurrence analysis was used to map and visualize the knowledge area at the interface between deep learning and construction, this having been conducted using VOSviewer. The keyword network visualization assists in illustrating the output results of the scientometric analysis of the relevant literature. VOSviewer tool uses a distance-based map to show the strength of the relationship between two knowledge areas (Perianes-Rodriguez et al. 2016). The shorter distance between two items indicates a stronger relationship and vice versa. The different colours separate different knowledge areas clustered with the help of the clustering technique in VOSviewer. The size of the label represents the number of publications containing keywords, i.e., the larger the label is, the higher the number of publications containing the keywords it represents (Oraee et al. 2017).

The keywords of 423 academic publications on deep learning in construction were fed into VOSviewer to generate the co-occurrence graph of the keywords. The minimum number of co-occurrences was set to ten. Out of 3,848 keywords, 82 met the threshold level. The selection of the threshold was based on generating an optimal number of clusters. The network of co-occurring

keywords is shown in Figure. 2-3. The network comprises 82 nodes, 1,844 links, and a total link strength of 6,470.

Each keyword retrieved from the literature is shown in Table 2-2. It can be seen from Table 2-2 that “construction industry”, “neural networks”, and “deep learning” are the keywords that appeared most frequently in the literature, which indicates that these topics were more widely researched in this field. Table 2-2 also shows the total link strength and year with the most publications related to a particular keyword. A “link” refers to the number of connections of one keyword and other keywords, and total link strength shows the total strength of connections of a keyword with other keywords (Eck and Waltman 2016). For instance, the total link strength of “construction industry” is 1,291, indicating a strong relationship between the keyword “construction industry” and other keywords captured in the literature. The year with the most publications related to a particular keyword, on the other hand, is reflective of the keyword frequency in articles published by researchers in that time period. For example, the keywords “construction equipment” and “neural network” garnered more attention from researchers in 2017, whereas the keywords “deep learning”, and “object detection” were researched more frequently in the year 2020, meaning that deep learning has garnered more attention in recent years. It is also of note that earlier contributions tend to focus on specific deep-learning methods employed in the construction industry, while more recent contributions are more focused on the specific problems encountered in the construction industry.

Table 2-2.

List of selected keyword and relevant network data.

Keywords	Occurrence	Year with most publications related to keyword	Links	Total link strength
Construction industry	230	2018	97	1291
Neural networks	172	2017	93	929
Deep learning	113	2020	84	620
Construction site	93	2019	84	457
Convolutional neural networks	48	2020	74	347
Artificial neural networks	62	2016	69	311
Construction equipment	45	2017	72	298
Machine learning	43	2019	76	285
Computer vision	35	2019	65	277
Object detection	41	2020	61	270

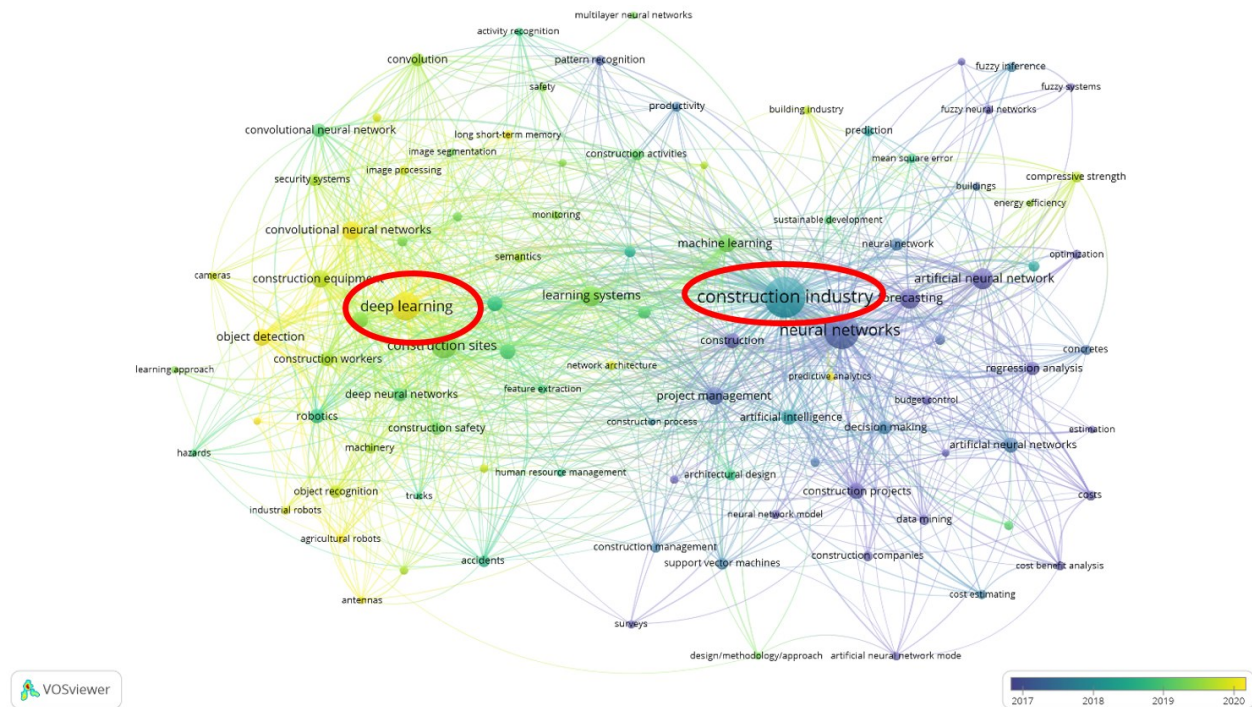


Figure 2-3. Network of co-occurring key words related to deep learning in the construction industry.

2.2.2.3 Co-Author analysis

In co-author analysis, a co-authorship network and network countries/regions were generated from the scientometric record containing authors from each article. The co-author analysis assists to identify the dominant researchers, and leading countries for deep learning in construction industry research.

2.2.2.3.1 Co-authorship network

Co-authorship network was used to represent the most productive authors based on the number of academic publications. On the bases of 423 academic publications on deep learning in construction, the top 10 authors with the most publications are given in Table 2-3. H. Li (Hong Kong), X. Luo (China), and S. Lee (United States) are the top three most productive authors. Figure 2-4 shows the co-authorship network, in which the nodes represent authors and the link between

authors indicates co-authorship collaboration in the selected articles. The redundant links in the network are removed through pathfinder, a technique recommended by (Chen and Morris 2003). The generated co-authorship network contains 298 nodes and 268 links. The size of the nodes is directly proportional to the number of publications.

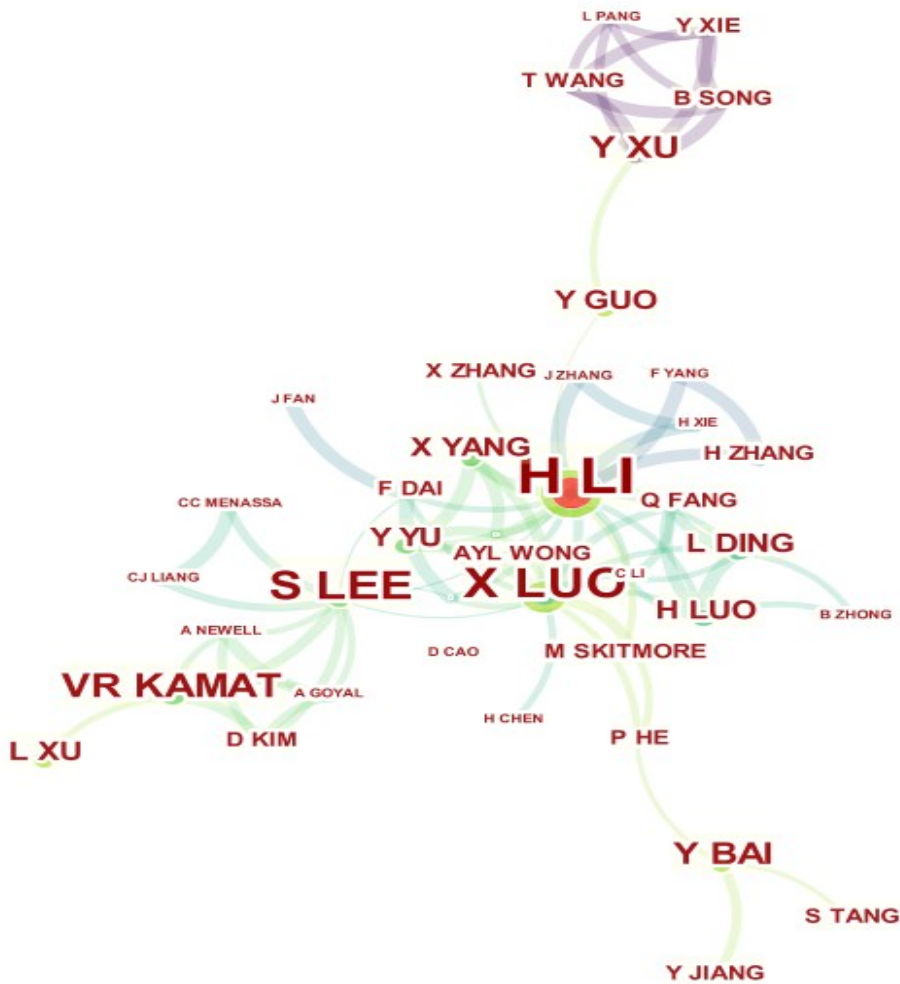


Figure 2-4. Network of co-authorship for publication related to deep learning in construction.

Figure 2-4 contains several close-loop orbits representing strong collaboration among the researchers in the orbit, such as the orbit of Y. Xu, B. Song, T. Wang, L. Pang, and Y. Xie. Furthermore, the co-authorship network identified a number of research communities, in which many authors worked with at least one or two most productive authors. For example, X. Luo and

H. Li were the two most productive authors of the research community that also included X. Yang, Y. Yu, H. Luo, etc., and V.R. Kamat was the most productive author of the research community that also included L. Xu, D. Kim, A. Goyal, etc. In a network, it should be noted, the betweenness centrality is defined as “the ratio of the shortest path between two nodes to the sum of all such shortest paths”(Linton 2016). The high betweenness centrality node connects a large group of nodes and can be identified by a red and green ring in CiteSpace. Such nodes can be of assistance to separate the clusters in the network and identify the revolutionary scientific publications (Girvan and Newman 2002, Chen 2006). In Figure 2-4, H. Li (centrality = 0.03), V.R. Kamat (centrality = 0.03), X. Luo (centrality = 0.03) and S. Lee (centrality = 0.03) are nodes with red and green rings, and they connect different groups of authors.

Table 2-3.

List of most productive authors from 2012 to 2021 involved in research on deep learning in construction.

Author	Country/Region	Number of publications
H. Li	Hong Kong	14
X. Luo	China	09
S. Lee	USA	08
A. Hammad	Canada	05
V.R. Kamat	USA	05
C. Kim	South Korea	04
Y. Xu	China	04
Y. Bai	USA	04
Y. Guo	China	03

2.2.2.3.2 *Network of countries/regions*

Similarly, a network of countries/regions was generated to explore the distribution of research publications on the application of deep learning in construction. The network consists of 128 nodes and 235 links. As shown in Figure 2-5, China (77 articles), the United States (62 articles), South Korea (31 articles), India (30 articles), Hong Kong (17 articles), Australia (17 articles), the United Kingdom (15 articles), and Canada (14 articles) are the major contributors to the publication in this area of research. The number of publications corresponds to the advancement of the research in the country/region. In comparison to the co-authorship network shown previously, the network of countries/ regions is relatively efficient and shows more homogeneity. The high centrality nodes are highlighted with red darker outer rings in Figure 2-5. Countries /regions including, China, United States, Hong Kong, and France with the centrality of 0.55, 0.42, 0.18, and 0.10, respectively, have gained top position in the network and linked research activities among several countries/regions.

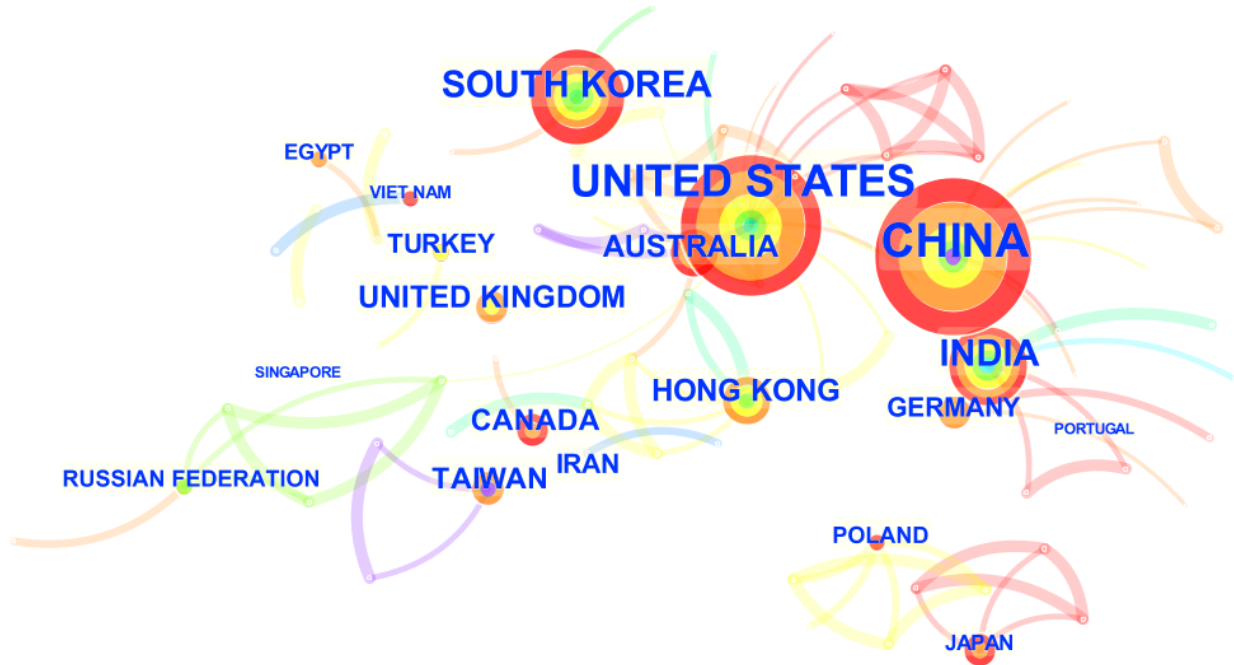


Figure 2-5. Network of countries/regions.

2.2.2.4 Co-citation analysis

According to Small (1973), co-citation can be defined as the “frequency with which two documents are cited together by other documents”. It measures the semantic similarity of the documents that use citation relationships. It is an efficient method for determining the similarity in the document. In this review article, the following co-citation analyses were performed: author co-citation, and document co-citation.

2.2.2.4.1 Author co-citation network

Author co-citation analysis was used to identify and visualize the intellectual structure of a given field of study by counting the frequency with which any work of an author is co-cited with another author in the references of citing documents. As Bayer et al. (1990) have noted, such an approach is helpful for analyzing the advancements of the research community within a given field of study.

As can be seen, the author co-citation network given in Figure 2-6 consists of 1,959 links and 466

nodes. The links reflect the indirect collaborations on the basis of co-citation frequency between authors, while the size of the node represents the number of co-citations of each given researcher. Accordingly, from the co-citation network, the highly cited authors were identified, including S. Ren (frequency = 53, China), K. He (frequency = 52, China), J. Redmon (frequency = 52, United States), Q. Fang (frequency = 46, China), R. Girshick (frequency = 43, United States), and H. Kim (frequency = 41, South Korea). The most cited authors are from China and the US, which shows that this field of research is new and still growing and only gaining attention among researchers in more developed countries.

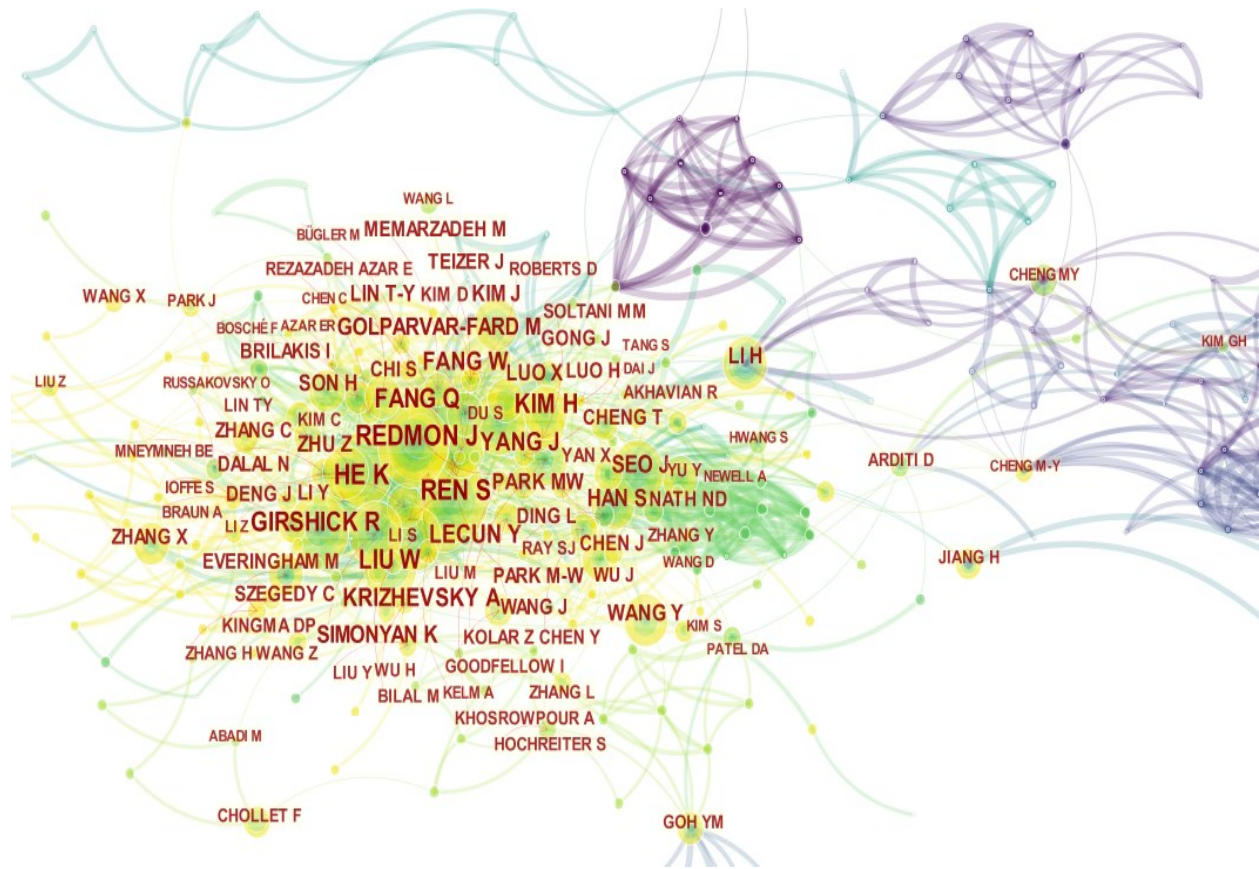


Figure 2-6. Network of author co-citation for publication to deep learning in construction.

2.2.2.4.2 Document co-citation analysis

In this chapter, document co-citation analysis was used to determine the core intellectual structures in the field of research, as well as to determine the authority and quantity of references cited by other publications. The generated document co-citation network represents the relationship between citations at an individual level. The document co-citation network given in Figure 2-7 has 396 nodes and 1,157 links. The nodes in the network represent the publications, where the labels show the first author's name along with the publication year. The links in the network show the co-citation relationship among the corresponding publications. The top five most highly cited articles as given in Figure 2-7 are (Redmon et al. 2016), (Fang et al. 2018b), (Fang et al. 2018c), (Ding et al. 2018), and (Lecun et al. 2015), which are also summarized in Table 2-4.

Table 2-4.

List of most cited documents from 2012–2021.

Article	Total citations	Year	Centrality
Redmon et al. (2016)	23	2016	0.12
Fang et al. (2018b)	19	2018	0.09
Fang et al. (2018c)	19	2018	0.07
Ding et al. (2018)	17	2018	0.08
Lecun et al. (2015)	14	2015	0.10
Fang et al. (2018d)	14	2018	0.06
Ren et al. (2017)	14	2017	0.04
Redmon and Farhadi (2018)	12	2018	0.06
Liu et al. (2016)	12	2016	0.01

Kingma and Ba (2015)	10	2014	0.01
----------------------	----	------	------

These articles have been widely recognized by researchers as adding significant value with respect to research on the application of deep learning in construction. For instance, Fang et al. (2018b) implemented a deep-learning method to detect and classify non-hardhat-use by the worker on the construction site to improve the supervision of construction workers in order to reduce the likelihood of accidents. The use of deep learning proved to be effective in detecting workers non-hardhat-use and thereby enhance safety inspection and supervision on the construction site. Based on the centrality scores, Redmon et al. (2016) and Lecun et al. (2015), with centrality scores of 0.12 and 0.10, respectively, have occupied the top two positions in terms of their significant influence on the body of knowledge on the application of deep learning in construction.

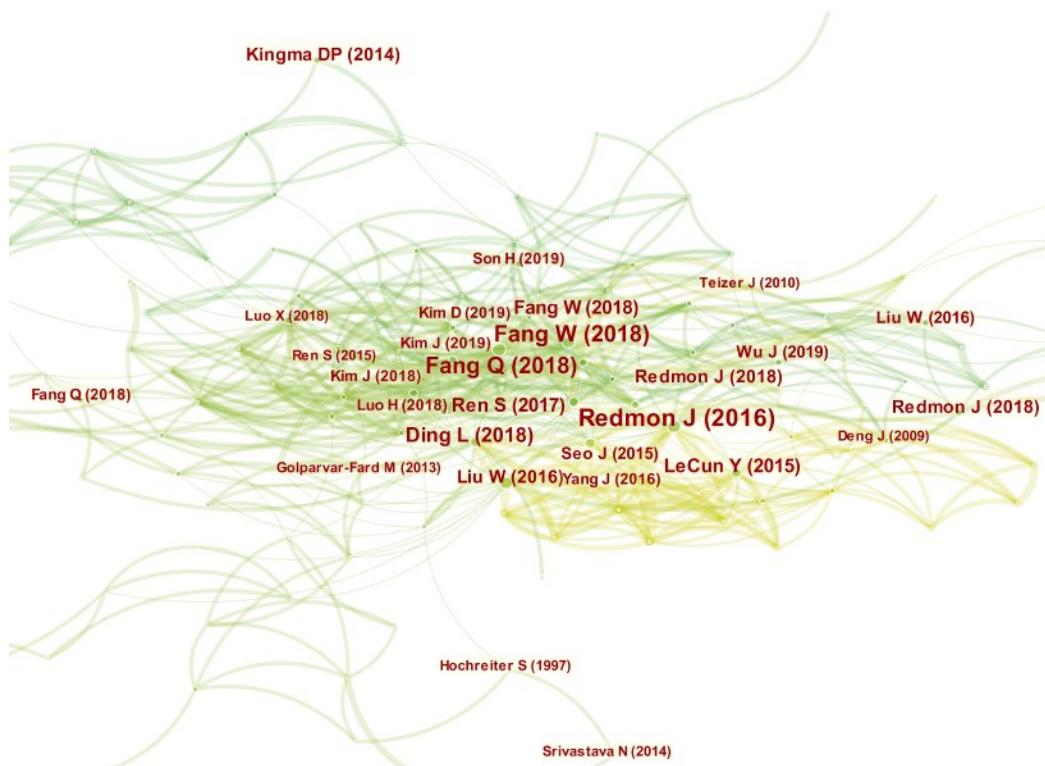


Figure 2-7. Network of document co-citation.

2.2.2.4.3 *Co-citation Cluster analysis*

Cluster analysis assists in identifying the hidden semantic themes in the textural data. The cluster analysis also determines the similar groups of the research data and labels them according to the focus of the cluster (Hossain et al. 2011). This will help to identify the trends in the research area based on the abstract and keywords of each of the documents cited. A total of 34 co-citation clusters were identified using the log-likelihood ratio (LLR) algorithm. The reason for using the LLR algorithm is its uniqueness, high quality of clusters, and wide coverage of the text data (Olawumi and Chan 2018). The most important identified clusters are shown in Figure 2-8.

Table 2-5.

Co-citation clusters of deep learning research for construction, 2012–2021.

Cluster ID	Size	Silhouette	Abstract cluster label	Alternative cluster label	Year with most publications	Representative documents
#0	64	0.787	Deep active learning approach	construction machinery monitoring using deep learning	2020	(Bang et al. 2019, Chen et al. 2020, Guo et al. 2020, Kim et al. 2020, Luo et al. 2020, Slaton et al. 2020, Xuehui et al. 2021, Zhou et al. 2021b, Lin et al. 2021, Sim et al. 2021,

						Xiao and Kang 2021a)
#1	34	0.908	Database- free vision- based monitoring	construction site activity and workers monitoring using deep learning	2019	(Ding et al. 2018, Fang et al. 2018b, Luo et al. 2018, Mneymneh et al. 2019, Kim et al. 2020, Li et al. 2020, Nath and Behzadan 2020, Neuhausen et al. 2020, Chen and Demachi 2021, Son and Kim 2021)
#2	26	0.988	joint-level vision-based ergonomic assessment	ergonomic assessment of construction workers using deep learning	2019	(Ryu et al. 2019, Yu et al. 2019b, 2019c, 2019a)
#4	18	0.936	Learning construction	Construction site text report monitoring	2018	(Baker et al. 2020, Deng et al. 2020, Li et al. 2020, Chi

			injury precursor	using deep learning		et al. 2013, Zhong et al. 2019)
#6	14	0.982	Deep semantic segmentation	Construction site resource monitoring using deep learning	2020	(Zheng et al. 2020, Wang et al. 2021, Yang et al. 2021, Zhou et al. 2021a)
#11	7	0.974	Industrial construction	Deep learning for construction site safety monitoring	2021	(Fan et al. 2020, Hou et al. 2020, Pereira et al. 2020, Chian et al. 2021, Li et al. 2021, Brilakis et al. 2020)
#13	6	0.979	Deep learning detection	Speech data and productivity analysis in construction industry using deep learning	2020	(Olanrewaju et al. 2020, Hong et al. 2020, Kamal et al. 2020, Liu et al. 2020, Scarpiniti et al. 2021)

Table 2-5 shows the detailed information concerning the identified co-citation cluster, e.g., cluster size, abstract cluster label, year with most publications in the cluster, and the silhouette that reflects

the homogeneity in the cluster. A silhouette of 1, it should be noted, means that the cluster is highly reliable, while a silhouette of 0 means that the cluster is unreliable. The highest silhouette score was achieved by cluster #02 and the lowest was by #0, with scores of 0.988 and 0.787, respectively. The largest cluster of the analysis was #0 with 64 documents labelled as “deep active learning approach” and the smallest was #13 with only 6 documents in the cluster and labelled as “deep-learning detection”, as shown in Table 2-5.

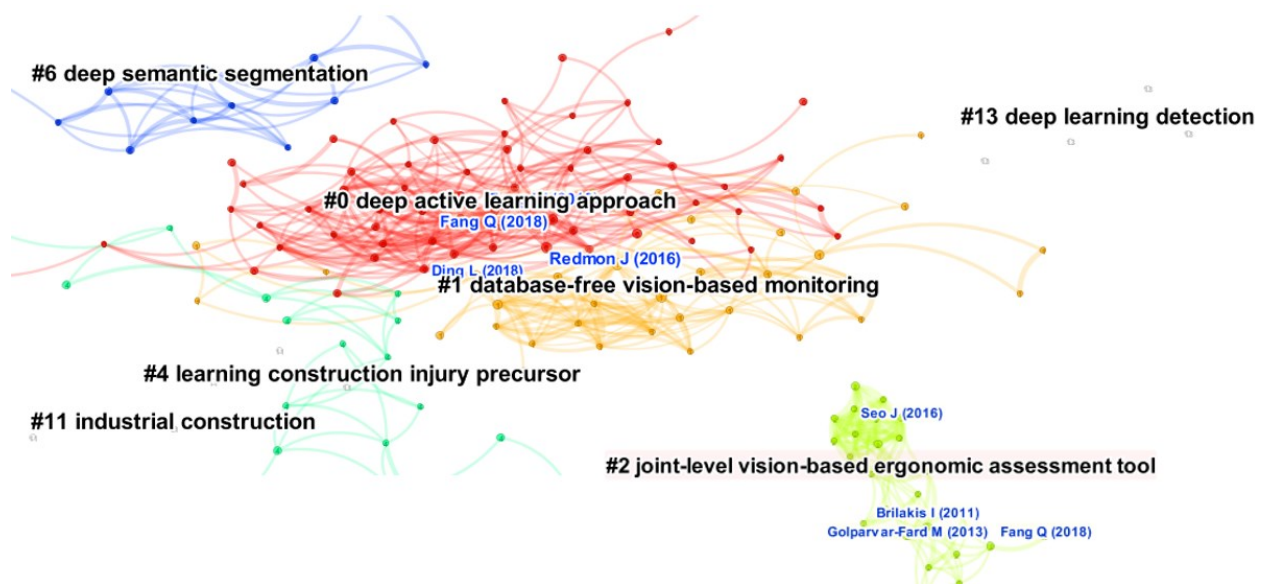


Figure 2-8. Network of document co-citation with clusters.

The timeline analysis of the cluster was also taken into consideration in the review, as shown in Figure 2-9. The timeline analysis identifies the time-period for the development of each cluster. The stretch of cluster #0 with the label “deep active learning approach” is from 2009 to 2021, cluster #1 with the label “Database-free vision-based monitoring” is from 1998 to 2021, cluster #2 with the label “joint-level vision-based ergonomic assessment” is from 1993 to 2021, cluster #4, cluster #6, cluster #11, and cluster #13 with labels “learning construction injury precursor”, “deep semantic segmentation”, “industrial construction”, and “deep-learning detection” is from 1998 to

2017, 2010 to 2021, 2005 to 2019, and 2012 to 2017, respectively. The identified clusters were relabelled based on the collective theme of the papers present in the cluster as given in Table 2-5.

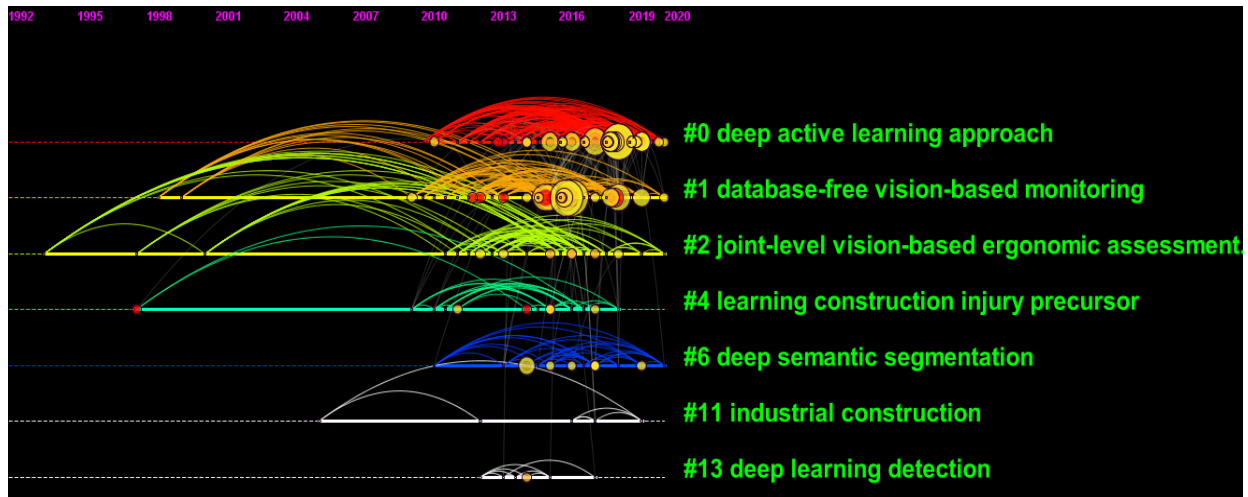


Figure 2-9. Timeline view of clusters.

2.2.3 Critical review of clusters (Phase2)

The critical review of the clusters given in Table 2-5 was conducted to provide an in-depth analysis of the application of deep learning in construction. The cluster identified in the literature was analyzed on the bases of overall research theme and the number of publications relevant to the theme. The detailed information on the research themes is given as follows.

2.2.3.1 Construction machinery monitoring using deep learning

The common link among the articles in this cluster is the concept that uncertainties in the dataset can be automatically evaluated in order to select meaningful learning features and train the model in the sequential process, with these articles centering on the detection, recognition, and tracking of construction machinery (Chen et al. 2020, Guo et al. 2020, Luo et al. 2020, Xiao and Kang 2021a, 2021b, Xuehui et al. 2021) and full pose estimation of construction equipment (Luo et al. 2020) using deep-learning approaches such as CNN, 3D-CNN, Faster region-based CNN (Faster

R-CNN), CNN+ Long short-term memory (LSTM), convolutional (Conv)LSTM, Single-shot detector (SSD), and You Only Look Once (YOLO). It was noted that deep-learning approaches provide the best possible results in detection, recognition, tracking, and full pose estimation of construction machinery on the construction site.

2.2.3.2 Construction site activity and worker monitoring using deep learning

This cluster encompasses the studies that use existing deep-learning models, such as YOLO, CNN, R-CNN, CNN+LSTM, and Faster R-CNN, for the recognition of diverse construction site activities (Luo et al. 2018, Kim et al. 2020, Mansoor et al. 2022), detection of unsafe worker behaviour on the construction site (Ding et al. 2018), detection of workers wearing PPE (e.g., safety helmet, safety waist) (Fang et al. 2018b, Mneymneh et al. 2019, Li et al. 2020, Neuhausen et al. 2020), and performance evaluation of construction workers on site. Each of these deep-learning models has been shown to achieve high accuracy and speed.

2.2.3.3 Ergonomic assessment of construction workers using deep learning

The representative article for this cluster was authored by Yu et al. (2019c), who proposed a tool for assessing the ergonomics of construction workers automatically based on construction videos of workers performing construction tasks. This research used a deep-learning model (i.e., CNN) to extract construction workers' skeletal data for the purpose of full-body ergonomic assessment (Yu et al. 2019c). Other publications in this cluster have used methods such as k -nearest neighbours (KNN), Multilayer perceptron, decision tree, and Support Vector Machine (SVM), to assess physical fatigue (Yu et al. 2019b) and workload estimation (Ryu et al. 2019, Yu et al. 2019a), among other applications.

2.2.3.4 Construction site text report monitoring using deep learning

The representative article for the cluster “construction site text report monitoring using deep learning” was authored by Baker et al. (2020). They used deep-learning models such as CNN, Gated Recurrent Unit (GRU), and Recurrent Neural Network (RNN) to automatically learn injury precursors from raw construction engineering reports by identifying textural patterns, asserting that such an approach can be used to understand, predict, and prevent injury occurrence in construction. Other applications for text monitoring in construction industry described in articles in this cluster include classifying construction safety accident reports automatically (Deng et al. 2020), using fastText-based classification to classify text reports of court compensation cases involving construction accidents (Li et al. 2020), classifying short text containing building quality complaints (Zhong et al. 2019), and text classification to automate job hazard analysis (Chi et al. 2013), to name a few.

2.2.3.5 Construction site resource monitoring using deep learning

The representative article for the “construction site resource monitoring using deep learning” cluster was authored by Wang et al. (2022). They proposed the use of deep learning in conjunction with semantic segmentation in order to visually understand the construction site—meaning to detect and classify/label construction site features such as machinery, material, and workers with the aid of a camera and deep-learning models—in order to improve safety on site. Deep semantic segmentation in deep learning, it should be noted, is a technique used to partition the image into semantically meaningful parts (Wang et al. 2022). This technique is being extensively used in robotics. Specific applications described in this cluster include structural crack detection and segmentation, construction site resource detection (Torok et al. 2014, Cha et al. 2017, Kang et al. 2020, Bang et al. 2020), classification of rock fragments produced by tunnel boring machines

(TBMs) (Yang et al. 2021), as well as segmentation of TBM muck images to estimate the size and shape of rock chips (Zhou et al. 2021a) using tools such as DeepLabV3, ResNET-18, VGG-16, and AlexNet.

2.2.3.6 Deep learning for construction site safety monitoring

The representative article for the “deep learning for construction site safety monitoring” cluster was authored by Pereira et al. (2020), who proposed the use of ANN and simulation-based analytics to allocate and prioritize safety resources in industrial construction as a way of enhancing safety management systems. They noted that, for example, consideration of factors such as awareness training, worker age, and long-term worker retention can be helpful in mitigating safety risks in industrial construction (Pereira et al. 2020). Other articles in this cluster have discussed the use of deep-learning techniques for mitigating safety risks by detecting structural components in a building (Hou et al. 2020, Brilakis et al. 2020, Lu et al. 2019, Bangaru et al. 2021), for structural health monitoring (Cha et al. 2018, Kang et al. 2018, Zhang et al. 2020, Fang et al. 2018a), for classification of waste materials (Davis et al. 2021), and for detecting missing barricades (Chian et al. 2021), with these studies employing deep-learning models such as YOLO, CNN, DCNN, and ANN.

2.2.3.7 Speech data and productivity analysis in the construction industry using deep learning

This cluster revolves around the use of models such as RNN, LSTM, CNN+LSTM, and SSD to classify speech data and productivity on construction sites. The applications described in the articles in this cluster include speech recognition technology for building quantity estimation (Olanrewaju et al. 2020), robotic gripper function on verbal commands (Follini et al. 2018), on-site conversation analysis (Zhang et al. 2018, Scarpiniti et al. 2021), and real-time construction

machinery checking (Xiao and Kang 2019), real-time onsite keyword identification (Mansoor et al. 2022), to name a few.

To summarize the critical review, it was noted that, within the construction engineering domain, deep-learning algorithms have been used in a diverse range of applications, such as detection, activity recognition, tracking and pose estimation of construction machinery, analysis of worker behaviour (e.g., PPE compliance), ergonomic analysis, text data classification, on-site conversation analysis, classification of rock chips produced by TBMs, safety management in industrial construction, construction resource management, construction site surveillance, and many more. As these examples demonstrate, deep learning has proved to be feasible for use in addressing construction site issues of this nature. Nevertheless, deep learning has received relatively little attention in construction practice because its accuracy is hampered by the limited availability of construction site-related data. It was further noted that almost all deep-learning applications in construction have been limited to image-based classification, whereas dynamic video-based classification, text and speech data classification have been very rare in construction, again because of the limited data available.

2.3 Challenges and Future directions

Despite the promising result, several challenges were also observed from the collected studies. This section discusses the identified challenges and future direction of the application of deep learning in the construction sector.

2.3.1 Challenges

The challenges were identified based on analysis of the 423 papers collected from the Scopus database. These challenges are the limited attention from researchers and industrial practitioners

in the area, and limitations of the deep-learning models or other models used for the specific purpose. The identified challenges are discussed in the following sections.

2.3.1.1 Data availability

The first challenge is the lack of data in the construction industry. Deep learning is a data-driven field. To achieve high accuracy in detection and classification, a large volume of label data is required to train the deep-learning models (Khan and Yairi 2018). However, there are few construction engineering-related datasets available to the public, and obtaining construction-related data can be a challenge. The data available to the public is domain-specific, where models are trained for a specific purpose and not necessarily transferrable to other applications. The researcher must therefore create their datasets and label the datapoints manually. This process is time-consuming, expensive, and tedious.

To tackle the issue of data availability, the transfer learning technique was employed. In transfer learning, it should be noted, the models are pre-trained with extensive common object datasets such as ImageNet and MNIST. Examples of the pre-trained model are VGG 16, YOLO series, ResNet, etc. These models have pre-trained weights, where the model has already learned the way to extract the features of the data. Transfer learning technology can assist in decreasing the demand for data volume; however, the lack of data is still a major challenge with respect to applying deep-learning technologies in the construction industry.

2.3.1.2 Accuracy of deep-learning models

The accuracy of the deep-learning model for safety-related and other critical construction site-related tasks is not sufficient for practical applications. This is due to the dynamic nature of the construction site the model trained on the construction site related datasets are lacking in generalization. Most of the data are captured on a single construction site or under the same

environmental condition. On the other hand, every construction site is different. Researchers tackled the data generalization problem by introducing data augmentation techniques. Data augmentation techniques assist in better generalizing and upsizing the dataset (Shorten and Khoshgoftaar 2019). The common data augmentation methods include flipping, cropping, varying the brightness and contrast of images, applying noise and blur to the images, adding noise to the speech data, change pitch of speech data, etc. By applying these transformations, the generalization of deep-learning models can be improved. These transformations helped to improve the accuracy of the deep-learning model (Shorten and Khoshgoftaar 2019) but were not adequate for practical application. Therefore, more efforts should be made to improve the accuracy of the deep-learning models.

2.3.1.3 Construction site Environment

Another challenge is the complicated construction site environment. The construction sites are congested, noisy and dynamic, and this makes the performance use of deep-learning approaches less promising. Complex site conditions also significantly affect the quality of images and other data. To achieve high-quality image data on construction sites more cameras and drones may be required, which help to capture images from every possible environment and angle in the construction sites.

2.3.1.4 Data privacy

Deep-learning models required a massive amount of data to train model the data privacy is one of the major challenges faced when conducting research on deep learning. Meanwhile, any data used for research conducted in Canada is subject to the Personal Information Protection and Electronic Document Act (PIPEDA 2022) and the Copyright Act of Canada (CAC 2022). As such, care had to be taken with respect to data usage and publication rights to ensure compliance.

2.3.1.5 Computational cost

The deep-learning models are complex and require high technology and powerful machines with GPU processors for fast processing (Sergeev and Del 2018). The deployment of powerful machines with the capability of fast processing on every construction site task is not possible. On the other hand, less powerful technological devices such as mobile devices and microcontrollers do not achieve the desired accuracy and speed for deep-learning models. Therefore, more studies should focus on finding approaches with high accuracy and speed on less computational power devices for the full adoption of deep learning in construction.

2.3.2 Future directions

2.3.2.1 Automated data collection and labelling

There are many opportunities in the application of deep learning yet to be explored within the context of construction. As already discussed in the previous section, there are few construction engineering-related public datasets available by which for the researcher to train deep-learning models. Hence, future work should focus on developing more construction engineering-related public datasets and finding alternate methods to train the model when the dataset is insufficient. One recommendation is to install additional cameras and microphones or use drones with capability to collect more images, video and speech data on the construction site. In this way, more datasets can be built, and data can be more generalized if they are captured from real construction sites. Generative Adversarial network (GAN) and data augmentation can also be used to increase the size and better generalize the dataset (Frid-Adar et al. 2018, Mansoor et al. 2022). GAN is being used by researchers to produce synthetic data in many applications, and it can produce more diverse and distinct datasets (Frid-Adar et al. 2018). The implementation of GAN might be a possible solution for the construction engineering-related dataset challenge in future. The labelling

of the dataset is mandatory for classification problems, and this task is typically performed manually by the researcher. This process is time-consuming and tedious, in the future automatic labelling of the construction site datasets should be developed because this can lower the amount of effort required developing high-performance deep-learning models.

2.3.2.2 Dynamic scenario/Scene classification

While deep-learning models have the potential to be leveraged to detect the dynamicity of the construction site, the potential of deep-learning models to detect and classify objects and dynamic construction site environments has yet to be fully explored. To detect and classify objects while moving, multiple deep-learning models need to be integrated to produce accurate results, for example, convolutional neural networks and recurrent neural networks are integrated to achieve spatial-temporal features from the moving objects. The integration of models using the spatial-temporal feature facilitates predictive modelling based on historical data. The integration will assist to detect and classify objects in motion such models will also improve the performance of deep-learning algorithms.

2.3.2.3 Deep learning for text and speech classification in construction industry

Another area warranting the attention of construction researchers is the use of a deep-learning approach to classify text and speech in construction. There have been a few studies published in this area, including such applications as construction site text report monitoring for accidents and claims (Deng et al. 2020, Li et al. 2020b), text classification to automate job hazard analysis (Chi et al. 2013), speech recognition technology for building quantity estimation (Olanrewaju et al. 2020), and on-site conversation analysis (Zhang et al. 2018, Scarpiniti et al. 2021). Deep-learning models are capable of classifying text and speech with high accuracy in a range of environmental conditions, and the use of deep learning in construction for text classification can enhance

understanding of construction reports and support timely analysis of complaints by automating the process. Speech classification in construction can also be useful for improving communication and keeping records of on-site meetings to avoid any misunderstanding in communication. Speech analysis, meanwhile, can be useful for analyzing construction equipment noise and monitoring the health of construction equipment.

2.3.2.4 Optimization of deep learning for low computational power devices

Future research can integrate deep learning with other technologies, such as virtual reality (VR), building information modelling (BIM), mobile devices, Google Glass, etc. Deep learning has traditionally required high computationally powerful machines, but in recent years models with a lower computational burden—e.g., MobileNet, Tensorflow Lite, etc.—have been proposed. These models perform satisfactorily and require less computational power. Using these models with other technologies can assist in decision-making on the construction site. With the help of 5G technology, it is easier than before to collaborate and optimize the model performance in real time. However, current lightweight deep-learning models with low computational power are not reliable for safety-critical task detection and classification due to the compromised accuracy and speed of models. Therefore, an optimized, highly accurate and fast deep-learning structure is still required in the future.

2.4 Conclusion

This study explores the current state-of-the-art research regarding the application of deep learning in construction. It was observed that the implementation of deep-learning methods has gained an increased amount of attention in recent years among researchers in the construction industry. The present study employs a science mapping approach to examine 423 journal articles and conference papers by conducting a scientometric analysis, which was followed by a critical review of the

identified themes. The reviewed literature emphasizes on issues that have been historically addressed by manual means, For example, inspection of construction sites, resources and construction equipment recognition, construction worker safety monitoring, text and speech data analysis. It was observed from research that work in this field is mostly done in isolation; this is particularly the case when considering the research themes and research studies of previous literatures. A conclusion that can be made is that collaboration and support among researchers in the future would be beneficial to increase the debates, dialogues and cross-pollination of ideas for the advancement of the research. Undoubtedly, an enriched understanding that certain practices, mainly the use of deep learning for communication, safety, productivity improvements, and advanced automation of construction processes, may convince the industry to support deeper and more advanced research in the domain, which may help in more funding and careful research planning efforts by policymakers and industry experts. Additionally, this study offers valuable information about research themes published on the topic of deep learning and its implementation in the construction industry through cluster analysis and critical review. Despite the contributions offered, the findings of the study are to be considered in light of certain limitations. As discussed, the findings are restricted by the initial selection of keywords and thus limit the coverage of the current literature. Second, data are retrieved from one single database (Scopus). In the future, a more exhaustive dataset can be obtained for scientometric analysis and critical review by combining data from various databases (e.g., Web of Science, Google Scholar, PubMed, and so on).

Chapter 3: A DEEP-LEARNING CLASSIFICATION FRAMEWORK FOR REDUCING COMMUNICATION ERRORS IN DYNAMIC HAND SIGNALLING FOR CRANE OPERATIONS²

The previous chapter described a detailed scientometric analysis and critical review of the literature on the application of deep learning in the construction industry, in which it was observed that deep-learning models have been shown to be capable of improving productivity, construction site monitoring, safety analysis, and ergonomic assessments in construction. However, the use of deep-learning models for classification of communication (hand signalling and speech) on construction sites has received little attention. Some of the reasons for this are the lack of open-access data, the dynamic nature of the construction work environment, and the questionable accuracy of deep-learning models when deployed in such applications.

The use of hand signals is a common practice in construction sites, especially in crane operations, where a signalman on the ground communicates essential instructions to the crane operator. However, communication errors are prevalent due to factors such as distance, noise, and visibility issues, leading to accidents and project delays. To reduce these errors, there is a need to develop a reliable and efficient framework that can classify hand signals accurately in real time.

In this chapter, a computer-vision-based deep-learning classification framework for reducing communication errors in dynamic hand signalling during crane operations is presented. The framework makes use of an integrated YOLOv4+LSTM model to classify 18 different dynamic hand signals based on their visual features. The framework can detect and classify hand signals in

² A version of this chapter has been published in *Journal of Construction Engineering and Management* as follows: Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., and Al-Hussein, M. (2023). "A deep learning classification framework for reducing communication errors in dynamic hand signaling for crane operation." *Journal of Construction Engineering and Management*, 149(2), 04022167.

real time with an accuracy of 93.5%, thereby improving the efficiency of communication and reducing errors during crane operations.

3.1 Introduction

The success of crane operations is highly dependent on the effectiveness of the communication between the crane operator and the signalman (Zavichi and Behzadan 2011). Hand signalling is the primary means of communication in crane operations. The crane operator operates the crane according to the signalman's hand signals and speech commands, and this is especially critical in blind lifts when the crane operator is unable to see the site. It is often said that the signalman is the "eyes and ears" of the crane operator (Chen et al. 2011, Zekavat and Bernold 2014). However, these means of communication (hand signalling and speech commands) are not always reliable, and breakdowns in communication can result in accidents, as discussed in Chapter 1. There is thus a pressing need for supplemental forms of communication between crane operator and signalman. With recent technological developments, deep-learning algorithms with the ability to classify dynamic hand signals can provide an added layer of communication between signalman and crane operator, making it more efficient and more reliable. In such an application, the deep-learning algorithms receive live camera data, process the data, and output the classification labels. In this way, the construction industry can leverage modern technology in much the same manner as the manufacturing and automobile industries have. For example, road safety has been improved through the introduction of advanced driver assistance systems (ADASs), which provide real-time information about the surrounding environment of a travelling motor vehicle (Pickering et al. 2007; Kukkala et al. 2018).

The outcome of a preliminary feasibility study to classify crane signalman static hand signals using deep learning (see Appendix 1) showed that deep-learning models are capable of classifying static

crane signalman hand signals. However, crane signalman hand signals tend to be dynamic in nature, and thus further investigation is warranted.

This chapter proposes an intelligent deep-learning model, trained with a dataset of dynamic crane signalman hand signals, to achieve a high level of accuracy in classifying dynamic crane signalman hand signals in training, validation, and test datasets, and to further validate the model for real-time crane signalman hand signal classification. The framework is capable of classifying dynamic crane signalman hand signals in real time. The classified hand signals and corresponding labels can assist the crane operator in making sound decisions with confidence as well as allowing the signalman to send commands to the crane operator without the need to hold any other communication device. Given that the signalman must maintain direct eye contact with the load at all times for safety reasons, they cannot look at a handheld device (e.g., tablet) even for a few seconds. The benefit of using the deep learning-based intelligent framework is that it can facilitate communication between the crane operator and the signalman without the need for such devices. The balance of this chapter is organized as follows: The next section provides background knowledge pertaining to hand signal classification. This is followed by an overview of the proposed intelligent framework, including a description of the data collection and data preprocessing techniques. Next is a description of the respective architectures of the deep-learning models used in this study and the modifications made to them, followed by an account of the implementation and performance of the developed models in training, validation, and testing for real-time classification. The chapter closes with a discussion, summary of the conclusions drawn, and a description of possible avenues of future work in this area of research.

3.2 Related work on hand signal classification

Deep learning-based hand gesture recognition is a hot topic in computer vision applications that plays a very important role in human–computer interaction (Su et al. 2020), with most studies in this area focusing on image data recognition due to the excellent performance of deep learning in such applications. Many deep learning-based methods have been developed to classify hand gestures. For example, Oyedotun and Khashman (2017) implemented a convolutional neural network (CNN) and stacked denoising autoencoder-based deep-learning model to recognize 24 hand gestures from American Sign Language (ASL). The model achieved an accuracy of 92.83% for static ASL hand gestures. Huang et al. (2018) used a 3D CNN for video-based sign language recognition without temporal segmentation. Okan et al. (2019) used a lightweight CNN and deep convolutional neural network (DCNN) to classify the hand gestures from EgoGesture and NVIDIA dynamic datasets. The model was found to be capable of achieving 94.04% accuracy in real-time video streaming recognition. Molchanov et al. (2016) introduced the concept of using a recurrent three-dimensional CNN to detect and classify dynamic hand gestures from multi-model data. Their model was trained to predict class labels from in-progress gestures in unsegmented input streams, achieving an accuracy of 88.4%. Many other models have been developed and have been found to perform exceptionally well in recognizing hand gestures.

Although the performance of recently developed methods for hand gesture recognition is promising, their actual implementation in the construction industry is not yet well established. This is due to the fact that the methods reported in the literature have typically been implemented in ideal conditions, i.e., in indoor environments where the signalmen are in static positions while signalling. On the contrary, construction sites are dynamic, complex, and often congested and disorderly. Moreover, the ambient conditions of the outdoor construction environment can also

have an impact on the accuracy of these hand signal interpretation methods. For the proposed intelligent framework, first the custom dataset of dynamic crane signalman hand signals are created. Then, the deep-learning models are developed and tested on the custom dataset. The developed model is validated in real-time classification and compared with the state-of-the-art deep-learning models. The performance of the models is evaluated based on the accuracy and inference time. The performance evaluation will help to analyze the application of deep-learning models as a supplement layer of communication in crane operations in different construction site environments.

3.3 Proposed intelligent framework

The overview of the proposed intelligent framework is shown in Figure 3-1. The framework uses You Only Look Once version 4 (YOLOv4) to extract spatial features (refers to features in static or object vector in a single frame) and the Long Short-Term Memory (LSTM) model to extract temporal features (refers to dynamic changes in features in object based on number of frames taken at different time) from the dataset, thereby improving the accuracy and inference time to facilitate real-time classification. The process underlying the framework includes dataset collection and preprocessing, model training with the dataset, and validation with real-time classification. The output of the framework is visible hand signals on screen with classified labels. These processes are discussed in detail in the following subsections.

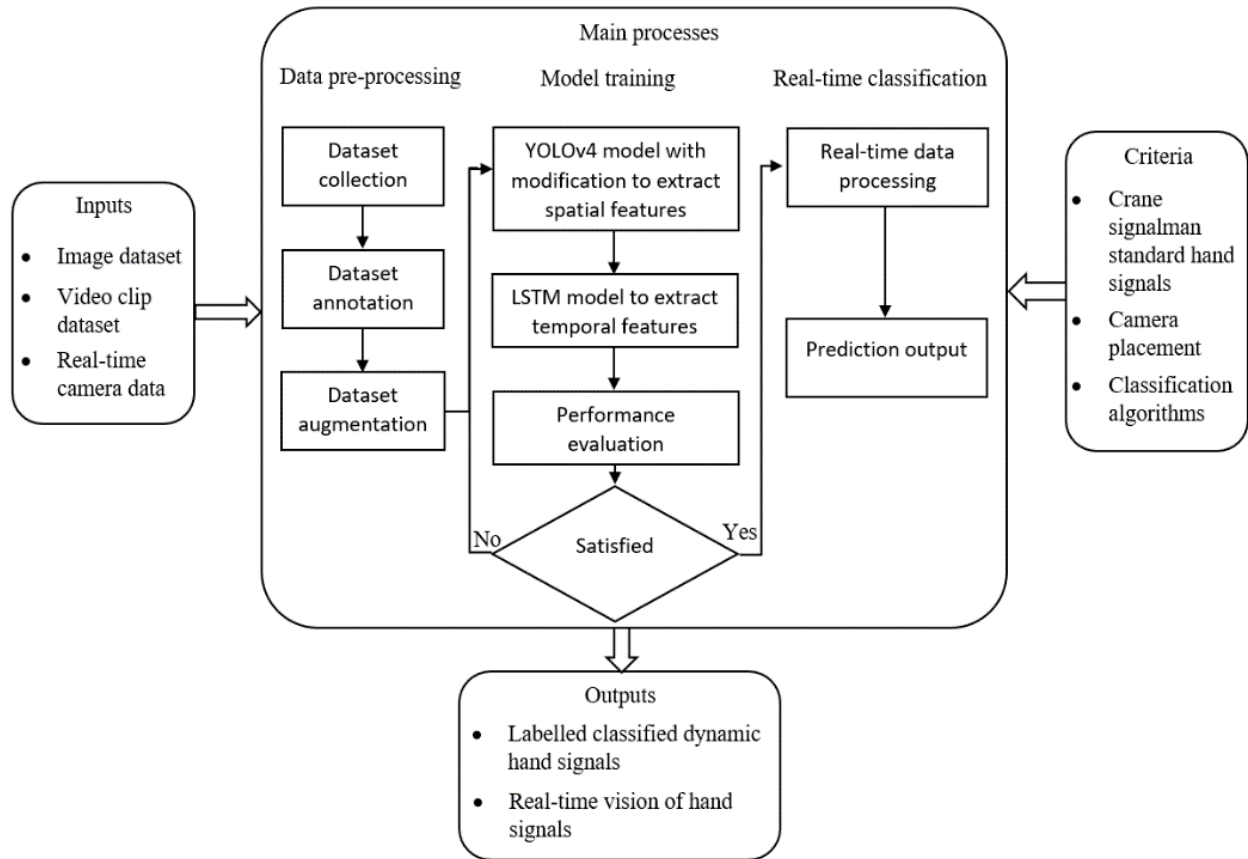


Figure 3-1. Overview of proposed intelligent framework.

3.3.1 Input data collection and preprocessing

Since there is no accessible dataset available for crane signalman hand signals, the priority is to create a video dataset of hand signals. The hand signals to be included in the dataset are based on Occupational Safety and Health Administration (2021) crane hand signals, as shown in Figure 3-2.

To create the dataset, a HD 720p (where p stands for the progressive scan) Logitech camera is used as the recording device. A total of 85 videos from 23 different volunteers are captured with a maximum resolution of 720×720 pixels at 30 frames per second (FPS) containing all 18 hand signals.


















 <p>HOIST With upper arm extended to the side, forearm and index finger pointing straight up, hand and finger make small circles.</p>	 <p>LOWER With arm and index finger pointing down, hand and finger make small circles.</p>	 <p>USE MAIN HOIST A hand taps on top of the head. Then regular signal is given to indicate desired action.</p>	 <p>USE WHIPLINE (Auxiliary Hoist) With arm bent at elbow and forearm vertical, elbow is tapped with other hand. Then regular signal is used to indicate desired action.</p>	 <p>TRAVEL With all fingers pointing up, arm is extended horizontally out and back to make a pushing motion in the direction of travel.</p>	 <p>DOG EVERYTHING Hands held together at waist level.</p>
 <p>BOOM UP With arm extended horizontally to the side, thumb points up with other fingers closed.</p>	 <p>BOOM DOWN With arm extended horizontally to the side, thumb points down with other fingers closed.</p>	 <p>MOVE SLOWLY A hand is placed in front of the hand that is giving the action signal. (Hoist slowly shown in example.)</p>	 <p>SWING With arm extended horizontally, index finger points in direction that boom is to swing.</p>	 <p>TELESCOPE OUT (TELESCOPING BOOMS) With hands to the front at waist level, thumbs point outward with other fingers closed.</p>	 <p>TELESCOPE IN (TELESCOPING BOOMS) With hands to the front at waist level, thumbs point at each other with other fingers closed.</p>
 <p>BOOM DOWN AND RAISE THE LOAD With arm extended horizontally to the side and thumb pointing down, fingers open and close while load movement is desired.</p>	 <p>BOOM UP AND LOWER THE LOAD With arm extended horizontally to the side and thumb pointing up, fingers open and close while load movement is desired.</p>	 <p>STOP With arm extended horizontally to the side, palm down, arm is swung back and forth.</p>	 <p>EMERGENCY STOP With both arms extended horizontally to the side, palms down, arms are swung back and forth.</p>	 <p>TRAVEL (BOTH TRACKS) Rotate fists around each other in front of body; direction of rotation towards body indicates travel forward; rotation away from body indicates travel backward. (For crawler cranes only)</p>	 <p>TRAVEL (ONE TRACK) Indicate track to be locked by raising fist on that side. Rotate other fist in front of body in direction that other track is to travel. (For crawler cranes only)</p>

Figure 3-2. Crane signalman hand signals. (Reprinted from OSHA 2021).

As shown in Figure 3-3, the videos are captured from multiple angles (0° to 360°), and the camera is placed at varying distances (as near as 3 m and as distance as 20 m, and as low as 1.5 m and as high as 30 m), under different weather conditions, and with multiple backgrounds (static versus dynamic), as outlined in Table 3-1.

Table 3-1.

Dataset collection in different scenarios.

Scenario	Camera position (varying distance)	Background	Weather (Winter)
Indoor	Right side of signalman	Static	-
	Left side of signalman	Static	-
	Front of signalman	Static	-
	Back of signalman	Static	-
	Helmet of signalman	Static	-
Outdoor	Right side of signalman	Static and dynamic	Sunny and cloudy
	Left side of signalman	Static and dynamic	Sunny and cloudy
	Front of signalman	Static and dynamic	Sunny and cloudy
	Back of signalman	Static and dynamic	Sunny and cloudy
	Helmet of signalman	Static and dynamic	Sunny and cloudy

It is observed from the videos that the crane signalman takes approximately 3 s to complete a cycle of a hand signal in most cases. Based on this observation, short video clips with a maximum duration of 3s are extracted from the collected videos. These videos clips are then separated based on their classification labels into separate folders corresponding to the 18 different hand signals. Table 3-2 shows the total number of video clips extracted for each crane signalman hand signal. Further data augmentation techniques are applied to the extracted video clips, as described in the following subsection.

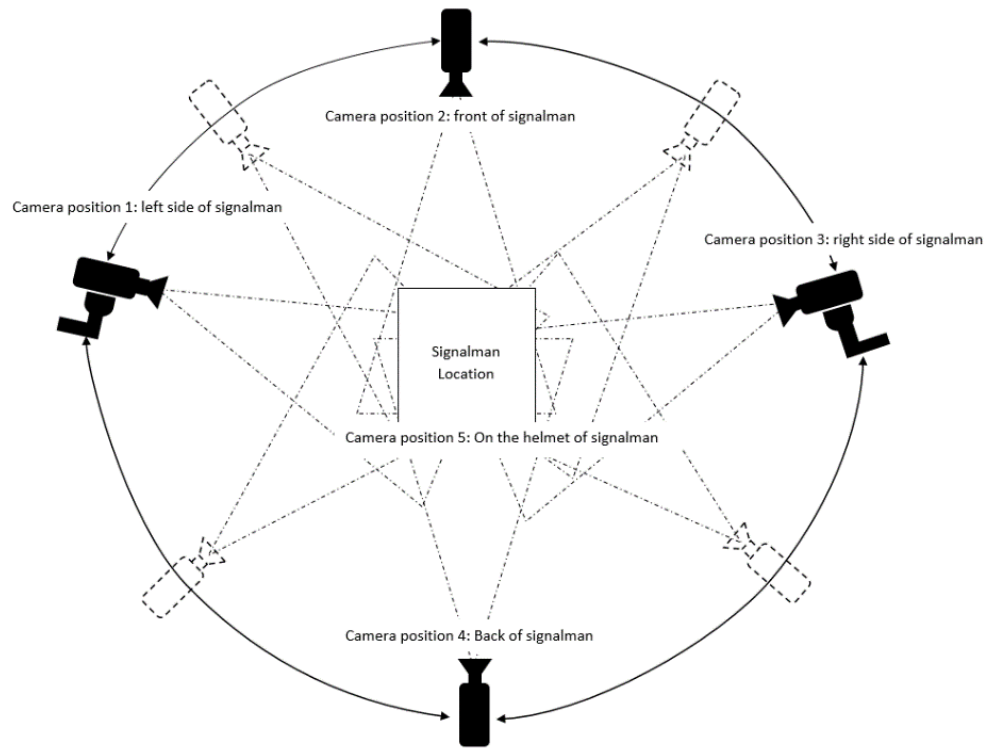


Figure 3-3. Camera placement for data collection and real-time crane signalman hand signal classification.

Data augmentation is a method to increase the size of the collected dataset artificially, by applying random but realistic transformation to each frame in the video clip. This technique helps to diversify the frames in the video dataset. The implementation of data augmentation can improve the performance of the model (Hernandez-Garcia and Konig 2018), reduce the likelihood of overfitting, as well as better generalize the dataset for model training purposes (Mikołajczyk and Grochowski 2018). The data augmentation techniques used for this research includes gamma transformation, Gaussian blur and salt-and-pepper noise. These augmentation techniques are most widely used technique because they are closest to mimic out-of-focus camera blur, changing weather conditions, and dusty conditions on the construction site. The use of these augmentation

techniques proved the increase of accuracy of the deep-learning model. Detail analysis of augmentation technique used in this chapter along with their effect on accuracy of deep-learning model are given in Appendix 2.

3.3.1.1 Gamma transformation

Gamma transformation is a technique to control the intensity of light in a given frame. With gamma transformation, we can adjust the contrast and brightness of the original frame (Bull 2014). This technique can bring more originality to the dataset and help the model to perform better in any lighting condition. For the current dataset, the range of gamma values in which to adjust the brightness and contrast is set to between -7 and $+7$, where a setting of -7 will result in darker frames and $+7$ will result in lighter frames. This range is chosen to make the conditions more realistic and representative of weather conditions ranging from sunny/clear to cloudy/overcast.

3.3.1.2 Gaussian blur

Gaussian blur is used to blur the frames in the video dataset; it has been widely used for multiple purposes such as reducing the detail in the frame. This technique smooths out the randomness in a frame based on a chosen blur radius. Each pixel in the frame will adopt a new value based on the weighted average of its surrounding pixels (Misra and Wu 2019), where more weight is given to pixels in closer proximity. The amount of blur in the frame is measured in pixels (px), where a higher value of px means more blur and a lower value means less blur will appear in the frame. The camera blur happens in moving objects in real time. Therefore, this transformation is used in this research. To add blur to the dataset, a blur value of 2.75 px is chosen for the present work. Adding blur to the frames in the dataset ensures that the model will be trained for a scenario where the camera may lose focus, thereby allowing the model to achieve better accuracy.

3.3.1.3 Salt-and-pepper noise

The “salt-and-pepper” noise data augmentation technique is used for frame degradation. In this technique, some pixels of the frame are kept very noisy. The effect is likened to sprinkling salt and pepper on the frame (Boncelet 2009), hence the name. The intensity of salt-and-pepper noise is measured in percentage, where 0% means there is no noise in the frame. For the current research, 3% salt-and-pepper noise is applied. The value of 3% is chosen based on the observations and considering the complexity of the construction-site scenarios. This helps the model to be trained for unintentional and unwanted changes in the scenes. It also helps the model to correctly classify hand signals captured from low-quality cameras (low-resolution pixels). Some examples data augmentation techniques is shown in Figure 3-4.



Figure 3-4. Example of data augmentation.

The reason for using these augmentation techniques is to better generalize the training dataset and to mimic the construction site environment, which is typically subject to changing weather, dusty conditions, among other dynamic and unpredictable conditions. After applying these transformations, the dataset size is increased by 25%, as the data augmentation randomly choose each frame and apply transformation and store the frame as a different frame. Table 3-2 shows the

number of video clips for each hand signal included in the dataset with and without data augmentation.

Table 3-2.

Number of video clips (duration = 3s) with and without data augmentation.

Standard hand signals	Dataset sample size without augmentation (90 frames in each video clip)	Dataset sample size with augmentation (90 frames in each video clip)
Hoist	187	231
Lower	180	220
Use main hoist	173	223
Use whipline	187	234
Boom up	184	229
Boom down	178	225
Move slowly	170	218
Swing	168	208
Boom down and raise the load	187	220
Boom up and lower the load	190	235
Stop	181	224
Emergency stop	189	233
Travel	177	229
Dog everything	178	225
Travel both tracks	167	216
Travel one track	176	224

Telescope out	181	222
Telescope in	187	238

3.3.2 The main process of the intelligent framework

The main process of the proposed intelligent framework consists of three parts, namely, data preprocessing, model training, and real-time classification as shown in Figure 3-1. The data preprocessing section includes data collection, annotation, and augmentation, which is discussed in detail in the previous section. In the model training part, YOLOv4 and LSTM deep-learning models are employed as the classification model that will be discussed in coming section. The classification model is trained and validated using the collected hand signal dataset and the performance of the model is evaluated. A satisfactory performance means the accuracy and speed of the classification model exceed the user-defined parameters. The model network and weights are saved and used for the real-time classification of hand signals. The output of the framework is divided into two parts, namely, real-time classification of dynamic hand signals and real-time visualization of hand signals, which will be discussed in the next section.

3.4 Deep-learning algorithms

3.4.1 YOLOv4 architecture

YOLOv4 is a deep learning-based object detection and classification model capable of inspecting an image to find the subset of object class, enclose it within bounding boxes, and identify its class when trained (Bochkovskiy et al. 2020). YOLOv4 is the updated version of the YOLO algorithm series, which has been used in many applications in various fields, such as in the classification of safety helmets (Hu et al. 2019), construction vehicles (Hou et al. 2020), defects in sewer pipes (Yin et al. 2020), etc. The YOLOv4 model is selected for the present work based on its superior

accuracy and speed compared to other deep-learning models (Bochkovskiy et al. 2020). For example, when the YOLOv4 model is compared with the other versions of the YOLO series the YOLOv4 outperform the other version of the YOLO series (Nepal and Eslamiat 2022; Rahman et al. 2021; Bochkovskiy et al. 2020). The YOLOv4 model also showed significantly higher accuracy and speed compared to other famous detection and classification models such as Single-Shot Detector (SSD) and Faster R-CNN (Kim et al. 2020). The YOLOv4 model architecture consists of three parts—the backbone, neck, and head—as shown in Figure 3-5.

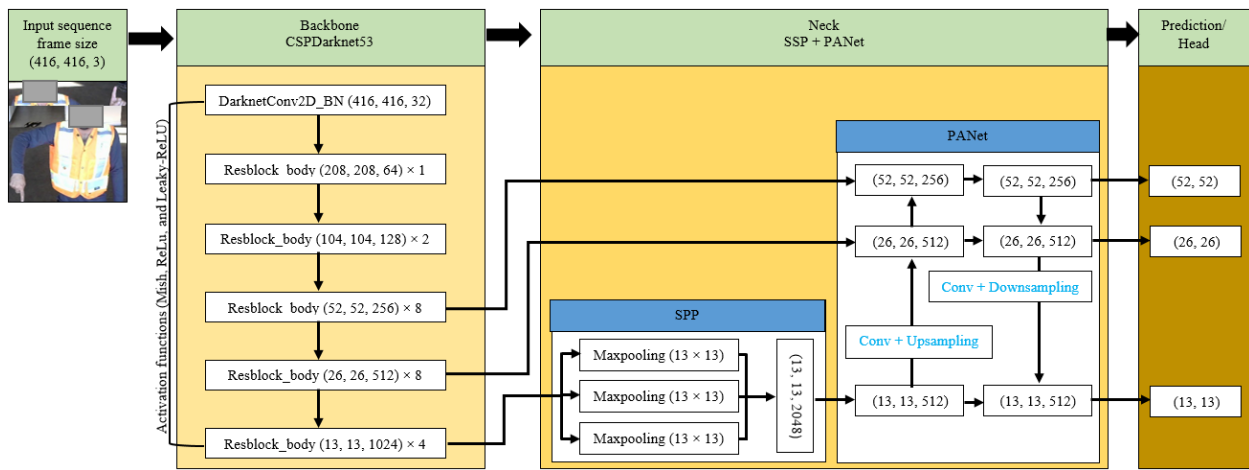


Figure 3-5. Detail architecture of proposed YOLOv4 model.

3.4.1.1 Backbone

The backbone of the YOLOv4 architecture, used in the research is composed of a cross-stage partial connection CSPDarknet53, Bag of freebies, and bag of specials. CSPDarknet53, it should be noted, is a CNN featuring residual connections to ensure the depth of the network and to address the vanishing gradient problem. Moreover, CSPDarknet53 has the capability to learn spatial features from the crane signalman hand signal dataset. Based on previous studies it has been noted that CSPDarknet53 performs better than other available models in object classification when using the Common Objects in Context (COCO) dataset, while the bag of freebies is employed to improve

the performance of the network without affecting inference time in classification problems (Bochkovskiy et al. 2020). Examples of freebies are data augmentation to increase the size of the dataset and minimize the risk of model overfitting, or editing the complete intersection over union (CIoU) loss, which predicts bounding box overlap with the ground truth bounding box and encourages the network to pull over the predicted box to the ground truth box (Zheng et al. 2019). The bag of specials is used to increase the inference time and performance of the network (Bochkovskiy et al. 2020). Bag of specials that are used in the proposed model are cross minibatch normalization to make the network capable of running on any GPU system (Bochkovskiy et al. 2020); dropblock regularization is also employed for the study, which forces the network to learn features that it otherwise may not capture (Ghiasi et al. 2018); and the Mish activation function, which moves the network feature creation towards its optimal point (Misra 2019). These improvements effectively improved the detection and classification speed and accuracy on the COCO dataset (Bochkovskiy et al. 2020). The detailed working philosophy and architecture of YOLOv4 and be found in the article by (Bochkovskiy et al. 2020).

3.4.1.2 Neck

The neck of YOLOv4 model architecture used in this study is responsible for collecting the feature maps from crane signalman hand signal at different stages in the network, and it comprises a spatial pyramid pooling (SPP) block (Huang et al. 2020) and a path aggregation network (PANet) (Liu et al. 2018). The SPP block used in the study is responsible for separating out the most important feature maps generated by different filters in the backbone of the developed YOLOv4 architecture. This is achieved with the help of pooling, which is performed using kernel sizes of 5, 9, and 13 and a stride of 1 while the padding is adjusted; in this manner, it can generate an output with the same shape as that of the input. The PANet used in the proposed YOLOv4 architecture serves to

concatenate the input and the vector from a previous layer to create a new vector. It thereby improves the propagation of layer information in the network from up to down and from down to up (Liu et al. 2018).

3.4.1.3 Head/ Dense prediction

The head of the YOLOv4 network architecture is responsible for performing the final dense prediction of crane signalman hand signal. It should be noted that the proposed YOLOv4 network features the same head as YOLOv3, as it is capable of multiscale object prediction. The prediction output is composed of a vector containing coordinates of the predicted bounding box (center, height, width) and the classified predicted label of crane signalman hand signals. The detailed architecture is given in Figure 3-5.

3.4.2 Modification in YOLOv4 model architecture

The activation function plays a key role in the performance of deep-learning models, especially when the dataset availability is limited (Ramachandran et al. 2017). It is used to determine the activation of neurons in the neural network model. As such, the selection of activation function plays a critical role in the performance of the model. The most widely used activation functions for training deep-learning models efficiently and improving the accuracy of predictions in the case of a limited custom dataset are selected for the proposed framework, which includes ReLU (Agarap 2018), Mish (Misra 2019), and Leaky-ReLU (Xu et al. 2020). In the proposed framework, the experiments are conducted by modifying the YOLOv4 backbone, in this case by changing the activation function and the performance of each proposed model is evaluated by using all the discussed activation functions one by one in the backbone of the YOLOv4 network architecture.

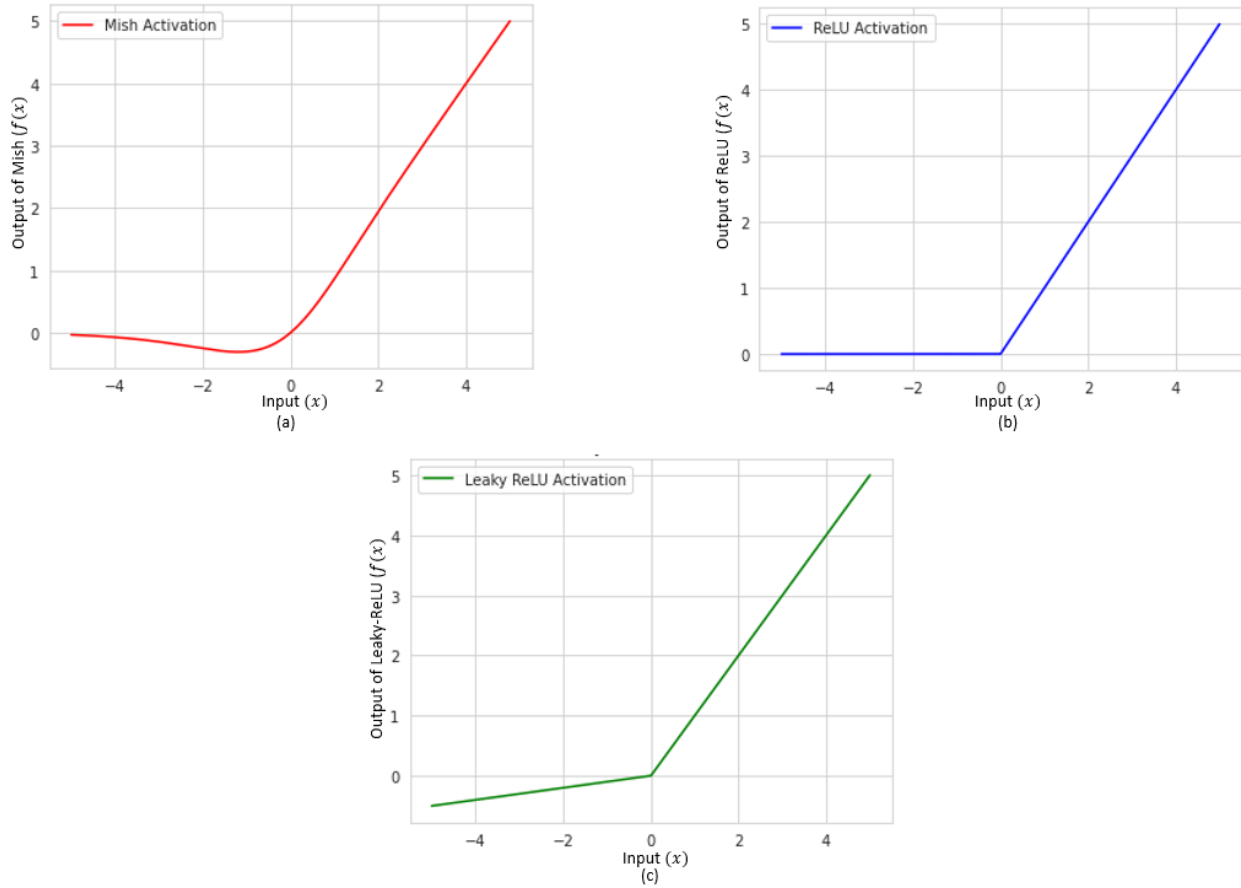


Figure 3-6. Graphical representation of activation functions: (a) Mish; (b) ReLU; and (c) Leaky-ReLU.

The Mish activation function Eq. (3.1) can improve the performance of the model; however, it requires a long training time and has a high computational cost. It has a range of $[\approx -0.31, \infty]$, which means it is bounded below and unbounded above. This property is important to improve the performance of the model training. A graphical representation of this function is given in Figure 3-6.

$$f(x) = x \tanh(\ln(1 + e^x)) \quad (3.1)$$

$$f(x) = \max(0, x) \quad (3.2)$$

$$f(x) = \begin{cases} x, & \text{for } x \geq 0 \\ \alpha x, & \text{for } x < 0 \end{cases} \quad (3.3)$$

The ReLU activation function Eq. (3.2), meanwhile, is an identity line where $y = x$ for all positive values and $y = 0$ for all negative values, as shown in Figure 3-6. Leaky-ReLU Eq. (3.3), finally, is one of the most widely used activation functions due to its good performance. This activation function has an alpha parameter that helps to keep the gradient away from zero all the time while the model is being trained. (A gradient of zero means that, for activations in that region, the weights are not being updated during back-propagation. This can create “dead” neurons that never become activated, thereby limiting the model and inhibiting further performance improvement.) It also solves the vanishing gradient problem and keeps updating the weights along the propagation process. The graphical representation of this function is given in Figure 3-6.

3.4.3 LSTM architecture

As discussed, that crane signalman uses dynamic hand signals while sending message to the crane operator. To accurately classify dynamic hand signals temporal features over time plays an important role. This study used LSTM to extract temporal features from dynamic crane signalman hand signals. The LSTM model, it should be noted, is a special type of RNN capable of learning information over a long period of time (Greff et al. 2016), and it can be employed to extract temporal information from video frames (Qasim and Pettirsch 2020). It converts the video frames into a sequence of feature vectors that accurately present the features from each video frame. Temporal feature extraction is an important tool for classifying objects in a dynamic image sequence, and temporal features are used to monitor dynamic changes in crane signalman hand signals. The LSTM architecture, meanwhile, is composed of hidden state, a memory cell, an input

gate, an output gate, and a “forget” gate, as shown in Figure 3-7, and described in greater detail below.

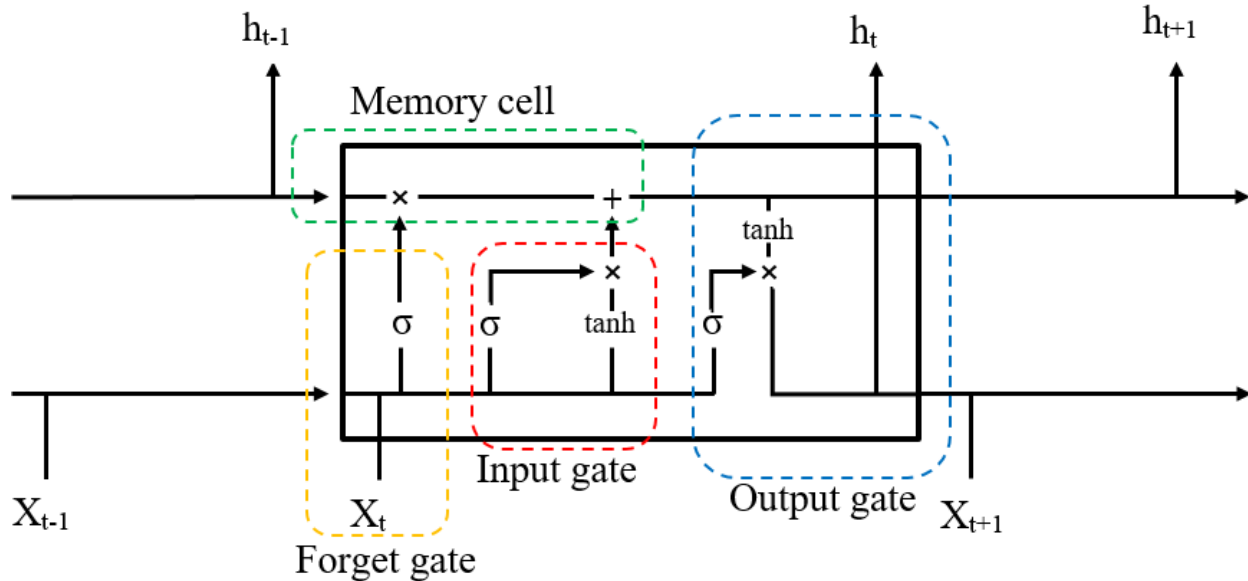


Figure 3-7. Architecture of LSTM model.

In the LSTM model, the hidden state is responsible to hold the previous information of the crane signalman data in the form of a vector from the YOLOv4 model for each video frame. In Figure 3-7, X_{t-1} represent the hidden state of the past timestamp and X_t is the hidden state of the current timestamp. The memory cell in the LSTM model is used to store the input data from YOLOv4 model in the form of a vector for each video frame of the crane signalman hand signal dataset. This cell also deals with the flow of information and handling of input data. The input vector from YOLOv4 model enter through the input gate of the LSTM model. The input gate is responsible to decide whether to allow the input data in or to remove the present state of the input data with the aid of a hyperbolic tangent activation function that updates the cell state. The output gate used in the LSTM model is responsible for filtering and regulating the output of the function, while, finally, the forget gate is used to discard the stored information that is no longer required. In this

way, LSTM model helps to extract, store, and learn the temporal features in a given time period for the sequence of frames from the crane signalman hand signal dataset.

3.5 The architecture of the integrated model

Crane signalmen use dynamic hand signals to send commands to the crane operator. To achieve high accuracy in dynamic hand signal classification, both spatial and temporal features of the dataset are required. The framework uses the proposed modified YOLOv4 model and LSTM model to extract both spatial and temporal features from the crane signalman hand signal dataset and improve the framework's accuracy. Here, the modified YOLOv4 is responsible for detecting and extracting the spatial features from crane signalman hand signal dataset, and LSTM is employed to use the temporal features and predict the final output. The integration of the improved YOLOv4 with LSTM (as shown in Figure 3-8) allows us to exploit the spatiotemporal features in the data and improve the accuracy of the model.

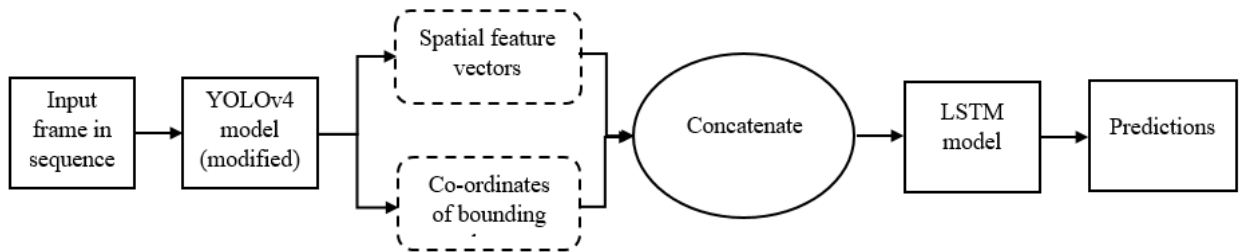


Figure 3-8. Proposed YOLOv4+LSTM model architecture.

The model receives the sequence of labelled frames from video clips as input. The labelled input frames are then fed into the modified YOLOv4 model, which processes the input frames to produce a spatial feature vector of each frame in sequence and coordinates of the predicted bounding box location, which is the detection box. The output information (spatial feature vector and the coordinates of the bounding box location) is then concatenated and, in turn, fed into the LSTM model. The LSTM model explores the information, learns the historical pattern based on the

sequence of frames, and, finally, generates the detection box coordinates, along with the classified label frame by frame, as a prediction output of the model.

3.6 Case Study

The proposed deep-learning algorithms are used to classify the dynamic crane signalman hand signals. The dataset of 4,054 video clips is used to train, validate, and test the classification accuracy of the deep-learning algorithm. The dataset contains 18 different hand signals labelled manually. The details concerning the data collection and preprocessing of the hand signal dataset as described below are provided in previous sections.

3.6.1 Model Training

Before the model can be used for real-time classification of crane signalman hand signals, it must be trained on the custom hand signal dataset. The dataset is split into training (70%), validation (15%), and test (15%) sets. The hyperparameters to train the model are determined by trial and error to achieve the best performance, which are given as follows, the spatial size of the frames, is reduced from 720×720 pixels/frame to 416×416 pixels/frame manually for the purpose of inputting them to the modified YOLOv4+LSTM model. The optimization algorithm employed in the model are Nesterov accelerated gradient, momentum, decay and dropout layers. A batch size of 64, momentum of 0.9, and decay of 0.0005 are used in the models, which are trained on 3,600 epochs. The highest training and validation accuracy is achieved by YOLOv4(Leaky-ReLU)+LSTM model with 96.9% and 96.02%, respectively. The lowest accuracy in model training and validation is achieved by YOLOv4 (Mish)+LSTM with 93.4% and 91.3%, respectively. The model accuracy is measured using Eq. (3.4).

$$Accuracy = \frac{\text{Correctly classified crane signalman hand signal of each class}}{\text{Total number of crane signalman hand signal in each class}} \times 100 \quad (3.4)$$

In the case of training and validation loss, the YOLOv4(Leaky-ReLU)+LSTM models are found to perform better than the YOLOv4 (Mish)+LSTM and YOLOv4 (ReLU)+LSTM models, achieving values of 0.038 and 0.062 in training and validation losses, respectively, as given in Table 3-3.

Table 3-3.

Training and validation accuracy and loss.

Models	Training accuracy	Validation accuracy	Training loss	Validation loss
YOLOv4(Mish)+LSTM	93.4	91.30	0.143	0.226
YOLOv4(ReLU)+LSTM	93.7	92.56	0.105	0.166
YOLOv4(Leaky-ReLU)+LSTM	96.9	96.02	0.038	0.062

The graphs given in Figure 3-9, meanwhile, show the training and validation accuracies and training and validation losses after each epoch. From Figure 3-9, it can be noted that the accuracy of the models is improving after each epoch and the value of loss is reducing accordingly to the minimum possible value. This Figure also illustrates the importance of the model training to the maximum number of epochs, results in improved accuracy of the model. The graphs in Figure 3-9, also demonstrate the progression of the model whereby the model has learned to respond appropriately to a set of training patterns within a specified margin of error. The output of the model draws closer and closer to a specific value close to the minimum. The model loss is

measured using Eq. (3.5). The detail about the loss function is found in the article by (Zheng et al. 2019).

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \quad (3.5)$$

$$IoU = \frac{|B \cap B^{gt}|}{|B \cup B^{gt}|} \quad (3.6)$$

$$\alpha = \frac{v}{(1 - IoU) + v} \quad (3.7)$$

$$v = \frac{4}{\pi^2} \left(\arctan \frac{w^{gt}}{h^{gt}} - \arctan \frac{w}{h} \right)^2 \quad (3.8)$$

“Here b and b^{gt} represent the centroid of B and B^{gt} given in Eq. (3.6). c shows the diagonal length of the smallest bounded box covering prediction and ground truth bounding boxes. $\rho(\cdot)$ is the Euclidean distance, α is a positive trade-off parameter Eq. (3.7), v computes the consistency of aspect ratio Eq. (3.8). $B_{gt} = (x_{gt}, y_{gt}, w_{gt}, h_{gt})$ is the centroid coordinate, width, and height of ground truth bounding box, and $B = (x, y, w, h)$ is the centroid coordinate, width, and height of the prediction bounding box. w_{gt} is the width of the ground truth bounding box, w is the width of the prediction bounding box. h_{gt} is the height of the ground truth bounding box, h is the height of the prediction bounding box” (Zheng et al. 2019).

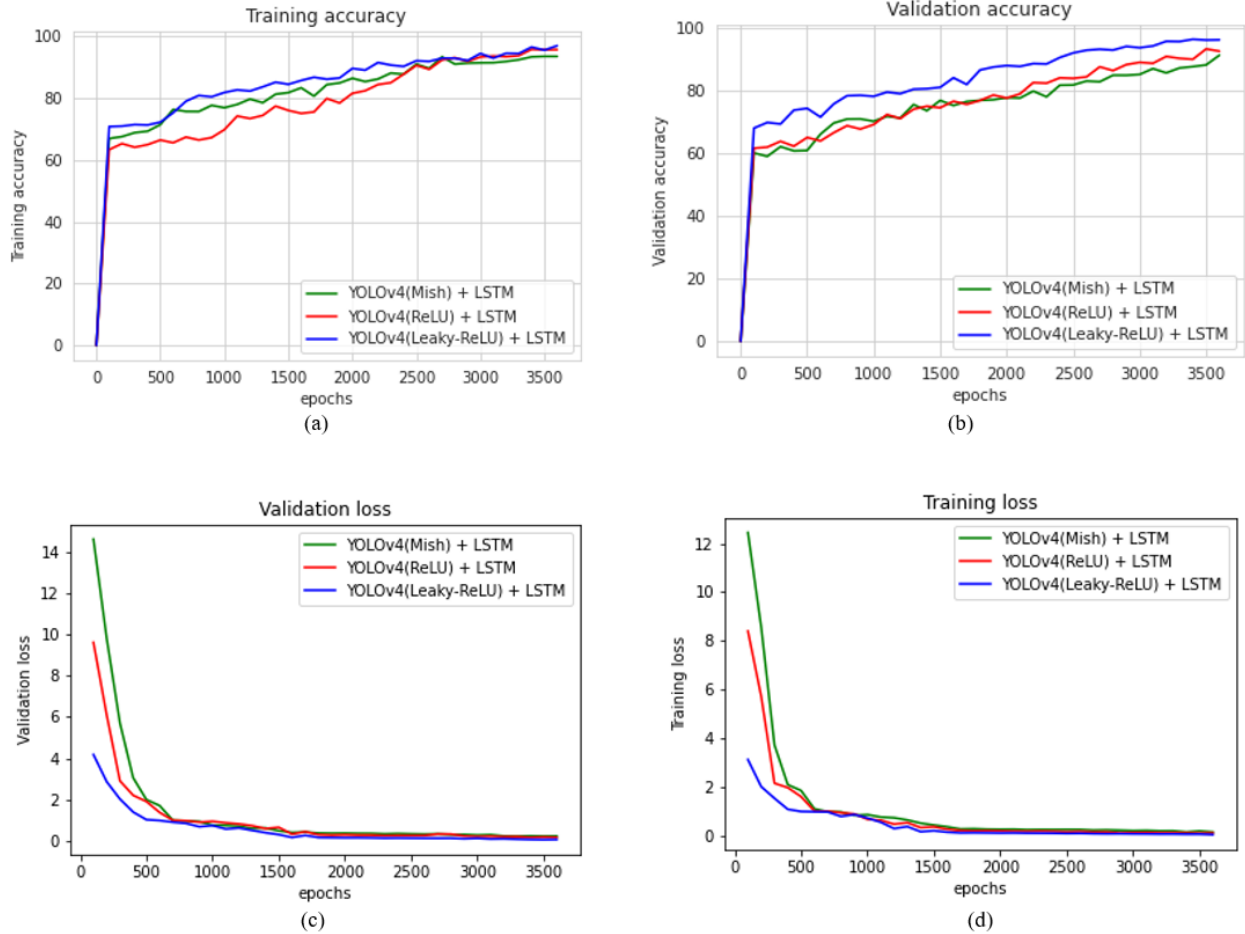


Figure 3-9. (a) Training accuracy; (b) validation accuracy; (c) validation loss; and (d) training loss of deep-learning models.

3.6.2 Model Performance evaluation on the test dataset

The performance of the model is further evaluated based on the confusion matrix, classification accuracy, and inference time obtained for the test dataset. A confusion matrix, it should be noted, is a table used to summarize the performance of a classification model. The rows in a confusion matrix represent the predicted class, while the columns represent the actual classes. The classification accuracy represents the percentage of samples correctly classified by the classification model. The inference time, meanwhile, refers to the time used by a pre-trained model

to make the prediction. The prediction time, finally, although it may differ with different computer hardware configurations, provides an idea of how fast the classification could be made.

		Predicted Class																	
		Hoist	Lower	Use main hoist	Use whipline	Boom up	Boom down	Move slowly	Swing	Boom down and raise the load	Boom up and lower the load	Stop	Emergency stop	Travel	Dog everything	Travel both tracks	Travel one track	Telescope out	Telescope in
Actual class	Hoist	85.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Lower	0	89.2	0	0	1.1	0	3.1	0	0	0	0	0	0	0	0	0	0	0
	Use main hoist	0	0	91.4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Use whipline	0	0	5.9	90.1	0	0	0	0	0	0	0	0	0	10.8	0	0	1.7	0
	Boom up	4.5	0	0	0	92.7	0	0	0	0	0	8.1	0	0	0	0	0	0	3.6
	Boom down	0	6.7	0	0	0	94.5	0	0	3.7	0	0	0	0	0	0	0	0	0
	Move slowly	10.1	0	0	0	2.1	0	83.8	0	0	0	0	0	0	0	8.9	0	0	0
	Swing	0	0	0	0	0	0	0	89.1	0	0	0.4	1.3	0	0	0	0	0	0
	Boom down and raise the load	0	0	0	0	0	5.5	0	0	93.4	0	0	0	0	0	0	0	0	6.7
	Boom up and lower the load	0	0	0	0	4.1	0	0	0	0	91.1	0	0	0	0	0	0	0	0
	Stop	0	0	0	0	0	0	0	10.9	2.9	0	96.4	6.1	0	0	0	0	0	0
	Emergency stop	0	0	0	0	0	0	0	0	0	3.2	92.6	0	0	0	0	0	0	0
	Travel	0	0	0	0	0	0	3.9	0	0	0	0	0	87.4	0	0	11.2	0	0
	Dog everything	0	0	0	0	0	0	0	0	0	0	0	0	0	97.8	0.9	0	0	0
	Travel both tracks	0	0	0	9.9	0	0	9.1	0	0	0	0	0	0	2.2	90.1	0	3.6	2.1
	Travel one track	0	0	2.6	0	0	0	0	0	0	0	0	0	1.8	0	0	87.1	0	0
	Telescope out	0	4.1	0	0	0	0	0	0	0	0.8	0	0	0	0	0	0	92.8	0
	Telescope in	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91.2

Figure 3-10. Confusion matrix of YOLOv4 (Mish)+LSTM on the test dataset.

		Predicted Class																	
		Hoist	Lower	Use main hoist	Use whipline	Boom up	Boom down	Move slowly	Swing	Boom down and raise the load	Boom up and lower the load	Stop	Emergency stop	Travel	Dog everything	Travel both tracks	Travel one track	Telescope out	Telescope in
Actual class	Hoist	91.7	0	3.4	1.6	0	0	3.5	0	0	0	0	0	0	0	0	4.9	0	0
	Lower	1.1	90.2	0	0	0	0.5	0	0	1.7	0	0	0	0	0	0	0	0	0
	Use main hoist	0	0	92.6	0	0	0	1.3	0	0	0	0	0	0	0	0	0	0	0
	Use whipline	0.7	0	0	93.7	2.6	0	0	2	0	0	0	0	1.2	0	0	1.9	0	0
	Boom up	1.3	0	0	0	94.5	0	0	0	0	5.1	0	0	0	0	0	0	6.8	3.1
	Boom down	0	5.8	0	0	0	96.4	0	0	10.1	0	0	0	0	0	0	0	0	2.1
	Move slowly	2.1	0	0	2.1	0	0	91.5	0	0	0	0	0	0	0	7.5	0	0	0
	Swing	0	0	0	0	0	0	0	90.1	0	0	0.8	0	0	0	0	0	0	0
	Boom down and raise the load	0	2.8	0	0	0	3.1	0	0	88.1	0	0	0	0	0	0	0	0	0
	Boom up and lower the load	0	0	0	0	2.9	0	0	0	0	94.2	0	0	0	0	0	0	0	0
	Stop	0	0	0	0	0	0	0	5.4	0	0	96.1	1.3	0	0	0	0	0	0
	Emergency stop	0	0	0	0	0	0	0	2.5	0	3.1	98.7	0	0	0	0	0	0	0
	Travel	1.8	0	0	2.6	0	0	0	0	0	0	0	95.7	0	0	0	0	0	0
	Dog everything	0	0	0	0	0	0	0	0	0	0	0	0	0	94.9	4.8	0	0	3.6
	Travel both tracks	0	1.1	0	0	0	0	3.7	0	0	0	0	0	5.1	87.7	0	0	3.5	0
	Travel one track	1.3	0	3.9	0	0	0	0	0	0	0	0	0	3.1	0	0	93.1	0	0
	Telescope out	0	0	0	0	0	0	0	0	0	0.7	0	0	0	0	0	0	89.7	0
	Telescope in	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91.2

Figure 3-11. Confusion matrix of YOLOv4 (ReLU)+LSTM on the test dataset.

		Predicted Class																	
		Hoist	Lower	Use main hoist	Use whipline	Boom up	Boom down	Move slowly	Swing	Boom down and raise the load	Boom up and lower the load	Stop	Emergency stop	Travel	Dog everything	Travel both tracks	Travel one track	Telescope out	Telescope in
Actual class	Hoist	92.8	0	3.4	0	0	0	0	0	0	0	0	0	0	0	0	1.4	0	0
	Lower	0	94.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Use main hoist	4.3	0	91.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Use whipline	1.3	0	0	95.8	0	0	0	0	0	0	0	0	2.9	0	0	7.6	0	0
	Boom up	0	0	0	0	98.2	0	0	0	0	3.5	0	0	0	0	0	0	0	0
	Boom down	0	5.9	0	0	0	98.7	0	2.2	0	0	0	0	0	0	0	0	0	4.5
	Move slowly	1.6	0	0	0	0	0	96.5	0	0	0	0	0	0	0	0	5.9	0	0
	Swing	0	0	0	0	0	0	0	92.5	0	0	0.9	0	0	0	0	0	4.1	1.1
	Boom down and raise the load	0	0	0	0	0	1.3	0	1.2	97.8	0	0	0	0	0	0	0	0	1.3
	Boom up and lower the load	0	0	0	0	1.8	0	0	0.6	0	96.5	0	0	0	0	0	0	0	1.7
	Stop	0	0	0	0	0	0	0	3.2	0	0	99.1	0	0	0	0	0	3	1.1
	Emergency stop	0	0	0	0	0	0	0	2.5	0	0	0	100	0	0	0	0	0	1.3
	Travel	0	0	0	0	0	0	0	0	0	0	0	0	93.1	0	0	2.2	0	0
	Dog everything	0	0	0	0	0	0	0	0	0	0	0	0	0	100	4.4	0	0	0
	Travel both tracks	0	0	0	0	0	0	3.4	0	0	0	0	0	0	0	89.7	0	0	0
	Travel one track	0	0	4.7	4.2	0	0	0	0	0	0	0	0	3.9	0	0	88.8	0	0
	Telescope out	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	91.2	0
	Telescope in	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	90.7

Figure 3-12. Confusion matrix of YOLOv4(Leaky-ReLU)+LSTM on the test dataset.

Figure 3-10, represents the confusion matrix of the YOLOv4 (Mish)+LSTM model, where *dog everything*, *stop*, *boom down*, and *emergency stop* are the hand signals with respect to which the highest classification accuracy (97.8%, 96.4%, 94.5%, and 92.6%, respectively) is achieved, while *move slowly* and *hoist* are the signals with respect to which the lowest classification accuracy is achieved (83.8% and 85.3%, respectively). Figure 3-11, shows the confusion matrix of the YOLOv4 (ReLU)+LSTM model. From Figure 3-11, it can be seen that *emergency stop*, *boom down*, and *stop* are the hand signals with respect to which the highest classification accuracy (98.7%, 96.4%, and 96.1%, respectively) is achieved, while the lowest accuracy is achieved with respect to the *travel both tracks* and *boom down and raise the load* signals (87.7% and 88.1%, respectively). The confusion matrix of YOLOv4(Leaky-ReLU)+LSTM is represented in Figure 3-12. From Figure 3-12, it can be seen that the highest accuracy is achieved with respect to the *emergency stop* and *dog everything* hand signals (an accuracy of 100% for both), while, on the low end, for the hand signals *one track* and *travel both tracks*, accuracies of just 88.8% and 89.7%, respectively, are achieved. Table 3-4 represents the overall classification accuracy achieved by each of the proposed models. The highest overall classification accuracy was achieved by YOLOv4(Leaky-ReLU)+LSTM with a classification accuracy of 94.8%. YOLOv4(Mish)+LSTM

model is found to be 89.75% accurate, which is 5.05% less than YOLOv4(Leaky-ReLU)+LSTM model. The classification accuracy achieved the YOLOv4(ReLU)+LSTM model is 92.6%, which is 2.85 more than YOLOv4(Mish)+LSTM model and 2.2% less than YOLOv4(Leaky-ReLU)+LSTM model. In terms of processing speed/inference time, the YOLOv4(Mish)+LSTM model is found to have an inference time of 16 milliseconds (ms), while the YOLOv4(ReLU)+LSTM model and YOLOv4(Leaky-ReLU)+LSTM model have inference times of 17 ms and 20 ms, respectively. Inference time is then converted from ms to FPS for better understanding of inference time, as given in Table 3-4, and the deep-learning models are compared based on their overall classification accuracy and inference time. From Figure 3-12, it is noted that the YOLOv4(Leaky-ReLU)+LSTM model has the highest classification accuracy at 94.8%—2.2% and 3.1% higher than the YOLOv4(ReLU)+LSTM and YOLOv4(Mish)+LSTM models, respectively. On the other hand, the YOLOv4(Mish)+LSTM model has the fastest inference time at 62.5 FPS, while the YOLOv4(ReLU)+LSTM model has an inference time of 58.8 FPS and the YOLOv4(Leaky-ReLU)+LSTM model has an inference time of just 50.1 FPS. The results achieved by the deep learning-based intelligent models are sufficient to justify their use for real-time classification, given that, according to Potter et al. (2014), a human eye can process an object with an average inference time of 50–60 FPS, while an ordinary smart phone camera can receive and produce results with an average inference time of 30–60 FPS (Bae et al. 2017), and the results produced by the deep learning-based intelligent model are similar in terms of processing speed to the human eye perceiving unprocessed data from an ordinary camera.

Table 3-4.

Overall classification accuracy and inference time of the models.

Model	Overall classification accuracy (%)	Inference time (ms / FPS)
YOLOv4(MISH)+LSTM	89.75	16 / 62.5
YOLOv4(ReLU)+LSTM	92.60	17 / 58.8
YOLOv4(Leaky-ReLU)+LSTM	94.80	20 / 50.1
Faster R-CNN	86.8	62 / 16.1

The results achieved by the proposed models are then compared with the state-of-the-art Faster Region-based Convolutional neural network (Faster R-CNN) model with respect to the dynamic crane signalman hand signal dataset. The state-of-the-art Faster R-CNN model is found to achieve an overall classification accuracy of 86.8%, which is 2.95%, 5.8%, and 8% less than the YOLOv4(Mish)+LSTM, YOLOv4(ReLU)+LSTM, and YOLOv4(Leaky-ReLU)+LSTM models, respectively. The inference time recorded by the state-of-the-art Faster R-CNN model, meanwhile, is 16.1 FPS, which is 46.4 FPS, 42.7 FPS, and 34 FPS less than those of the proposed YOLOv4(Mish)+LSTM, YOLOv4(ReLU)+LSTM, and YOLOv4(Leaky-ReLU)+LSTM models, respectively. As these findings demonstrate, the proposed models are found to outperform the state-of-the-art Faster R-CNN model in classification accuracy and speed with respect to the custom dynamic crane signalman hand signal dataset.

3.6.3 Real-time implementation

For real-time classification of crane signalman hand signals, the YOLOv4(Leaky ReLU)+LSTM model is selected due to this model having achieved the highest classification accuracy during training, validation, and testing. For the real-time classification, six volunteer participants are

instructed to perform hand signals in both indoor and outdoor scenarios. Camera is placed accordingly as shown in Figure 3-3, and all the construction site scenarios are captured. The complexity of the construction site background (static and dynamic construction site) and weather conditions (cloudy and sunny) as given in Table 3-1, are also taken into consideration. Table 3-5 shows the results of the real-time crane signalman hand signal classification model. In the implementation, the highest accuracy is found to be have been achieved with respect to the *dog everything*, *stop*, and *emergency stop* hand signals, with classification accuracies of 98.5% 98.3%, and 97.2%, respectively. The overall classification accuracy achieved by the YOLOv4(Leaky-ReLU)+LSTM models in real time is found to be 93.5%, while the inference time is 44 FPS. It is noted that, during real-time classification, the accuracy of the model is found to be lower at just 93.5% (compared to 94.8% achieved during the training of the model). The reduction in accuracy when implementing the models in real time may be due to several factors, such as the subject being unable to perform the hand movements precisely (given that the participants performing the hand signalling tasks are not trained signalmen). Meanwhile, the inference time is also found to be lower at just 44 FPS to 50.1 FPS. This may be due to the hardware configuration, as the hardware used in the real-time classification has a different configuration than that used in training, validating, and testing the model. The model is trained, validated, and tested on a PC running the Windows 10 operating system, Intel core i9 processor, and 2-GeForce RTX 3090 GPUs. For the real-time classification, the model is deployed on a laptop running a Windows 10 operating system, Intel corei7 processor, and GeForce RTX 3060 GPU. Some examples of real-time classification using the improved YOLOv4+LSTM model are shown in Figure 3-13.

Table 3-5.

Classification performance of YOLOv4(Leaky-ReLU)+LSTM models in real time.

Standard crane signalman hand signal classification labels	Number of times hand signal is shown in the camera	Number of times hand signal is correctly classified	Number of times hand signal is incorrectly classified	Number of times hand signal is missed	Accuracy (%)
Hoist	65	59	3	3	90.8
Lower	62	57	3	2	91.9
Use main hoist	66	60	4	2	90.9
Use whipline	69	65	2	2	94.2
Boom up	71	65	4	2	91.5
Boom down	60	58	0	2	96.7
Move slowly	58	54	1	3	93.1
Swing	69	60	8	1	86.9
Boom down and raise the load	62	59	3	0	95.2
Boom up and lower the load	67	64	3	0	95.5
Stop	59	58	1	0	98.3
Emergency stop	71	69	0	2	97.2
Travel	60	56	2	2	93.3
Dog everything	65	64	1	0	98.5
Travel both tracks	70	65	5	0	92.9
Travel one track	68	63	3	2	92.6
Telescope out	64	60	4	0	93.8

Telescope in	62	56	5	1	90.3
--------------	----	----	---	---	------

3.7 Discussion

The proposed deep-learning framework showed promising results in the classification of dynamic crane signalman hand signals at high speed in real time. In terms of notable contributions of this research, first, the state-of-the-art classifier, (YOLOv4 model), was modified by altering the activation function in the backbone of the network to improve the classification accuracy of the classifier with respect to the custom crane signalman hand signal dataset. Second, the modified classifier was integrated with the LSTM recurrent neural network, making the framework capable of classifying dynamic hand signals with high classification accuracy and improved processing speed. The proposed model was further validated for real-time classification, achieving high accuracy in crane signalman hand signal classification, which has not been addressed in previous studies. The aim of this research was to improve on-site communication between crane operators and crane signalman. It should be noted in this regard that the proposed framework is not a replacement for the current practice, but rather a supplement to the current practice that automates the hand signal classification process based on deep-learning techniques. The proposed framework presents in detail a dynamic object classifier, and, in this respect, the framework can serve as a guide for the development of similar systems.



Figure 3-13. Examples of real-time classification using improved YOLOv4+LSTM model.

The justification for choosing YOLOv4 as the classifier in the proposed framework was its high classification accuracy and processing speed compared to similar deep-learning algorithms based on the experiments conducted by the developer of the YOLOv4 model (Bochkovskiy et al. 2020). The model proposed herein was trained on a Windows 10 platform with a high-speed GPU. In this research, videos were processed instead of images due to the dynamic nature of hand signals.

3.8 Conclusions and future work

The framework proposed in this chapter employed YOLOv4(Mish)+LSTM, YOLOv4(ReLU)+LSTM, and YOLOv4(Leaky-ReLU)+LSTM models to classify crane signalman hand signals as the basis for improving communication between crane operator and signalman. The models were trained using video clips of 18 hand signals used by signalmen to communicate instructions to the crane operator. Data augmentation techniques were applied to better generalize the dataset and mitigate the risk of model overfitting. A total of 4,054 video clips of the 18 crane signalman hand signals were used to train the model to classify the dynamic hand signals. The YOLOv4(Mish)+LSTM achieved an overall classification accuracy of 89.75% and YOLOv4(ReLU)+LSTM achieved an overall classification accuracy of 92.6%, while the YOLOv4(Leaky-ReLU)+LSTM performed best in overall classification accuracy, achieving 94.8% on the test dataset. In terms of inference time, the YOLOv4(Mish)+LSTM, YOLOv4(ReLU)+LSTM, and YOLOv4(Leaky-ReLU)+LSTM models achieved values of 62.5 FPS, 58.8 FPS, and 50.1 FPS, respectively. The performance of the proposed models was then compared with that of the state-of-the-art Faster R-CNN model, with the proposed models being found to outperform the Faster R-CNN model in classification accuracy and inference time with respect to the custom dynamic crane signalman hand signal dataset. The YOLOv4(Leaky-ReLU)+LSTM models are identified as the models best suited for real-time classification due to their high classification accuracy. YOLOv4(Leaky-ReLU)+LSTM achieved an overall classification accuracy of 93.5% with an inference time of 44 FPS. These results illustrate that the deep-learning models are capable of classifying dynamic hand signals in complex environments and under different weather conditions on construction sites.

In future research, more crane signalman hand signals and other construction site-related hand signals and construction signs will be added to the dataset to train the model to classify more classes of construction site-related hand signals and signs. The model will be further optimized to reduce the computation burden so that it can be implemented on micro-controllers and other mobile portable devices for on-site field classifications. Moreover, effective means of transmitting hand signal labels across long distances between crane operator and signalman in the construction site will be investigated. The models described in this chapter can improve communication on construction sites during crane operations and help to move the industry towards automation of crane operations.

Chapter 4: CRANE SIGNALMAN HAND SIGNAL CLASSIFICATION FRAMEWORK USING SENSOR-BASED SMART CONSTRUCTION GLOVE (SCG) AND MACHINE-LEARNING ALGORITHMS³

In the previous chapter, a vision-based framework to classify 18 dynamic crane signalman hand signals using YOLOv4+LSTM model was proposed. The research presented in this chapter employs a different approach, developing a sensor-based smart construction glove (SCG) and proposed machine-learning models to classify 18 dynamic crane signalman hand signals with the aid of sensor data. The study begins with the development of the sensor-based custom SCG, which is used to collect sensor data for each crane signalman hand signal. The collected data is then used to train four machine-learning models: k -nearest neighbour (KNN), support vector machine (SVM), decision tree (DT), and convolutional neural network–long short-term memory (CNN-LSTM). The models are validated and tested on unseen datasets. Furthermore, a mobile-based application is developed, and the top-performing models are deployed in real time to classify crane signalman hand signals. In real-time crane signalman hand signal classification, CNN-LSTM is found to outperform the other machine-learning models, achieving an accuracy of 93.87%. The proposed framework offers an accurate and efficient method of communicating with crane operators, thereby reducing the risk of accidents and improving overall construction site safety.

4.1 Introduction

In the construction industry, communication between workers is of paramount importance. The use of heavy machinery such as cranes can pose significant risks if the communication commands

³ A version of this chapter has been submitted to the *Journal of Construction Engineering and Management*.

are not conveyed correctly. Crane signalmen play a vital role in the safety of construction sites by communicating with crane operators using hand signals. However, traditional hand signals can be prone to misinterpretation, especially in noisy or visually obstructed environments (as discussed in Chapter 1).

Sensor-based technologies have been used in various industries, such as manufacturing, transportation and the healthcare sector, to increase efficiency, productivity, and safety, and they have also been used in a variety of sectors to reduce the risk of accidents and injuries. These technologies allow devices to communicate with one another and with humans by detecting and responding to changes in their surroundings and by relaying data to other devices. They also assist in reducing human error and miscommunications (Oztemel and Gursev 2020; Ghobakhloo 2018; Guerrero-Ibáñez et al. 2018; Tewolde 2012; Qureshi and Abdullah 2013; Qamar et al. 2018; Minaie et al. 2013). In this context, this chapter proposes a sensor-based SCG equipped with various sensors to capture the signalman's hand signals and classify them in real time with the help of machine-learning models.

A number of different algorithms for sensor-based hand signal classification and have been proposed and implemented in such applications as sign language classification for speech-impaired people (Gurbuz et al. 2020; Jani et al. 2018; Al Mamun et al. 2017; Sriram and Nithiyanandham 2013), gaming (Tu et al. 2020), hand motion classification (Chuang et al. 2019; Dong et al. 2020; Kim et al. 2019) to promote human–computer interaction. These applications, which rely on such technologies as accelerometer and gyroscope sensors, IMU sensors, Electromyography (EMG) sensors, microelectromechanical system (MEMS) sensors, and flex sensors, have achieved excellent results in recognizing hand signals. Nevertheless, and in spite of its potential benefits in terms of improving on-site communication and safety, the use of sensor-based technologies in the

construction industry has been very limited (Awolusi et al. 2018). The present study thus has two main objectives. The first is the development of an SCG equipped with a 3-axis accelerometer, 3-axis gyroscope, 3-axis magnetometer, and flex sensors. These sensors record the orientation of the hand and the magnitude of the bend in the fingers. The second objective is to train various different machine-learning models on the data collected from the SCG and assess their performance for real-time crane signalman hand signal classification. An android-based mobile application is also developed that receives the decision from the SCG in text format and then audibly outputs it.

This chapter is organized as follows: The next section discusses the related literature on wearable sensors in the construction industry and the use of sensors for hand signal classification. This is followed by an overview of the research framework, including the development of the SCG, data collection, and preprocessing. Next is a description of the four machine-learning models trained and evaluated using the collected dataset. Finally, the real-time performance of the proposed machine-learning models is evaluated.

4.2 Related literature

4.2.1 Sensors for hand signal classification

Hand signalling can be a powerful means of communication in a number of different contexts. Hand signal classification is the process of recognizing the meaningful signals communicated using the hands. Many efforts have been made by researchers to classify hand signals using vision- and sensor-based hand signal classification techniques. Vision-based hand signal classification devices use deep learning and classification labelling to interpret the images and videos in which the hand signals are captured. However, the effectiveness of deep learning in such applications is subject to weather conditions and dynamic changes on the construction site, and high

computational power is required in order to train the deep-learning models (Mansoor et al. 2023; Bonato 2009).

Sensors are a technology that can improve the accuracy of real-time hand signal classification (Awolusi et al. 2018). Sensors have been used for hand signal classification in a range of different applications. For instance, Rosero-Montalvo et al. (2018) presented an intelligent glove for classifying sign language as a way of automating communication for speech-impaired people. They proposed using a glove with five flex-sensors (one flex-sensor affixed to each finger) and a KNN classifier. Their intelligent glove achieved high accuracy in classifying hand gestures. Taneja and Sukhija (2020) also developed a sensor-based glove featuring flex-sensors, this one being used to facilitate simple communication commands between the speech-impaired and other people. Their glove was found to be capable of lowering the communication barrier between the speech-impaired and other people. Pathak et al. (2016) developed a framework by fusing the data from flex-sensors, contact sensors, and accelerometer sensors to recognize 26 alphabets of American Sign Language. Gupta et al. (2016), meanwhile, developed a technique for continuous hand gesture recognition using an accelerometer and gyroscope sensor, while Zhang et al. (2011) developed a framework that used an accelerometer and EMG sensor to recognize Chinese sign language. Other applications of the sensor-based glove technology with specific hand signals have included robotic controls (Sathiyarayanan et al. 2014), robotic arms (Raheja et al. 2010), health care (Kim et al. 2019; O'Flynn et al. 2015) and augmented reality (Minh et al. 2019).

4.2.2 Wearable Sensors in the construction industry

Wearable technology is beginning to be developed, tested, and used for a range of applications in the construction industry. For example, Akhavian and Behzadan (2016) used accelerometer, gyroscope, and magnetometer sensors affixed to the arm of the construction worker to calculate

the durations of construction tasks. Their framework was validated by comparing the sensor data with real-world experiments, and the sensor data was found to be very close to reality. Jebelli et al. (2019) proposed a wearable biosensor technology to assess the stress level of construction workers. Sensors were used to collect physiological signals from the worker's body and, based on the collected signal data, the body stress level was measured. Their framework achieved an accuracy of 84.48% in predicting low and high stress levels. Nath et al. (2017) used smartphone sensors for ergonomic analysis of construction workers' body postures. They found the data from the smartphone sensors to be very close to the measurements made by visual observation. Other applications of wearable technology in construction have included classification of worker activities on construction sites (Srikanth et al. 2021; Kanan et al. 2018; Jo et al. 2019); measurement of construction workers' physiological indicators, such as heart rate (Hwang et al. 2016) and breathing (Roelofs et al. 2011); environmental sensing to detect hazardous gases and chemical leaks (Lai et al. 2019); proximity detection and alerting system for construction site safety (Marks et al. 2013; Park et al. 2016); and detection of workers losing their balance (Ahn et al. 2019), to name a few.

Although sensor-based technologies assist the construction industry in keeping Job sites safe and more productive, but the sensors are not being used to improve communication on the construction site. This chapter presents a framework to enhance communication on the construction site, especially between crane operators and signalmen using multiple sensor and machine-learning algorithms. The proposed framework can assist in classifying in real time 18 different hand signals used in communication between crane operator and signalman during crane operation.

4.3 Research framework

The overview of the research framework is shown in Figure 4-1. As can be seen, the research consists of three steps: [1] develop a custom-built SCG for the purpose of data collection; [2] acquire the dataset using the developed SCG, preprocess the data, and remove any sensor noise; and [3] train four machine-learning models—*k*-nearest neighbour (KNN), support vector machine (SVM), decision tree (DT), and convolutional neural network / long short-term memory (CNN-LSTM)— to classify the crane signalman hand signals with high accuracy in real time in order to enhance communication between crane operator and signalman.

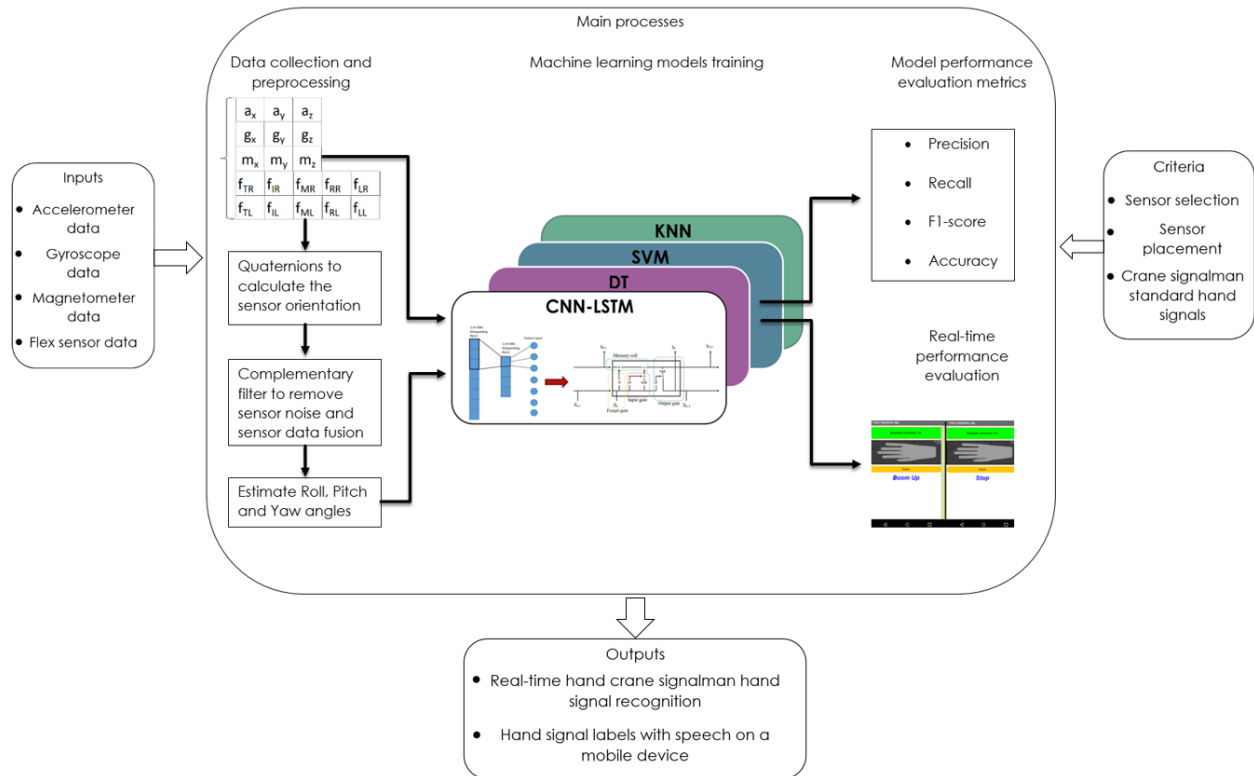


Figure 4-1. Overview of the research framework.

4.3.1 Custom build SCG

A custom-built SCG to capture the crane signalman hand signals on the construction site is developed. The crane signalman hand signals are shown in Figure 3-2. The selection of the sensors to measure the orientation of the hand and bend in the finger based on the given hand signals is of utmost importance. To measure the orientation of the hand, a 3-axis accelerometer, 3-axis gyroscope, and 3-axis magnetometer are affixed to the SCG, while flex-sensors affixed to the SCG are used to measure the bend of the finger. An accelerometer sensor, it should be noted, measures the directional movement or linear acceleration of an object. A gyroscope sensor, meanwhile, is used to measure the lateral orientation or angular velocity of an object with respect to the Earth's gravity and how fast the object is spinning about an axis. A magnetometer sensor is used for measuring the Earth's magnetic field intensity and other magnetic anomalies, and flex-sensors, finally, use a resistive carbon element to produce a resistance output correlated to the amount of bending in the fingers (Saggio 2012). These sensors are selected because they are lightweight, low-cost, flexible, easy to install, have low power consumption, and are readily available. An Arduino microcontroller is also affixed to the SCG to assist in collecting, reading, processing, and interpreting the sensor's data and providing meaningful output.

The selected sensors are then placed on the glove, satisfying the following constraints.

- The sensors must be placed such that they can detect the movements and orientation of the hand.
- The sensors must be placed such that they do not restrict the worker's hand from performing other tasks.
- The sensors must be placed such that they are not disturbed by external pressure on the hand while other tasks are being performed.

The regions of the hand identified for sensor placement based on these constraints, as well as the layout of the SCG, are shown in Figure 4-2.

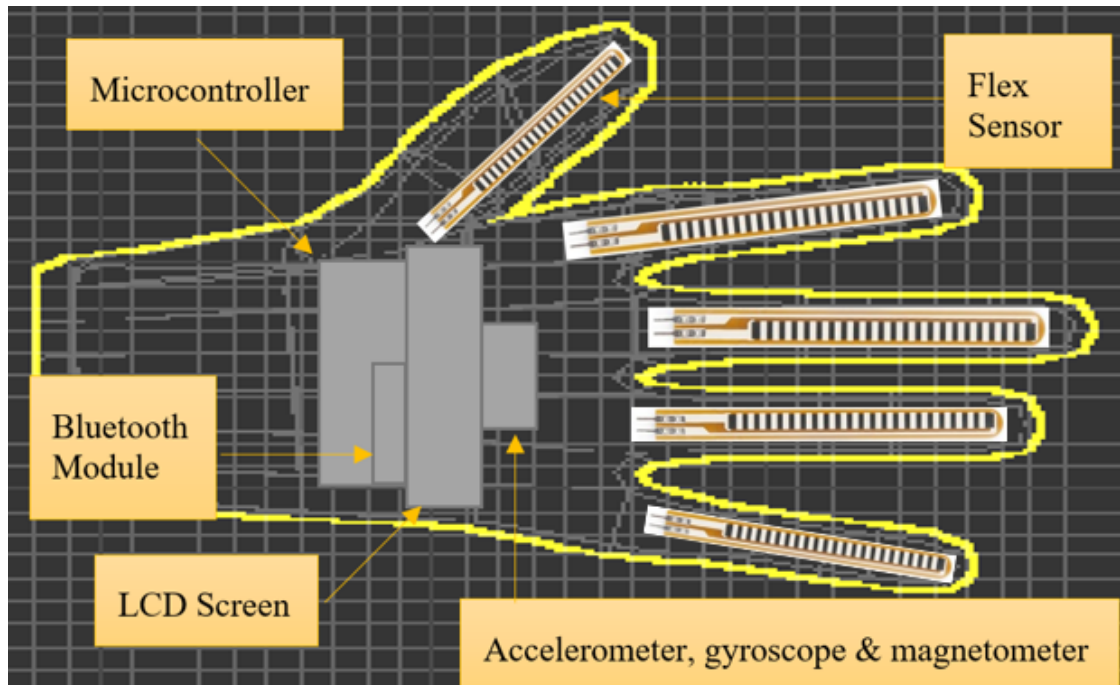


Figure 4-2. Layout of smart construction glove.

4.3.1.1 Estimation of crane signalman hand orientation

There are various methods for measuring the orientation of a rigid object in 3D space, with two of the most widely used being Euler angles and quaternions. Euler angles are among the most widely used methods for measuring the orientation of an object in a set of three consecutive rotations in a coordinate system. Although Euler angles are favoured for their ease of implementation, they do suffer from singularity and gimbal lock issues (Hemingway and Oliver 2018). A gimbal lock, it should be noted, is the condition of losing one degree of freedom in 3D space, and it occurs when the angle of pitch is near $\pm 90^\circ$. When the pitch angle reaches nearly 90° , the roll and yaw make the sensor move in the same direction. A gimbal lock can occur in either of two possible forms in a sequence of rotation: (1) pitching and then yawing, and (2) rolling and then pitching. This

phenomenon can lead to inaccuracy or failure when estimating the orientation of the sensor using Euler angles (Hemingway and Oliver 2018).

To overcome this problem, this research uses quaternions to measure the orientation based on the roll, pitch, and yaw angles of the hands, as this method overcomes the problems of singularity and gimbal lock. A quaternion is a four-element vector—one real and three complex elements—that can measure the rotation and direction of any object in a 3-axis coordinate system (Zhang. 1997). The orientation is represented as a unit vector quaternion as expressed in Eq. (4.1).

$$\mathbf{q} = (q_0, q_1x, q_2y, q_3z)^T \quad (4.1)$$

where q_1x , q_2y , and q_3z represent the vector about which the rotation is performed, while q_0 represents the scalar quantity defining the amount of rotation about the vector when α is the angle made by the rotation and (V_x , V_y , and V_z) are the unit vectors in the axis of rotation. The matrix is given in Eq. (4.2).

$$\begin{pmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{pmatrix} = \begin{pmatrix} \text{Cos} \left(\frac{\alpha}{2} \right) \\ V_x \text{Sin} \left(\frac{\alpha}{2} \right) \\ V_y \text{Sin} \left(\frac{\alpha}{2} \right) \\ V_z \text{Sin} \left(\frac{\alpha}{2} \right) \end{pmatrix} \quad (4.2)$$

Detailed calculations of quaternions can be performed following the method described by Valenti et al. (2015). Furthermore, to achieve the roll pitch and yaw angles (φ , θ , and ψ) in the z - y - x sequence, the quaternions are converted into Euler angles using Eq. (4.3).

$$\begin{bmatrix} \varphi \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} \text{atan2}(2(q_0q_1 + q_2q_3), 1 - 2(q_1^2 - q_2^2)) \\ \text{asin2}(2(q_0q_2 - q_3q_1)) \\ \text{atan2}(2(q_0q_3 + q_1q_2), 1 - 2(q_2^2 + q_3^2)) \end{bmatrix} \quad (4.3)$$

Obtaining reliable and accurate output data from sensors is of critical importance. The raw data from the accelerometer sensor is noisy and unreliable when exposed to external vibration and sudden movements, and the 3-axis rotation of moving objects cannot be easily sensed (Chaudhury et al. 2014). To overcome the accelerometer problem, a gyroscope sensor is used. However, the gyroscope sensor is subject to drifting, whereby the angle measurement gradually changes and error increases as the sensing time increases (Nonnarit and Barreto 2016). To correct the drifting phenomenon, a magnetometer is used. The concept underlying the use of multiple types of sensors is to address the weaknesses of a given sensor by leveraging the strengths of other types of sensors. To overcome the issue of sensor noise, finally, complementary filters with quaternions are used, as described in the following section.

4.3.1.2 Complementary Filter

The use of complementary filters is an effective method of reducing sensor noise. Complementary filters require comparably little computational power, are easy to implement, and feature user-friendly algorithms, making them a preferred choice for use in systems with low computational capabilities (McGinnis et al. 2014). In complementary filters, the accelerometer and gyroscope are the primary sensors, and the magnetometer is used as a corrective sensor. In the present study, the force acting on the SCG is measured by an accelerometer that provides reliable results on long-term movements; on the other hand, though, the accelerometer is sensitive to slight short-term vibrations, and as such the data given by the accelerometer is not reliable in the short term. To address this, a low-pass filter is used to correct small, short-term vibrations. Meanwhile, a high-pass filter is used to correct the gyroscope drifting problem (mentioned above) in the long term. Both the low- and high-pass filters are complementary filters, and integrating the data of the accelerometer and gyroscope using these complementary filters results in precise outputs.

In the present study, complementary filters are used as a sensor fusion technique that integrates the accelerometer and gyroscope data and reduces sudden vibration and drifting noise in the sensor data. The quaternions are used to mitigate the gimbal lock problem, and are further converted into Euler angles to compute the roll, pitch, and yaw angles as the basis for estimating the orientation of the crane signalman's hands. The equations to compute roll, pitch, and yaw using a complementary filter are given in Eq. (4.4), Eq. (4.5), and Eq. (4.6).

$$roll = k_f \times (roll_{prev} + (x_{gyr} \times \Delta t)) + (1 - k_f) \times roll_{acc} \quad (4.4)$$

$$pitch = k_f \times (pitch_{prev} + (y_{gyr} \times \Delta t)) + (1 - k_f) \times pitch_{acc} \quad (4.5)$$

$$yaw = k_f \times (yaw_{prev} + (z_{gyr} \times \Delta t)) + (1 - k_f) \times yaw_{mag} \quad (4.6)$$

Here k_f represents the filtering coefficient given in Eq. (4.7).

$$k_f = \frac{t_{const}}{t_{const} + \Delta t} \quad (4.7)$$

where the value of k_f lies between 0 and 1. The complementary filter is mainly dependent on the k_f value because of the time delay due to the gyroscope's high-pass filter and the accelerometer's low-pass filter. The k_f value is used to overcome the time delay, and the filter performs better when the value of k_f is close to 1 and more poorly when the value of k_f is close to 0. The yaw angle is obtained using a magnetometer sensor, as an accelerometer sensor cannot be used to obtain the yaw angle because its vertical axis is parallel to Earth's axis, while the yaw obtained from the gyroscope sensor is used in the complementary filter to correct the drifting problem. The detailed algorithm for the complementary filter is expressed as Algorithm 1 below. The process starts with

the input data from the accelerometer, gyroscope, and magnetometer sensors, implementation of quaternions, and fusing of the sensor data to achieve roll, pitch, and yaw angles as outputs.

Algorithm 1

1. Acc: a_x, a_y, a_z
 2. Gyro: g_x, g_y, g_z
 3. Mag: m_x, m_y, m_z
 4. Quaternion $q = (q_0, q_1x, q_2y, q_3z)^T$
 5. Euler ϕ, θ, ψ
 6. Filtering coefficient:
 7. $k_f = \frac{t_{const}}{t_{const} + \Delta t}$
 8. $k_f = 0.94, t_{const} = 0.159, \Delta t = 0.01$
 9. Gyro - high pass filter
 10. Acc - low pass filter
 11. Data fusion:
 12. $0.94 \times (roll_{prev} + (x_{gyr} \times \Delta t)) + 0.06 \times roll_{acc}$
 13. $0.94 \times (pitch_{prev} + (y_{gyr} \times \Delta t)) + 0.06 \times pitch_{acc}$
 14. $0.94 \times (yaw_{prev} + (z_{gyr} \times \Delta t)) + 0.06 \times yaw_{mag}$
 15. **Output:**
 16. Roll, Pitch, Yaw
-

4.3.2 Data collection and preprocessing

Due to the lack of an available sensor dataset for crane signalman hand signals, the developed SCG is used to collect the crane signalman hand signal dataset. The SCG is equipped with an IMU sensor that consists of a 3-axis accelerometer, 3-axis gyroscope, 3-axis magnetometer, and flex-sensors. Prior to data collection, the sensors are calibrated. To calibrate the accelerometer and gyroscope, the sensor is kept in a stationary position relative to the x -, y -, and z -axes for a few seconds. To calibrate the magnetometer, some random movements of the hand are made. The offsets observed from the accelerometer, gyroscope, and magnetometer during the calibration

process are removed automatically, as this allows the sensor to be fully calibrated. An example of sensor calibration is shown in Figure 4-3.

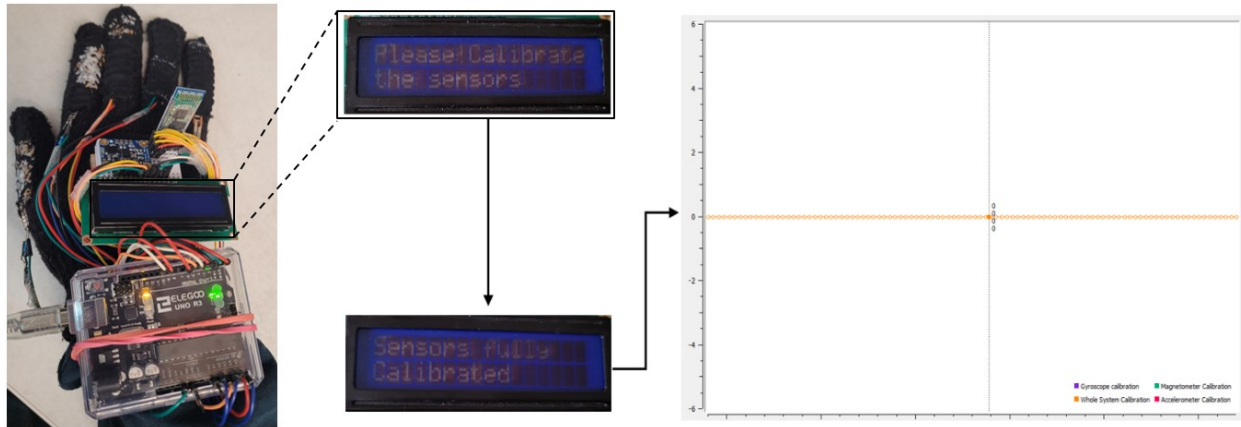


Figure 4-3. Example of sensor calibration.

The SCG yields 14 types of raw data: accelerometer data in the x -, y -, and z -axes (a_x , a_y , and a_z), gyroscope data in the x -, y -, and z -axes (g_x , g_y , and g_z), magnetometer data in the x -, y -, and z -axes (m_x , m_y , and m_z), and flex-sensor data from the thumb, index, middle, ring, and little finger (f_T , f_I , f_M , f_R , and f_L , respectively). Furthermore, the IMU sensor data is fused and filtered to achieve roll, pitch, and yaw angles as the basis for measuring the orientation, while the flex-sensor is used to measure the bend in the finger. A combination of IMU sensor data and flex-sensor data helps to determine the crane signalman hand signals. A total of 16 volunteers are engaged to participate in the process of data collection; two SCGs—one on each hand—are used. Each of the 18 different hand signals used by crane signalmen is clearly explained to the participants, and they are asked to perform them for approximately 9 min—30 s for each hand signal—with a sampling rate of 100 Hz. The IMU and flex-sensor data collected from the participants is transmitted to a computer (via a Bluetooth module affixed to the SCG) in the form of an Excel file, as shown in Figure 4-4. Approximately 864,000 sample datapoints are collected—54,000 datapoints from each participant

4.3.3.1 k-nearest neighbours (KNN)

KNN is a machine-learning algorithm that is easy to implement and that has been shown to be capable of achieving high accuracy in sensor data classification. Because it does not go through a learning process, it is considered a direct classification method that only requires the whole dataset to be stored in the model (Kaghyan and Sarukhanyan 2012; Mandong and Munir 2018). The KNN model predicts the class of new datapoints based on the labelled datapoints already stored in the model. To be more specific, it calculates the Euclidean distance between the new datapoint and the labelled training datapoints and makes the prediction accordingly, assigning the new datapoint to a class based on the majority votes of its k -nearest neighbours. In the present study, the hyperparameter selected in the KNN model includes a k -value of 5, and Euclidean distance is used as a distance function to measure the distance between new and training datapoints in order to make the final prediction. To obtain the optimum value of k , multiple values of k ranging between 1 to 20 are analyzed, with a k -value of 5 being found to achieve the highest accuracy.

4.3.3.2 Support vector machine (SVM)

SVM is another machine-learning classification algorithm that has strong generalization capabilities. SVM requires less training data and is computationally less expensive compared to other machine-learning models. For these reasons, it has been used in a number of sensor-based data classification applications, such as gait classification from sensor data (Begg et al. 2005) and human fall detection from sensor data (Shibuya et al. 2015), to name a few. The SVM algorithm finds a hyperplane, which separates the datapoints from different classes. The algorithm then finds the datapoint nearest to the line from different classes, this datapoint being known as the support vector. The algorithm then computes the distance between the support vector and the line. The goal of this procedure is to maximize the distance between the support vector and the line in order

to achieve the optimal hyperplane (i.e., the one that provides the greatest distance between different classes). In the present study, a penalty coefficient C of 3 is used, and the gamma (also known as the coefficient of kernel function) is set at -7 .

4.3.3.3 Decision Tree (DT)

DT is a popular machine-learning algorithm that has been used in a wide range of sensor-based data analysis applications, such as human activity recognition (Fan et al. 2013) and diagnosis of coronary artery diseases (Ghiasi et al. 2020), to name a few. Some of its benefits are its low computational complexity, its ability to handle irrelevant feature data, and its insensitivity to missing data. A DT establishes the nodes based on available features in the overall dataset and further selects the split with the most homogenous feature sub-nodes, then repeats the process until the classification decision is displayed within the tree structure. The most important hyperparameter of the DT is to choose the maximum depth of the tree. For the present study, the depth of the tree is set to 17 based on multiple experiments. This hyperparameter is found to achieve the best decision output for the given dataset.

4.3.3.4 Convolutional Neural Network – Long Short-Term Memory (CNN-LSTM)

The CNN-LSTM model is a combination of two state-of-the-art deep-learning models. The function of the CNN component of the CNN-LSTM model is to extract features from the sensor data, while the LSTM component of the model is responsible for learning from the recurrent features to produce accurate outputs. In the present study, the sensor data is sampled using a fixed-width sliding window of 2.56 s, an overlap of 50% is used, and a 1D CNN model using Keras with a TensorFlow backend is employed. The model is trained, validated, and tested on raw and preprocessed data collected using the SCG. The raw dataset consists of 28 attributes of sensor data that include data from both the right hand ($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, f_T, f_I, f_M, f_R, f_L$) and left

hand ($a_x, a_y, a_z, g_x, g_y, g_z, m_x, m_y, m_z, f_T, f_I, f_M, f_R, f_L$). The preprocessed dataset comprises 16 attributes (i.e., the roll, pitch, yaw, f_T, f_I, f_M, f_R , and f_L for both the right hand and the left hand). The architecture of the 1D CNN model consists of two 1D convolutional layers with 128 filters and a kernel size of 5, followed by two 1D max-pooling layers. Moreover, rectified linear unit (ReLU) is used as an activation function. ReLU, it should be noted—the most widely used activation function due to its high performance—is an identity line where $y = x$ for all positive lines and $y = 0$ for all negative values (Agarap 2018). A flatten layer, finally, is added to achieve the flatten feature maps from the sensor data. The output of the flatten layer is fed to the LSTM layer, which is responsible for learning from the recurrent temporal features of the sensor data. The output of the LSTM layer is then fed to the softmax activation function responsible for the prediction of the final output with the classification score. The CNN-LSTM model is trained, validated, and tested on 864,000 samples. The other hyperparameters used for training the model are a learning rate of 0.005, a mini-batch size of 128, and a dropout of 0.2, and the model is trained in 100 epochs.

4.3.4 Models performance evaluation

Multiple experiments are conducted to evaluate the performance of the machine-learning models discussed in the previous sections. The matrices used to measure the performance of the model in the test dataset include precision, recall, F1-score, and accuracy. The respective mathematical equations for the evaluation matrices are given in Eq. (4.8), (4.9), (4.10), and (4.11), respectively. The definition of TP, FN, FP and TN are given in Table 4-1.

$$Precision = \frac{TP}{TP + FP} \quad (4.8)$$

$$Recall = \frac{TP}{TP + FN} \quad (4.9)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.10)$$

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \quad (4.11)$$

Table 4-1.

Definition of TP, FN, FP, and TN.

		Predicted	
		Positive	Negative
Actual	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

The proportion of the data used to train the model is kept at 70% to allow the model to learn important features from the sensor data, while the model is validated and tested on the remaining 30% of the data (i.e., 15% each for validation and testing).

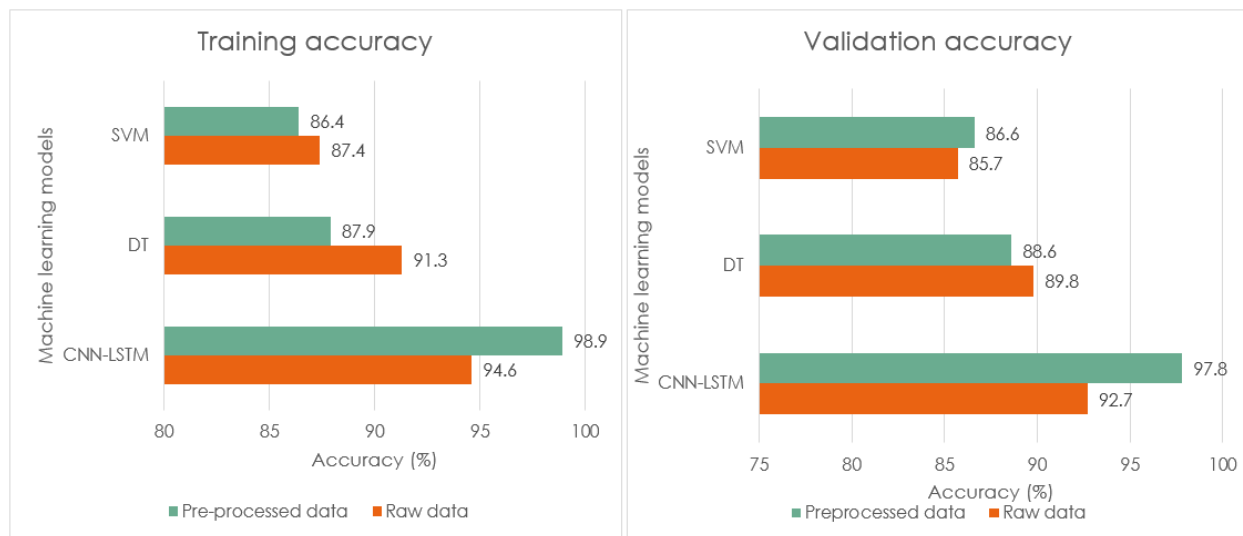


Figure 4-5. Training and validation accuracies of machine-learning models.

The machine-learning models are trained, validated, and tested using both raw and preprocessed datasets. After training, the CNN-LSTM model is found to outperform the other models, achieving training and validation accuracies of 98.9% and 97.8%, respectively, on the preprocessed data and 94.6% and 92.7%, respectively, on the raw dataset, as shown in Figure 4-5. The poorest performing model, SVM, is found to achieve training and validation accuracies of just 87.4% and 85.7%, respectively, on the raw dataset and 86.4% and 86.6%, respectively, on the preprocessed dataset. From the model’s training and validation results, it is also observed that the machine-learning models achieves higher accuracy with the preprocessed data than with the raw dataset. The reason for this is that, in the preprocessed dataset, the sensors are calibrated and sensor noise is removed by fusing the sensor data, whereas the raw data has not undergone any preprocessing.

After training and validation, the models are further validated on the remaining 15% of the test dataset. For testing, only preprocessed data is used, since the models perform better on the preprocessed dataset than on the raw dataset. As noted above, the machine-learning models are evaluated based on precision, recall, F1-score, and accuracy.

Table 4-2.

Results of the machine-learning model on the test dataset.

Evaluation matrices	KNN	SVM	DT	CNN-LSTM
Precision (%)	68.7	54.9	59.1	84.3
Recall (%)	68.4	55.6	61.8	83.9
F1-score (%)	68.1	55.3	60.4	84.0
Accuracy (%)	86.5	76.2	80.7	94.22

Table 4-2 shows the precision, recall, F1-score and accuracy achieved by the model evaluated on the test dataset. Based on the results in Table 4-2, it can be seen that CNN-LSTM is found to

inventor platform. The mobile application connects the SCG's Bluetooth module to the Android device (via Bluetooth). Once the connection is established, the mobile application receives the data from the SCG in text format as "labels of crane signalman hand signals". The Android-based mobile application also features a text-to-speech conversion function, which converts the resulting output text into speech that can be outputted by the Android device's speaker. The interface of the mobile application is shown in Figure 4-8.

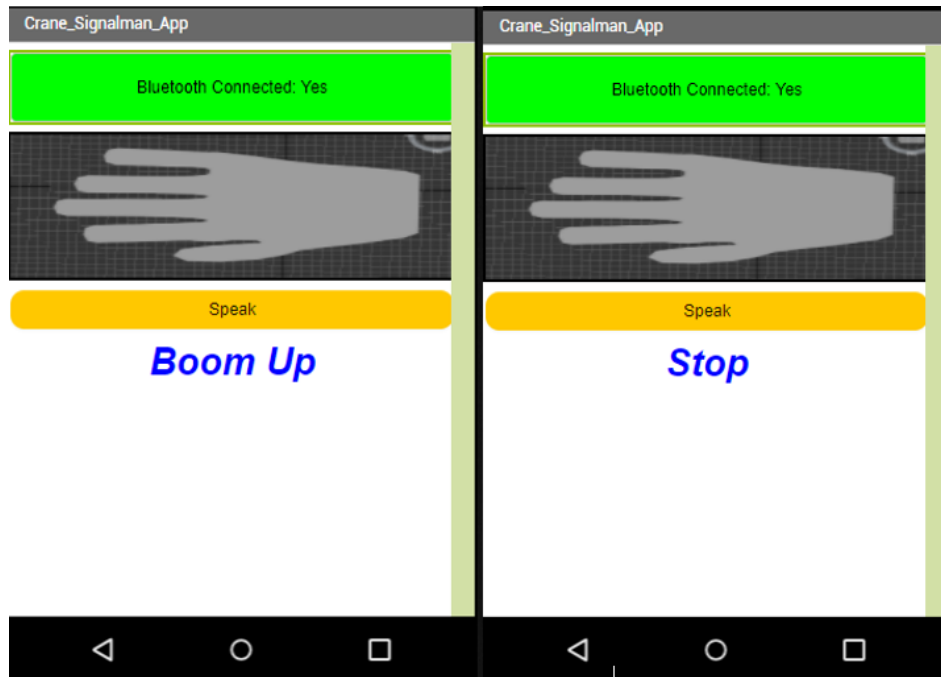


Figure 4-8. Layout of developed mobile application.

Trained CNN-LSTM and KNN models are deployed to evaluate their performance in real time using the developed Android application. Six participants are trained to perform the crane signalman hand signals wearing the developed SCG. Each of the participants performs the hand signals for about 2 min each, and the accuracy of each hand signal is calculated for each iteration. The equation used to calculate the accuracy of the models in real time is given in Eq. (3.4).

The CNN-LSTM model is found to achieve an overall accuracy of 93.87% in real time, while the KNN is found to achieve an accuracy of 85.37%, meaning that the CNN-LSTM model outperforms the KNN model in hand signal classification. Both the overall and individual hand signals classified by both models are given in Figure 4-9.

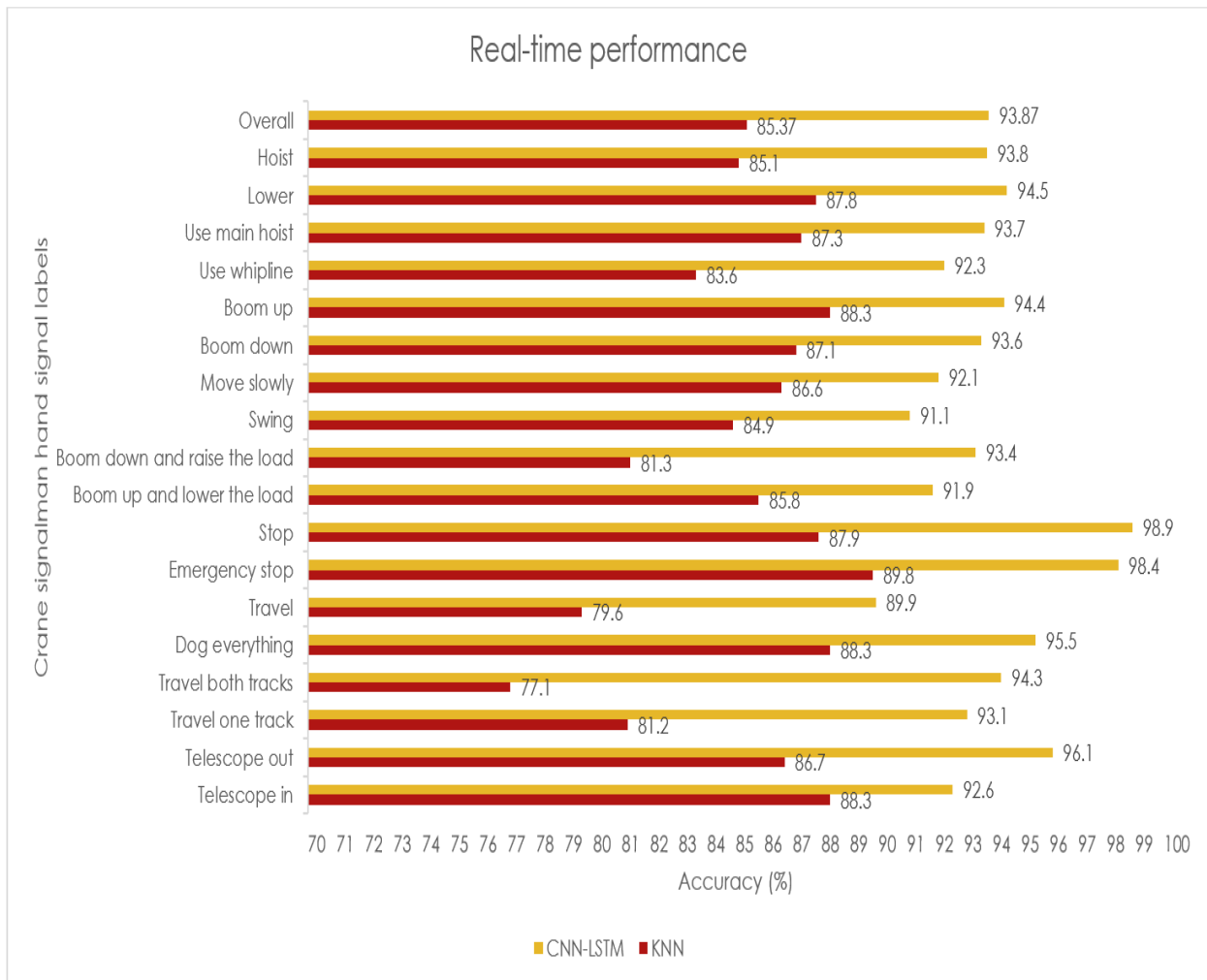


Figure 4-9. Real-time performance of KNN and CNN-LSTM model.

4.5 Conclusion

Hand signalling is well established as the primary means of communication between crane signalman and crane operator on construction sites due to its proven effectiveness. However, hand

signals are often not captured and interpreted correctly due to the long distances between the crane signalman and operator or due to obstacles obstructing the sightline, and this can lead to errors and accidents. This chapter describes the development of a sensor-based SCG capable of classifying in real time 18 different crane signalman hand signals used on construction sites. It begins with a literature review of sensor-based hand signal classification technologies, as well as a discussion of the wearable sensing technologies in use within the construction industry. This is followed by a description of the selection of sensors for detecting the crane signalman's hand movements. The selected sensors are affixed to the SCG, then calibrated, and any noise in the sensor readings is removed using a complementary filter fusion algorithm. Data is collected for 18 different crane signalman hand signals using the developed SCG. Four machine-learning models (KNN, SVM, DT, and CNN-LSTM) are trained, validated, and tested using the sensor data collected from the SCG. The CNN-LSTM model is found to outperform the other investigated models on the test dataset, achieving a precision of 84.3%, a recall of 83.9%, an F1-score of 84%, and an average accuracy of 94.22%, while the KNN model is found to be the second-best performing model, achieving a precision, recall, F1-score, and accuracy of 68.7%, 68.4%, 68%, and 86.5%, respectively. Moreover, an Android-based mobile application is developed that receives the data from the glove via Bluetooth in text format and outputs it from the Android device's speaker. The two best-performing machine-learning models (i.e., CNN-LSTM and KNN) are validated in real time with respect to their ability to classify 18 different crane signalman hand signals in real time, and the CNN-LSTM model is found to outperform the KNN model, achieving an overall accuracy of 93.87% (compared to 85.37% in the case of the KNN model). Given the high accuracy (i.e., 93.87%) of the CNN-LSTM model it employs, the SCG can be considered a promising solution as an additional layer of communication to complement hand signalling and two-way radio

communication, thereby improving communication and reducing the likelihood of communication errors on the construction site.

Chapter 5: KEYWORD IDENTIFICATION FRAMEWORK TO FACILITATE THE SPEECH COMMUNICATION ON CONSTRUCTION SITES⁴

Construction sites are inherently complex and dynamic environments, where efficient communication is vital for ensuring smooth operations and maintaining a safe working environment. However, the presence of high levels of noise, workers from multiple linguistic backgrounds, and visual obstructions can make it challenging to communicate effectively, especially when using traditional methods such as hand signalling and speech.

In the research described in previous chapters, computer vision and sensor-based techniques were used to classify crane signalman hand signals. However, crane signalmen also use two-way radio speech communication when there is an obstruction in the line of sight between crane operator and signalman or when hand signalling is otherwise not possible. There are several challenges in two-way radio speech communication (e.g., speaker accent, noise in the site, etc.), as discussed in detail in Chapter 1.

To address these challenges, this chapter proposes a keyword identification framework for speech communication on construction sites, drawing upon the case of communication between crane signalman and crane operator. This study uses a 1D CNN model to classify 18 speech commands/keywords used to guide the crane operator in crane operation. The model is trained, validated, and tested using crane signalman speech datasets. The model is further validated in real time to identify

⁴ A version of this chapter has been published, as follows: Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., Al-Hussein, M., and Hassan, I. (2022). “Keyword identification framework for speech communication on construction sites.” *Proceedings of the 2022 Modular and Offsite Construction Summit*, Edmonton, AB, Canada, Jul. 27–29, pp. 106–113.

the crane signalman speech commands/keywords from live speech, with the model achieving an accuracy of 95.3% in identifying crane signalman commands from live speech data.

5.1 Introduction

Effective communication is vital across all industries for establishing and maintaining a productive working environment. In current practice in construction, as discussed in Chapter 1, workers on construction sites typically rely on face-to-face verbal, hand signalling, and two-way radio speech communication systems (Mansoor et al. 2020). However, when there is a visual obstacle or a significant distance between workers, face-to-face verbal or hand signalling communication may not be reliable or even feasible. In such cases, two-way radio speech communication is the most common way of conveying the message (Stevenson, 2019). In two-way radio speech communication, radio units are used to send and receive audio data (Carbonell et al. 2020). On construction sites, this communication technique is primarily used to establish communication between ground workers and heavy machinery operators, such as in crane operations. In order for this communication method to work effectively, though, a dedicated radio channel must be allocated for the purpose of exchanging signals between the operator and signalman, and it must remain active at all times.

However, construction sites can be noisy due to construction activities such as drilling and the operation of heavy equipment (Kwon et al. 2016), making it difficult for the listener to hear speech commands. Furthermore, construction workers typically represent a diverse range of different ethnic and linguistic backgrounds and have different accents, meaning that it may be difficult for the listener to understand the speaker in some cases, leading to misjudgments in decision-making, as well as safety and productivity issues (Bust et al. 2008, Mansoor et al. 2022).

There is an opportunity in this regard for the construction industry to benefit from recent developments in information technology. Intelligent and automated communication systems can significantly enhance the communication between heavy machinery operators and ground workers on construction sites, ultimately leading to improved safety and productivity. In this context, the present study develops a framework for identifying keywords in speech on construction sites. The developed framework provides an intelligent, advanced, and reliable communication system that can reduce the risk of miscommunication on construction sites.

5.2 Related studies

Speech is the primary means of communication among human beings; as such, speech recognition systems have received considerable interest among researchers in recent decades. However, due to reliability issues, the systems developed have not been widely implemented (Latif et al. 2021; Otter et al. 2020; Strehl et al. 2006). Nevertheless, the major advancements in machine learning and deep learning in recent years have led to accurate speech recognition with high reliability which has increased the practicability of speech recognition systems (Hinton et al. 2012; Meftah et al. 2018). Speech recognition systems are now being used for various applications, including (1) keyword identification/spotting (Lopez-Espejo et al. 2021; Michaely et al. 2017; Werchniak et al. 2021; Momeni et al. 2020); (2) automatic recognition of the content of words and phrases in order to direct computer tasks as an alternative to typing, facilitating human-machine interaction as a support for the disabled, supporting smart home functions, etc.; (3) emotion recognition, i.e., for recognizing the emotion of the speaker based on speech signals (Fragopanagos and Taylor, 2005; Petrushin, 2000); (4) in intelligent health care systems to provide information on patient health status (Zhou et al. 2001); (5) to assist in deciphering the speech of people with various accents (Biadsky, 2011); (6) in estimating a speaker's age (Bocklet et al. 2008) and gender (Vogt

and André, 2006); and (7) for spoken language translation (Schultz and Waibel, 2001), spoken document retrieval (Chelba et al. 2008), and multilingual speech recognition (Kannan et al. 2019).

In the area of construction, Zhang et al. (2018a) used a speech recognition framework to analyze onsite conversations. Their framework used a naïve Bayes classifier to translate speech captured on site into text scripts, and to further classify the text scripts into construction activities and operations. The framework achieved an overall accuracy of 90.9%. In another study, Zhang et al. (2018b) developed a supervised machine learning-based sound identification framework, using it to identify six different sounds common to construction sites (concrete-grinding, hammering, concrete-pouring, drilling, excavator operation, and dozer operation). The framework achieved a maximum accuracy of 94.3%. Speech recognition has also been used to retrieve BIM data from BIM software (Shin and Issa, 2021) and to identify heavy construction equipment operating at a construction site (Cheng et al. 2017). The use of speech recognition systems to identify keywords in speech communication on construction sites, however, has yet to be explored. In this research, therefore, a keyword identification framework to facilitate communication by identifying keywords related to crane operations in noisy construction sites is developed. The framework is also capable of identifying keywords in speech by workers with different accents.

5.3 Research methods

In this research, a crane signalman speech dataset is collected using an audio-recording device affixed to the crane signalman's helmet. The speech dataset is preprocessed by adding representative construction noises, and data augmentation is implemented by altering the pitch and adding random noise to the speech. This helps to generalize the dataset and reduces the likelihood of model overfitting, per Lei et al. (2019). The waveform speech is then converted into Mel-frequency cepstral coefficients (MFCCs) in order to extract unique features from the speech dataset

(Mahmood and Utko, 2021). MFCCs, it should be noted, are the most widely used feature extraction algorithm in the field of speech recognition. The purpose of using MFCCs is to reduce the complexity of the model and achieve higher accuracy (Mahmood and Utko, 2021). The extracted features obtained using MFCCs are then used to train the one-dimensional convolutional neural network (1D CNN) classifier to identify the keywords in the speech. The 1D CNN model is validated, tested, and deployed for real-time identification of keywords in speech as an output. An overview of the keyword identification framework is given in Figure 5-1.

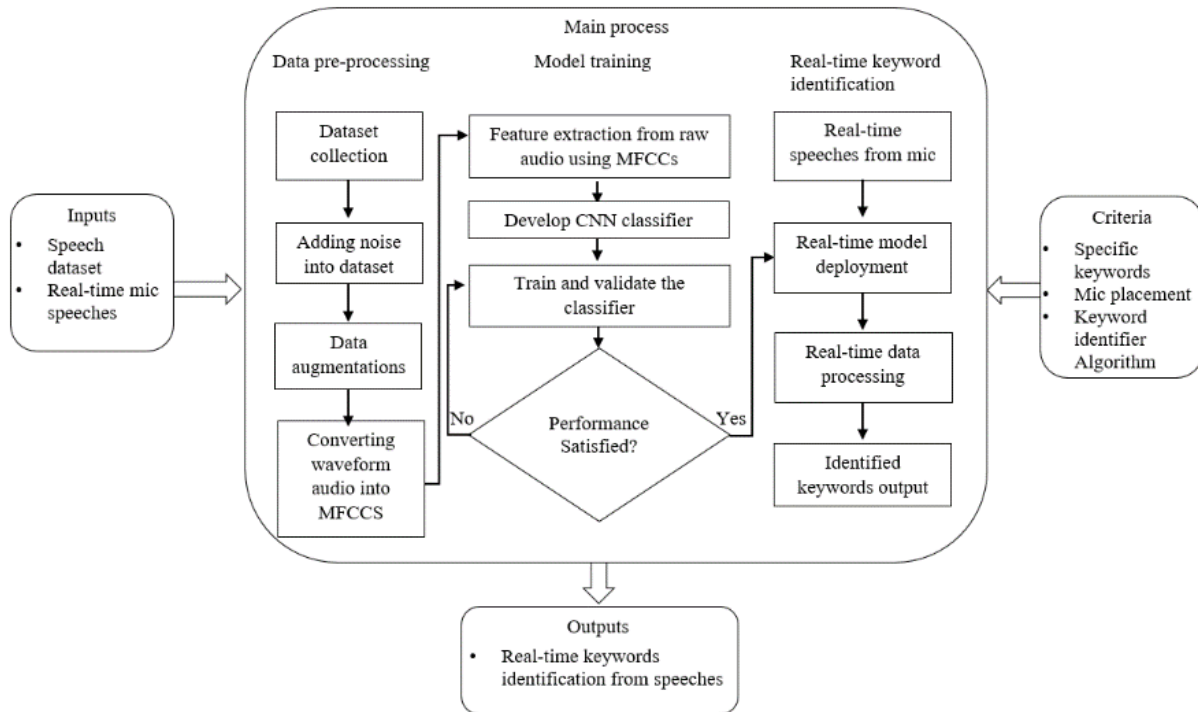


Figure 5-1. Overview of keyword identification framework.

5.4 Implementation and case study

The framework is implemented and tested using the crane signalman speech commands used to guide the crane operator in the crane operations on the construction site (see Figure 5-2).

Boom up and lower the load	Boom down	Dog everything	Hoist	Move slowly	Swing
Boom down and raise the load	Boom up	Emergency stop	Lower	Stop	Travel
Use main hoist	Telescope out	Travel both tracks	Travel one track	Telescope in	Use whipline

Figure 5-2. Crane signalman speech command/keywords.

5.4.1 Dataset collection and preprocessing

With no existing dataset for crane signalman speech commands available, the speech command data is collected manually at a 16,000 Hz frequency using an audio-recording device affixed to the crane signalman’s helmet. The dataset collected contains 12 hours of crane signalman speech commands made by 45 volunteers (30 male, 15 female) representing 13 different ethno-linguistic backgrounds and having different accents. The dataset is resampled into 2-second duration speech files for each command, referred to as a “keyword”. Each sample is further normalized to adjust the range of speech, equalized to remove bumps from the speech, and compressed to modify the range of loudness of the speech. Furthermore, construction site-related noise, collected from the Mixkit (2022) and Zapsplat (2022) datasets, is added to the speech. The incorporation of construction site noises helps to generalize and reduce the likelihood of model overfitting. Data augmentation is then applied to artificially alter the pitch of the speech.

5.4.2 Model development

The set of 21,600 samples of 2-second audio commands representing a total of 12 hours of data is converted from waveform into MFCCs, which are capable of representing the amplitude spectrum of the sound wave in a compact vectorial form (De Pinto et al. 2020). In this technique, it should be noted, the audio file is divided into frames, usually using a fixed window size, in order to obtain statistically stationary waves and, in turn, frames. The frames having been obtained, discrete Fourier transform is applied, and only the logarithm of the amplitude spectrum is retained. The amplitude spectrum is normalized with a reduction of the Mel frequency scale. This operation is

executed for the purpose of identifying the frequencies (Logan, 2000). The interested reader may refer to Davis and Mermelstein (1980) and Huang et al. (2001), in which the MFCC calculations are thoroughly explained. To extract features from waveform audio, the main parameters used in MFCCs are the number of coefficients (referred to as the static features) that contain the information in a given audio frame, the fast Fourier transform length (which represents the number of samples in each window), the number of filters (which reflects the number of features extracted from the audio file), the frame stride, and the frame length. The values chosen for the developed framework are given in Table 5-1. These values are selected in a trial-and-error based on the performance of the model. Figure 5-3, shows examples of waveform audio and corresponding MFCCs.

Table 5-1.

Parameter of MFCCs.

Parameters	
Number of coefficients	40
Fast Fourier transform length	512
Number of filters	40
Frame stride	0.02
Frame length	0.02

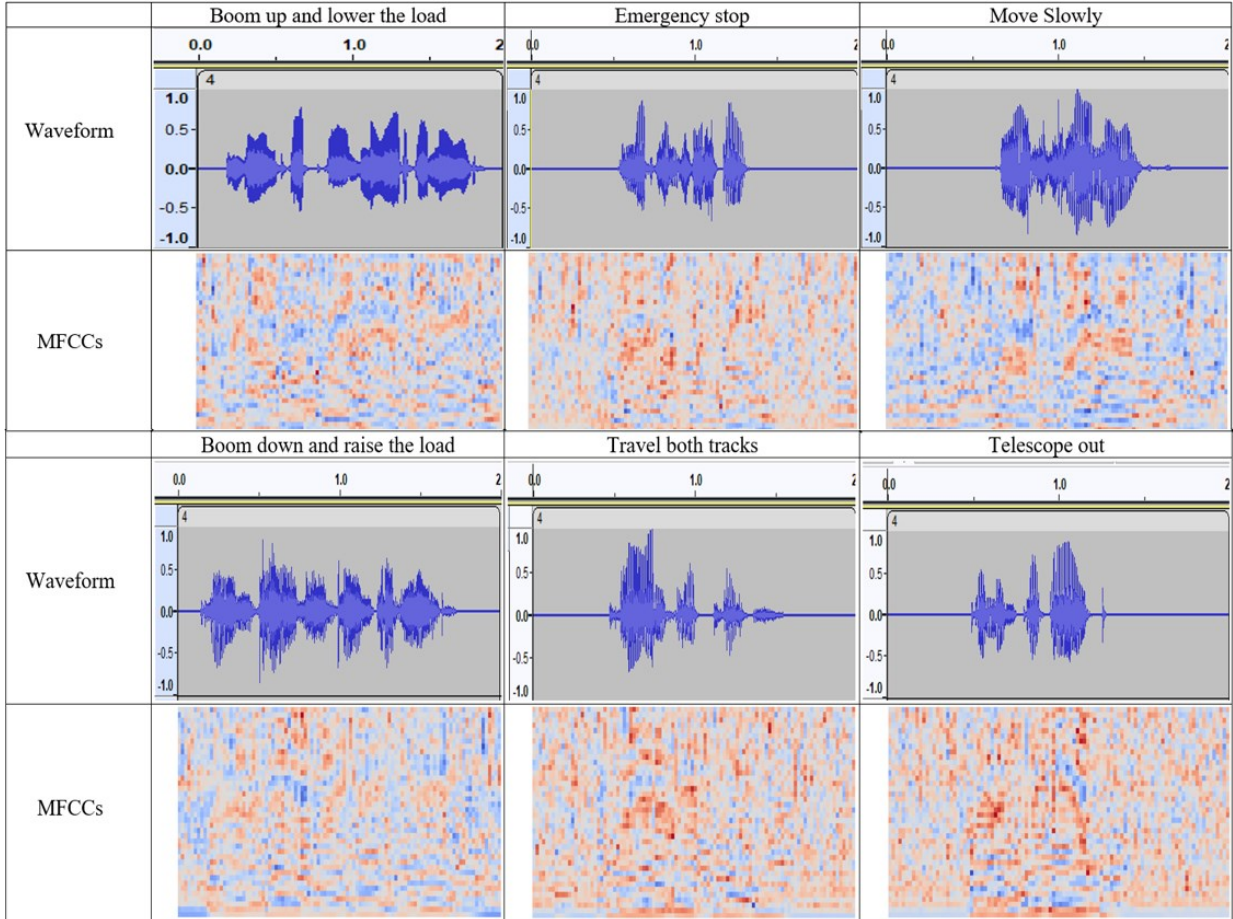


Figure 5-3. Waveform audio sample of keywords and corresponding MFCCs.

The features extracted from the MFCCs are then fed into 1D CNN classifiers, which a classifier can operate on vectors of the features for each audio file provided as input. Here, the values represent the compact numerical form of the audio frames of 2-second duration. The compact numerical form of the audio frame is input to the 1D CNN, the architecture of which includes four convolutional layers that are responsible for extracting and learning features from the input data. Each convolutional layer is followed by an activation function to add non-linearity to the output neuron. For this work, the rectified linear unit (ReLU) activation function is used in the convolutional layers. The ReLU activation function, it should be noted, is an identity line for which $y = x$ for all positive lines and $y = 0$ for all negative values. Each convolutional layer and activation

function is followed by a pooling or subsampling layer and dropout layer. The pooling layer helps the model to focus on the principal characteristics of each portion of speech data, making the portion of data them invariant by their position, while the dropout layer activates and deactivates the neurons with respect to their weights. (This technique helps to better generalize the predictive capabilities of the model.) The output from the dropout layer is then flattened to make it compatible with the subsequent layers. Finally, a SoftMax activation function is applied to one dense layer (i.e., fully connected layer) in order to estimate the probability distribution of each of the classes properly encoded in the model.

5.5 Results

The dataset is randomly split into training, validation, and test sets. The proportion of the training set is kept at 80% (17,280 samples) while the validation and test sets are kept at 10% (2,160 samples) each. The reason for using more samples in the training set is to allow the 1D CNN model to learn more features from the dataset (as this will lead to a more accurate model for identifying keywords in the validation and test sets). The 1D CNN model is trained on 100 epochs while keeping the model-learning rate to 0.005 and the dropout rate to 0.25. The number of epochs, values of learning, and dropout rate are determined experimentally to boost the accuracy of the validation and testing. The model is found to achieve average accuracy of 97.3% and 96.1% and losses of 0.12 and 0.18 in the training and validation processes, respectively. Moreover, the model is found to achieve an accuracy of 93.8% in the test set. The accuracy and loss are measured using Equations 1 and 2.

$$Accuracy = \frac{\text{Correctly identified keywords in speech}}{\text{Total number of keywords in speech}} \times 100 \quad (5.1)$$

$$\text{cross entropy loss} = \frac{-1}{N} \times \sum_{x=1}^N \sum_{y=1}^M Z_{xy} \times \log(p_{xy}) \quad (5.2)$$

where N is the number of samples and M is the number of classes; Z_{xy} specifies whether or not sample x belongs to class y ; and p_{xy} represents the probability of sample x belonging to class y . The loss has no upper limit and falls within the range $[0, \infty]$, where a value of loss near 0 indicates high accuracy.

The model is then deployed for real-time keyword identification from live speech captured using an audio-recording device affixed to the signalman's helmet. Based on 650 iterations, the model is found to achieve an overall accuracy of 95.3% in real time. This result demonstrates that the developed model is capable of accurately identifying keywords in speech in the context of a construction site environment. As such, the model can be considered suitable for use as an additional layer of communication on noisy construction site.

5.6 Conclusion

In this chapter, a keyword identification framework is developed that is capable of identifying 18 different crane signalman speech commands (i.e., "keywords"), in real time. First, a dataset of 12 hours of crane signalman speech commands is collected using an audio-recording device affixed to the signalman's helmet. Construction site noise is then added to generalize the dataset. Short audio clips of 2-second duration (i.e., the approximate duration of a keyword/command) are then separated from the dataset, and features are extracted from the audio dataset using MFCCs. The extracted features are used as an input to train the 1D CNN model, which is found to achieve training, validation, and testing accuracies of 97.3%, 96.1%, and 93.8%, respectively. The model is further validated in the real-time identification of keywords in live speech, achieving an accuracy

of 95.3%. In future work, more data will be added, and the model will be further optimized to improve its accuracy in performing real-time keyword identification.

Chapter 6: AN ENSEMBLE TECHNIQUE TO IMPROVE THE ACCURACY OF ONSITE COMMUNICATION CLASSIFICATION. A CASE OF COMMUNICATION BETWEEN THE CRANE OPERATOR AND CRANE SIGNALMAN.

Crane operations are an essential aspect of many construction sites. However, they also pose a significant risk to worker safety if there is not effective communication between crane operator and signalman. In this regard, the previous chapters have described the use of various technologies to facilitate effective communication in crane operations by classifying hand signals and speech communication.

In this chapter, an ensemble model approach is proposed to improve the accuracy of communication classification between crane signalmen and operators in crane operations. The ensemble model uses the models proposed in the previous chapters and combines their respective outputs to create a more robust and accurate classification system. The proposed ensemble model approach has several advantages over traditional methods, including higher accuracy, increased reliability, and enhanced flexibility. In this chapter, two ensemble techniques—weighted average and majority voting—are used to combine the classification decision outputs of the models described in the previous chapters and produce a single, more accurate classification decision output. The ensemble models are evaluated, and all ensemble models are found to outperform the individual models in terms of accuracy. Among the several ensemble models, the hard majority voting ensemble model is found to achieve the highest accuracy in crane signalman and operator communication classification (99.40%), suggesting that this model could be used on construction sites to add a supplement layer of communication in crane operations.

6.1 Introduction

In current practice, the primary modes of communication among workers on construction sites are direct verbal communication, hand signalling, and two-way radio communication. However, the reliance on these communication methods is increasingly untenable, given the complex and congested nature of modern construction sites. In previous chapters, three different methods of classifying different data to facilitate communication on construction sites were proposed: YOLOv4+LSTM to classify dynamic crane signalman hand signals in real time using cameras, a sensor-based smart construction glove (SCG) to classify crane signalman hand signals in real time using sensor data, and a 1D CNN speech recognition model to identify the speech commands of crane signalman. These models are proposed to add more layers of communication to the current practice with classification capabilities that are achieved with the help of advanced technologies (as described in detail in previous chapters). These classification models have been found to achieve high accuracy in classifying dynamic crane signalman hand signals and speech data. However, to improve the performance of the framework, in this chapter the concept of redundancy—where the models developed as discussed in the previous chapters are integrated with the aid of ensemble models in which weighted average and majority voting are used as classification criteria for the final decision output—is proposed.

Ensemble modelling, it should be noted, is a machine-learning approach that combines multiple models to provide a more accurate and robust prediction system. The idea behind this approach is that combining multiple models can produce better results than any single model operating alone. This is due to the fact that each model has various strengths and weaknesses, and combining them can reduce the adverse impact of the limitations of each of the constituent models making up the ensemble model. In ensemble modelling, similar or different datasets can be used to train multiple

models, and their output predictions are combined to produce a more accurate and stable decision output prediction. Majority voting is an ensemble technique that predicts the new datapoint based on the majority class label predicted by the individual models and produces a single classification decision. Majority voting is commonly used in classification problems to increase the robustness and accuracy of the model prediction. The objective of the research described in this chapter is to combine the decisions of multiple modalities (camera, sensor, and speech data) to enhance the performance and reliability of the classification system.

The remainder of this chapter is organized as follows: the next section describes the relevant literature on the application of majority voting and weighted average ensemble models in various industries, including the construction industry. This section is followed by a description of the methodology. Next the different types of ensemble models used in this research are introduced and their implementation described, followed by a summary of the results of each ensemble model. The chapter then closes with a discussion and conclusion.

6.2 Related work

The use of majority voting in ensemble modelling has been thoroughly researched and published in several articles. These studies have demonstrated that the majority voting ensemble model can increase the accuracy and reliability of decision-making systems when compared to the individual models it comprises. For example, Attallah and Al-Mousa (2019) presented a majority vote ensemble approach for predicting the existence of heart disease in humans. They trained and evaluated four machine-learning models—Stochastic Gradient Descent (SGD), k -Nearest Neighbour (KNN), Random Forest (RF), and Logistic Regression classifier—then combined all four machine-learning models in an ensemble method in which the majority voting classification model is used. The ensemble model was found to outperform the individual models in terms of

accuracy. Mehanović et al. (2020), Raza (2019), and Kumari et al. (2021) also used majority voting ensemble models to improve the prediction of heart disease and diabetes mellitus, with each of these studies also achieving favourable results in terms of the accuracy of the developed ensemble model. Neloy et al. (2022) achieved a weighted average ensemble model by combining three machine-learning models—random forest, decision tree, and naïve bayes—to predict heart disease. Their weighted average ensemble model achieved 100% accuracy in training, and was found to outperform six different machine-learning models.

Mukerjee et al. (2020), meanwhile, proposed an ensemble model (EnsemConvNet) that combines CNN-Net, Encoded-Net, and CNN-LSTM models to classify human activities such as running, walking, and cooking. They evaluated the developed model by comparing it with existing individual deep-learning models such as CNN-LSTM, with the results demonstrating the superiority of the EnsemConvNet model over the individual models it comprised. In a similar vein, Irvine et al. (2019) proposed a neural network ensemble approach to recognize human daily activities. The performance of the proposed model was evaluated using two non-parametric benchmark classifiers—KNN and Support Vector Machine (SVM), with the developed ensemble model being found to outperform the KNN and SVM models. The ensemble approach has also showed better accuracy than the individual models in applications ranging from facial recognition (Choi and Lee 2019; Renda et al. 2019), to emotion recognition (Wei et al. 2020) and sentiment analysis (Alrehili and Albalawi 2019).

In the construction industry, Zhang et al. (2019) proposed an ensemble approach to classifying the causes of construction site accidents from construction site reports using text mining and natural language processing techniques. They compared the proposed ensemble model with individual models such as support vector machine (SVM), linear regression (LR), KNN, decision tree (DT),

Naïve Bayes (NB). The results indicated that the proposed ensemble model outperformed the individual model in the classification of construction site accidents. George et al. (2022) used ensemble and individual machine-learning models to predict the risks of the construction sites and noted that the ensemble model achieved superior accuracy. The ensemble models achieved higher accuracy in predicting concrete tensile strengths (Hu et al. 2022), predicting the compressive strength of fly ash concrete (Barkhordari et al. 2022), Estimating the Residual Value of Heavy Construction Equipment (Milošević et al. 2021), Prediction of Unit Price Bids of Resurfacing Highway Projects (Cao et al. 2018), to name a few.

These studies demonstrated that the ensemble model outperformed the individual models in every application. Despite the potential benefits, though, the application of ensemble models to classify hand signals and speech communication in construction sites has yet to be extensively explored. Therefore, this chapter proposes a redundancy-based ensemble model to classify hand signals and speech communication in the construction industry.

6.3 Methodology

The overview of the proposed framework is given in Figure 6-1.

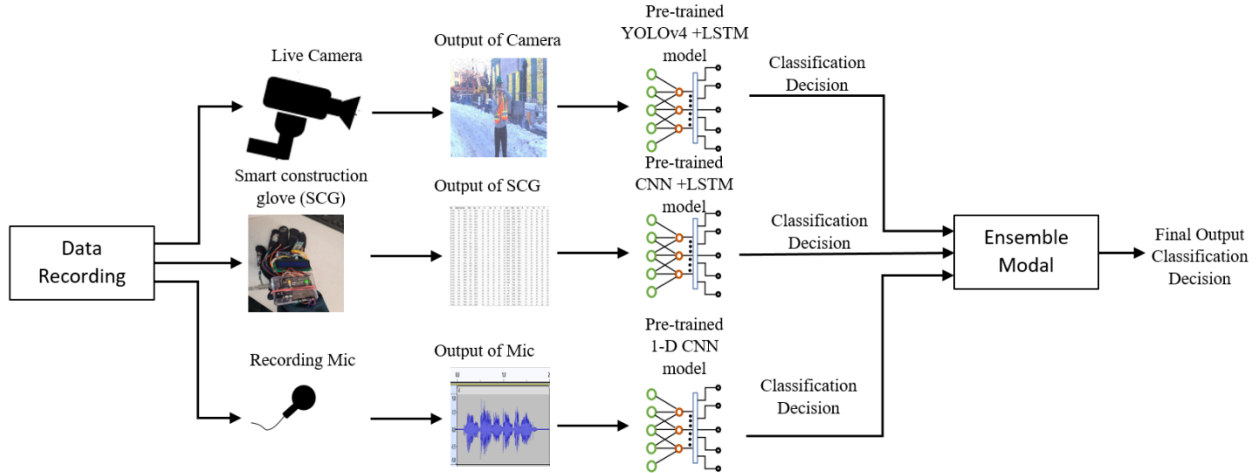


Figure 6-1. overview of the proposed ensemble framework.

In this framework, initially data is recorded using three different techniques. Cameras placed in multiple positions and angles as shown in Figure 3-3 are used to capture the camera live video of the crane signalman's dynamic hand signals. A sensor-based SCG is used to capture the sensor data including the orientation of the crane signalman's hand and bend the fingers of the crane signalman. An audio-recording device affixed to the crane signalman's helmet is used to record the speech command used by the crane signalman to guide the crane operator in the crane operations.

Furthermore, pre-trained models including YOLOv4+LSTM model discussed in detail in Chapter 3 is deployed to classify the dynamic crane signalman hand signals from camera live feeds. CNN+LSTM model discussion Chapter 4 is deployed to classify the dynamic crane signalman hand signals with the aid of sensor data from an SCG. 1D CNN model is deployed to classify the crane signalman speech commands to the crane operator captured by an audio-recording device

affixed to the crane signalman’s helmet. The classification results of each of the models discussed above are then integrated using an ensemble model to generate a single, more accurate and robust final output.

6.4 Ensemble models

In this framework, two types of ensemble models—weighted average and majority voting—are used. The algorithm underlying the ensemble model used in this framework is presented as Algorithm 2 below.

Algorithm 2

1. **Initialize (M_n):**
 2. Pre-trained YOLOv4+LSTM model (M_1)
 3. Pre-trained CNN+LSTM model (M_2)
 4. Pre-trained 1-D CNN model (M_3)
 5. **INPUT (X_n):**
 6. Video data (X_1)
 7. Sensor data (X_2)
 8. Speech data (X_3)
 9. **Process:**
 10. **for** each X in stream
 11. **do** collect data (X_1 , X_2 , and X_3)
 12. **Fed** X_1 to M_1 , X_2 to M_2 , and X_3 to M_3
 13. **do** predict P_1 , P_2 , and P_3 from M_1 , M_2 , and M_3
 14. Ensemble the prediction (C) = (P_1 , P_2 , and P_3)
 15. **Output:**
 16. Final output (C)
-

6.4.1 Weighted average ensemble model

The weighted average ensemble is a type of ensemble model that combines predictions from several models by assigning varying weights to each model's output prediction, where the weight signifies the importance or confidence of each model's prediction. The final output prediction is then calculated as a weighted average of the individual model predictions. This method can be used to enhance the performance of a single model by combining the strengths of multiple models

and mitigating the models' respective weaknesses. The model weights are positive values ranging between 0 and 1, and the sum of the weights must be equal to 1. To find the optimum weights of each model, this framework uses model performance-based weights selection. In this technique, higher weights are assigned to the model that performed better. In this framework, the optimum weights of the individual models are 0.30, 0.33, and 0.37 for YOLOv4+LSTM, CNN+LSTM, and 1D CNN, respectively. A higher weight assignment means that the given model is more reliable due to its higher accuracy in crane signalman hand signal and speech classification. Figure 6-2 shows an overview of the developed weighted average ensemble model.

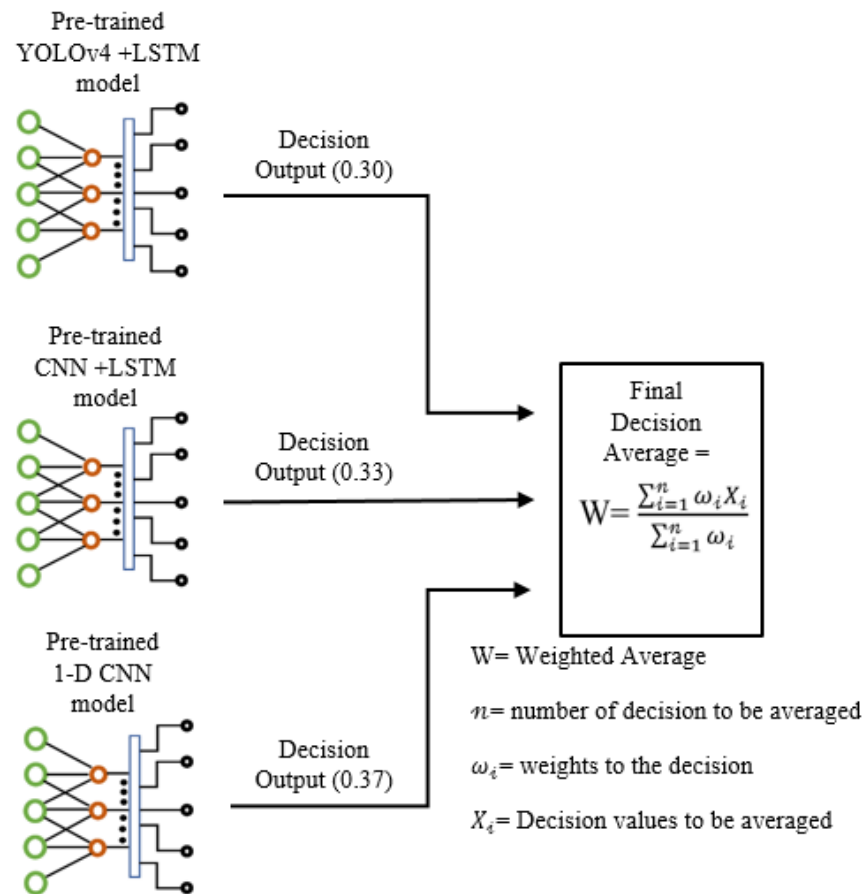


Figure 6-2. Overview of the developed weighted average ensemble model.

6.4.2 Majority voting ensemble model

Majority voting is an effective and simple technique used to combine the predictions of several models and provide a single accurate and robust output prediction. This technique functions in such a way that each model taking part in the ensemble makes a prediction for an input and the majority voting technique makes the final prediction by choosing the class that has received the most votes from the model. This technique capitalizes on the strengths of the various models making up the ensemble by combining their predictions to achieve a higher accuracy than any of the individual models it comprises. There are two main types of majority voting techniques—soft voting and hard voting. These are discussed in detail in the following subsections.

6.4.2.1 Soft voting

Soft voting is a type of majority voting technique in ensemble modelling that uses the prediction probabilities of each class in the model and calculates the average of the prediction probability score of each class in the model to achieve one final output decision. For example, suppose three models are used to make a prediction of a class of new input data based on soft majority voting techniques. Model-1 predicts the inputs as Class A = 60%, Class B = 20%, and Class C = 20%. Model-2 predicts the class of new inputs as Class A = 80%, Class B = 20% and Class C = 0%. Model C classifies the inputs as Class A = 20%, Class B = 50%, and Class C = 30%. In the above case, the average probabilities of Class A, B, and C are 53%, 30%, and 17%, respectively, and the final output will be Class A, since it has the highest probability. An overview of soft majority voting is provided in Figure 6-3.

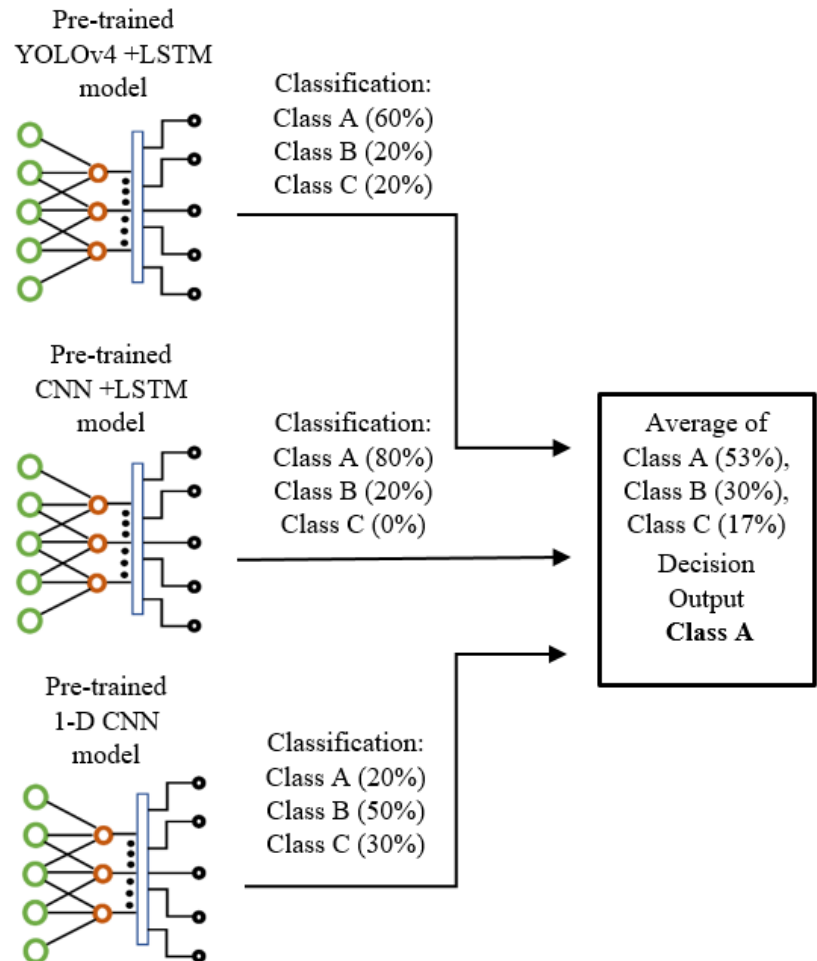


Figure 6-3. Overview of soft majority voting.

In the present research, soft voting is carried out in two scenarios. In the first scenario, it is assumed that the crane signalman is only using hand signals (i.e., the signalman is not using speech commands to guide the crane operator), and that the data is coming from vision- and sensor-based classification models. In the second scenario, it is assumed that the crane signalman is using hand signals and speech commands at the same time, and that data is coming from vision, sensor, and speech classification models.

6.4.2.2 Hard voting

In the hard majority voting technique, the ensemble model makes a classification prediction based on the number of votes received by the individual label/class as the final decision output. In the hard majority voting technique, two or more similar or dissimilar models are used to generate individual decisions, and the output decisions of each of the individual models are then combined in an ensemble hard majority voting model that generates a single prediction output. An overview of hard majority voting is provided in Figure 6-4.

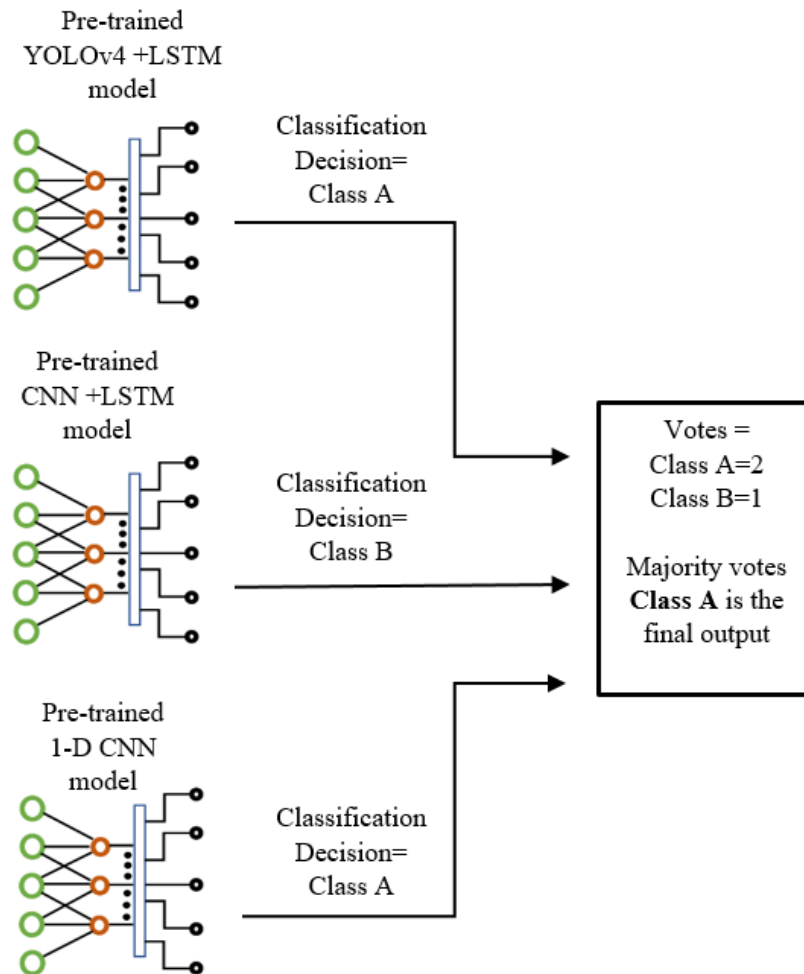


Figure 6-4. Overview of hard majority voting.

In the ensemble hard majority voting technique, the number of votes received by each class predicted by the individual models is counted, and the class with the most votes is the output. In the present research, three different individual models are trained on three different types of datasets—i.e., the YOLOv4+LSTM model is trained on the video dataset, the CNN+LSTM model is trained on the sensor dataset, and the 1D CNN model is trained on the speech dataset—to classify the crane signalman hand signals and speech commands. Each model produces a decision output based on the given input data, and the hard majority voting technique is applied to count the number of votes of each class predicted by the individual models and produce the final output decision accordingly. It should be noted that, in this technique, the model is developed in such a way that the final output decision is of the class that receives the most votes. In the case that the decision outputs are different, meaning that each model produces a different output decision, the model will not produce any output decision.

6.5 Results

Three individual pre-trained models—YOLOv4+LSTM, CNN+LSTM, and 1D CNN, trained on video data, sensor data, and speech data, respectively—are deployed and combined in an ensemble to provide improved classification accuracy in classifying crane signalman hand signals and speech commands and generate a single decision output. The respective performance of three different ensemble models—weighted average ensemble, soft majority voting, and hard majority voting—is evaluated on a new test dataset of 18,270 datapoints to which the models have not yet been exposed. The evaluation metric used to evaluate the performance of the models is accuracy. The accuracy calculation is expressed in Eq. (3.4) (see Chapter 3).

6.5.1 Results of weighted average ensemble model

The individual models used in the ensemble have different architectures and hyperparameters and different datasets to increase their diversity and improve the model performance. The predictions of the individual models are combined using a weighted average ensemble technique. Based on the assigned weights discussed above, the model is found to achieve an average overall accuracy of 99.05% on a held-out test set, outperforming each of the individual models and achieving state-of-the-art performance on the dataset. Figure 6-5 shows the accuracy achieved for each hand signal using the weighted average ensemble model.

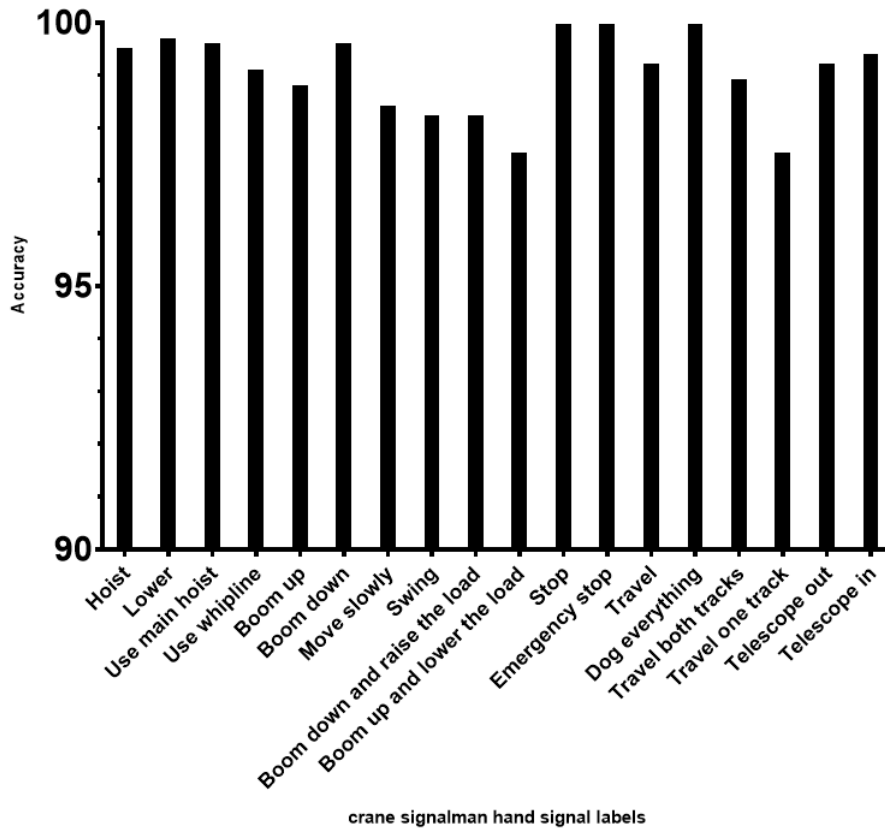


Figure 6-5. Performance of weighted average ensemble model.

It can be seen from Figure 6-5 that the model is found to perform well for all classes, with an accuracy of over 98% for all classes with the exceptions of *Boom up and lower the load* and *Travel one track*, which are slightly lower (97.54% and 97.49%, respectively). Overall, the ensemble model is found to have achieved high accuracy and to have outperformed the individual models, demonstrating the effectiveness of combining multiple models to improve performance.

6.5.2 Results of Soft Majority Voting ensemble model

As discussed in Subsection 6.4.2.1, the soft majority voting is applied to two different scenarios. Figure 6-6 shows the performance achieved by the model when the sensor- and vision-based models are combined to classify crane signalman hand signals. In this ensemble model, the classification model for crane signalman speech commands is not combined with the sensor- and vision-based models. The model is found to achieve an overall classification accuracy of 94.8% in the test dataset, with *Dog everything* achieving the highest classification accuracy (98.22%) and *Swing* achieving the lowest accuracy (90.73%) in classifying hand signals. However, the ensemble model is found to outperform the individual models in classification accuracy.

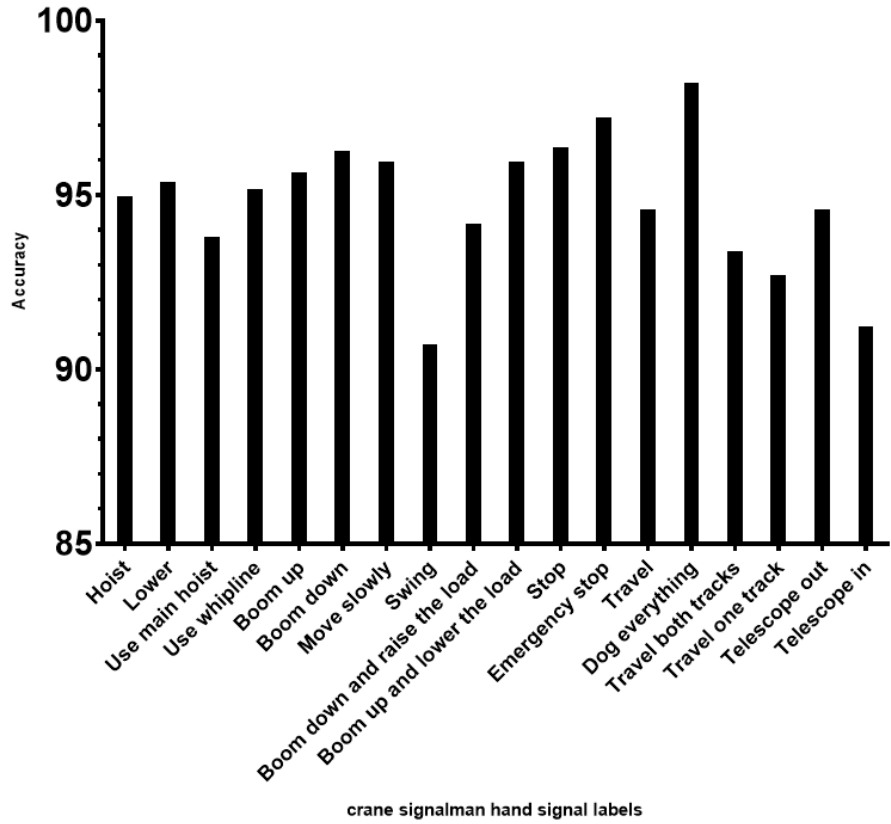


Figure 6-6. Performance of Soft majority voting (2-medium) ensemble model.

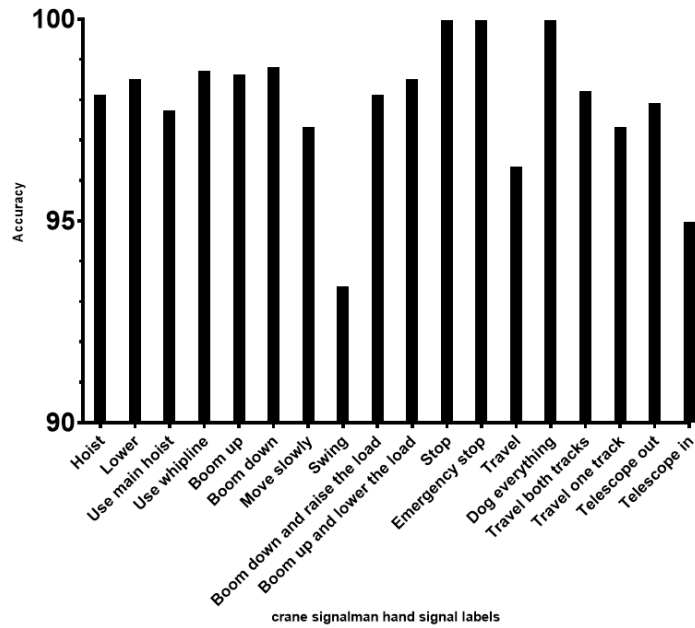


Figure 6-7. Performance of Soft majority voting (3-medium) ensemble model.

In the second soft majority voting ensemble model, all three models for classifying crane signalman hand signals and speech commands are combined and evaluated. This model is found to perform better than the soft majority voting ensemble model in which just two models are combined, achieving an overall classification accuracy of 97.92% on the test dataset. Figure 6-7 shows the performance of the model. From Figure 6-7, it can be seen that some hand signal labels, such as *Stop*, *Emergency stop*, and *Dog everything*, are found to achieve 100% classification accuracy.

6.5.3 Results of hard majority voting ensemble model

The hard majority voting ensemble model is evaluated on the test dataset. In the hard majority voting ensemble model, the output decisions of the three proposed models—YOLOv4+LSTM, CNN+LSTM, and 1D CNN—are combined and, based on majority voting, the final decision is determined. The hard majority voting ensemble model is found to achieve an overall classification accuracy of 99.45%. Figure 6-8 shows the classification accuracy of each hand signal and speech command used by the crane signalman to guide the crane operator. From Figure 6-8, it can be seen that all the classes are found to achieve accuracies higher than 98%, with 8 classes of the 18 classes achieving 100% classification accuracy.

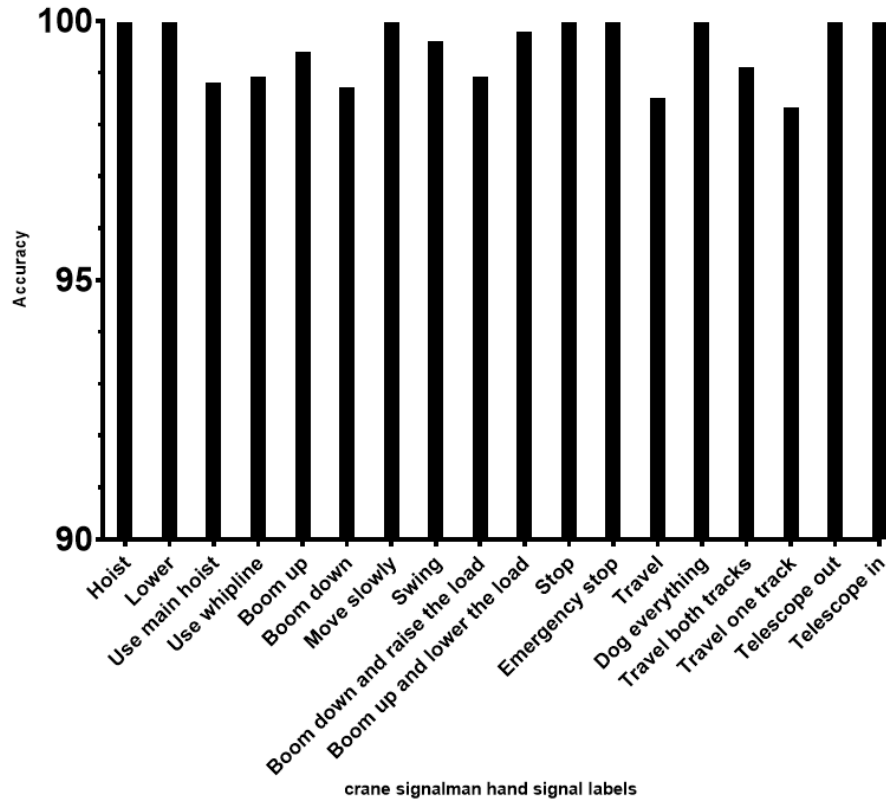


Figure 6-8. Performance of hard majority voting ensemble model.

6.6 Discussion

In this chapter, crane signalman hand signals and speech commands are classified using an ensemble of three individual pre-trained models, namely, YOLOv4+LSTM, CNN+LSTM, and 1D CNN, with each being trained on a different dataset (video, sensor, and speech, respectively). The performance of each of the three different ensemble techniques—namely, weighted average, soft majority voting, and hard majority voting—is evaluated on a separate test dataset of 18,270 datapoints to assess the classification accuracy on the test dataset. Figure 6-9 demonstrates the classification accuracy achieved by each of the ensemble models considered. From Figure 6-9, it can be seen that the highest accuracy is achieved by the hard voting majority ensemble model with an average classification accuracy of 99.45%, while the model with the lowest classification

accuracy is the soft majority-voting model (2-medium—combining camera and sensor data only) with an average classification accuracy of 94.8%. Figure 6-10, meanwhile, compares the classification accuracy of each of the ensemble models with respect to each hand signal.

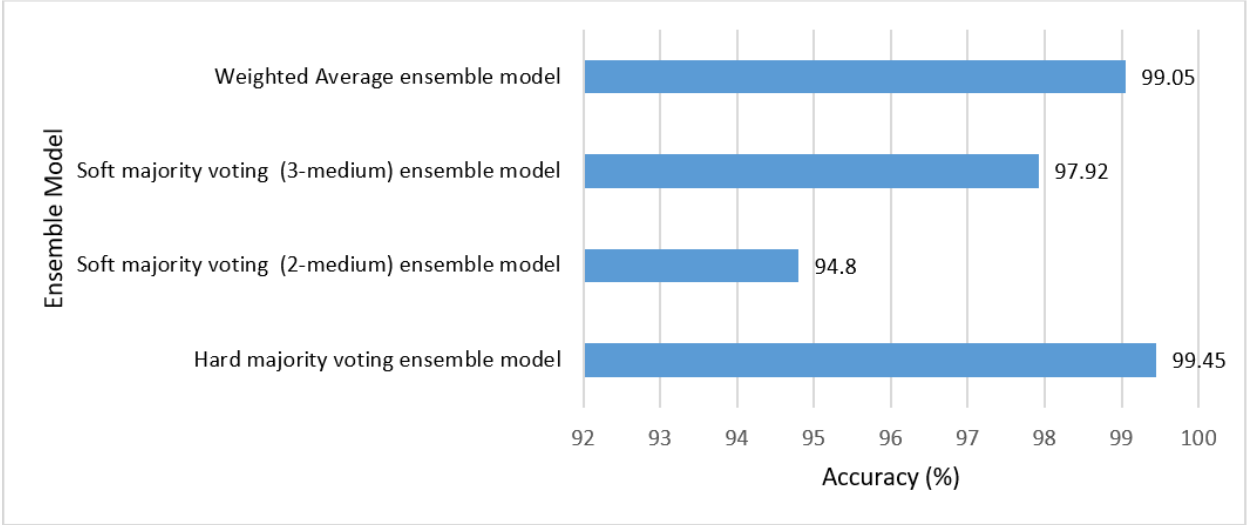


Figure 6-9. Performance of proposed ensemble model on test dataset.

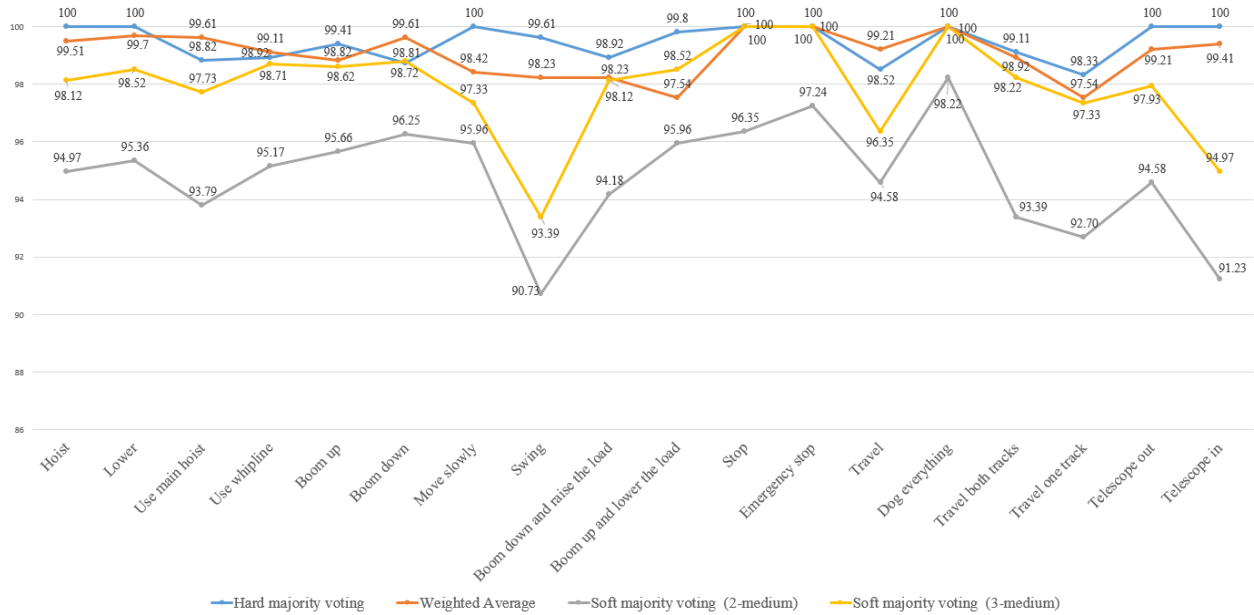


Figure 6-10. Comparison of the classification accuracy of the ensemble models with respect to each hand signal.

Overall, the results of this study demonstrate the effectiveness of ensemble models in improving the performance of classification tasks. It is demonstrated that combining multiple models with different architectures, hyperparameters, and datasets can increase the diversity of the models and lead to improved performance. The study also highlights the importance of evaluating different ensemble techniques to identify the most effective approach for a given task.

6.7 Conclusion

In this chapter, three different ensemble models for the classification of crane signalman hand signals and speech commands are proposed. The proposed ensemble models are composed of three different deep-learning models: YOLOv4+LSTM to classify dynamic crane signalman hand signals using cameras, a sensor-based model to classify crane signalman hand signals using sensor data from an SCG, and a 1D CNN model to classify crane signalman speech commands. Each

model is trained on a different dataset, and their respective decisions are combined to produce a single classification decision. Each ensemble model is evaluated based on the average classification accuracy. The weighted average ensemble technique is found to achieve an average overall accuracy of 99.15% on the held-out test set. The performance of the soft majority voting ensemble model is also evaluated, based on two different scenarios. In the first scenario, in which crane signalman hand signals are classified using a combination of sensor- and vision-based models, the overall classification accuracy is found to be 94.8%. In the second scenario, in which all three models are combined to classify both hand signals and speech commands, the overall classification accuracy is found to be 97.9%. Finally, the performance of the hard majority voting ensemble model is evaluated the, with the model achieving an overall classification accuracy of 99.45%.

Chapter 7: CONCLUSIONS

7.1 Research Summary

Effective communication between crane operator and signaller is essential for safety, coordination, productivity, quality, and cost management in construction projects involving crane operations. It is critical that communication is open and frequent, and that both have the information they need to perform their tasks safely and effectively. The current practice is to rely on hand signals to send commands to the crane operator in most situations, and to use two-way radio speech communication when hand signalling is not possible. However, these means of communication are not reliable in modern construction due to site congestion, visibility issues, background noise, language barriers, and other distractions. In this regard, recent technological developments can be leveraged to achieve more effective communication and improved safety and productivity in the construction sites. This research thus employs advanced technologies such as deep learning in a multimodal communication classification framework that can assist in improving communication, safety, and productivity on the construction site.

As described in Chapter 2, a detailed scientometric and critical review of the literature on the application of deep learning in the construction industry is conducted. In the scientometric analysis, co-word, co-author, co-citation and cluster analysis is performed. Co-word analysis is used to analyze and visualize the course of development of the body of knowledge, co-author analysis is used to identify the dominant countries and researchers working in this research area, co-citation analysis is used to detect similarity in the literature, and cluster analysis is applied to identify themes in the existing textual data. The purpose of the critical review, meanwhile, is to analyze the depth of the literature. Based on this review, it is observed that deep learning has been successfully deployed in the construction industry to monitor construction machinery, construction

site activities, and construction site reports, and to assess the worker ergonomics on construction sites. Deep learning has also been used for safety and productivity improvement on construction sites. However, multiple challenges, including lack of data availability, inaccuracy of the deep-learning models, data privacy issues, computational cost, and the dynamic nature of the construction site environment, are also identified in the critical review. Moreover, it is observed that the use of deep-learning models for communication classification in construction sites has yet to be thoroughly explored, yet is a matter of pressing concern considering that communication errors are responsible for disastrous accidents in construction sites. Chapter 2 thus proposes the use of multiple deep-learning models for communication classification in construction sites, drawing upon the case of communication between crane operator and crane signalman.

In Chapter 3, a deep learning-based classification framework to reduce communication errors in dynamic hand signalling for crane operations is proposed. The objective of this framework is to add a supplementary layer to the communication between crane operator and signalman with the aid of computer-vision-based deep learning hand signal classification models. It begins with the collection of dynamic crane signalman hand signal data, since there is no publicly available dataset of dynamic crane signalman hand signals. Then, two state-of-the-art models, YOLOv4 and LSTM, are deployed and modified to achieve the final classification decision. The YOLOv4 model is modified to collect spatial features from the dynamic crane signalman hand signal dataset, while the LSTM model is deployed to extract temporal features and to predict the final classification decision. The model is found to achieve high accuracy in classifying dynamic crane signalman hand signals in complex construction sites and under variable weather conditions in real time.

Chapter 4 describes a framework for the development of a smart construction glove (SCG) furnished with sensors and deployed in conjunction with machine-learning algorithms to classify

crane signalman hand signals. In this framework, sensors affixed to the SCG are used to measure the orientation of the crane signalman's hand during the execution of hand signals. A complementary filter fusion algorithm is used to remove sensor data noise, and the data collected from the SCG is transmitted to an Excel file with the aid of a Bluetooth module affixed to the SCG. The collected dataset is labelled manually according to the classification labels. Four machine-learning models—KNN, SVM, DT, and CNN-LSTM—are then trained on the collected sensor datasets. Finally, a mobile application is developed, and the top-performing models are deployed in the developed app to classify dynamic crane signalman hand signals in real time. The models are found to achieve high accuracy in hand signal classification, and thus could be used as an additional layer of communication to reduce the likelihood of communication errors.

The framework discussed in Chapters 3 and 4 uses vision- and sensor-based techniques to classify crane signalman hand signals, whereas Chapter 5 proposes a framework to identify speech commands used in the communication between crane signalman and crane operator in crane operations. This framework begins with speech data collection in waveform, and, for this purpose, an audio-recording device is affixed to the crane signalman's helmet. The collected data is preprocessed by adding construction site noises and data augmentation. The waveform preprocessed data is then converted to MFCCs in order to extract features from the datasets, and a 1D CNN model is developed and trained on the collected dataset. Finally, the 1D CNN model is deployed to identify keyword/crane signalman speech commands in real time. The findings suggest that the framework is capable of identifying crane signalman speech commands in real time, of aiding communication in noisy construction sites, and of helping to reduce linguistic barriers in understanding speech commands.

The aim of the research described in Chapter 6 is to combine the respective decisions of the models proposed in Chapters 3, 4, and 5 to produce a single classification decision output. This is achieved using a redundancy-based ensemble model. Two types of ensemble models—weighted average and majority voting—are considered. Both constituent models of the redundancy-based ensemble model are evaluated on a test dataset, with the ensemble model showing high accuracy in classifying communication between crane signalman and crane operator in crane operation and outperforming the individual models it comprises.

7.2 Research Contributions

This research proposes a multimodal hand signal and speech communication framework for the construction industry. In this research, the communication between crane signalman and crane operator in crane operations is used as a case study. The industrial and academic contributions of this research are summarized as follows.

7.2.1 Industrial contributions

- 1) A deep-learning model is proposed that integrates two state-of-the-art models (i.e., YOLOv4+LSTM) to classify dynamic crane signalman hand signals in real time. The model is capable of classifying dynamic crane signalman hand signals on complex construction sites and under variable weather conditions, and can be used on construction sites as an added layer of communication to reduce communication errors and improve safety and productivity.
- 2) An SCG featuring multiple sensors is developed and deployed in conjunction with trained machine-learning models to classify the dynamic crane signalman hand signals in real time. The model is tested in real time using a mobile application developed for this research. The model is found to be capable of accurately classifying crane signalman hand signals in real

time. The SCG can be used as a supplementary layer of communication between crane operator and signalman in crane operations.

- 3) A 1D CNN model is developed to identify keyword/crane signalman speech commands from speech communication data. The model is shown to have the ability to identify keywords used by crane signalman to guide the crane operator during crane operations, and to do even under noisy construction site conditions and irrespective of the speaker's pitch or accent. As such, this model can assist in communication between crane operators and signalmen on construction sites.

7.2.2 Academic contributions

- 1) The scientometric analysis and critical review of applications of deep learning in construction serves as a valuable contribution for researchers by providing a thorough understanding of the current landscape of deep learning applications in construction. By identifying the existing applications, encountered challenges, and highlighting prevailing gaps in the field. This research study laid a foundation for the further development of deep-learning models to facilitate hand signal and speech communication on construction sites.
- 2) The proposed computer-vision-based deep learning model (i.e., YOLOv4+LSTM) enables real time crane signalman dynamic hand signals classification with high accuracy and speed in dynamic and complex construction sites and under variable weather conditions. The proposed model can assist the crane operator in decision making in crane operation. The findings also demonstrate the feasibility of using deep-learning models for classification on construction sites, even under complex and adverse conditions to improve construction site safety and enhance productivity.

- 3) The developed of sensor-based SCG enables the data collection of hand signal without interfering with the user's activities, meaning that it can be widely used for recording different hand signal datasets on construction sites. Furthermore, the addition of machine-learning models to the glove allows the SCG to classify the crane signalman hand signals in real time. The SCG can serve as an additional communication method to complement the current practice on construction sites. Moreover, this technology could be used in other applications ranging from other construction-related hand signal applications to hand signal classification of speech-impaired people
- 4) The concept of redundancy is adopted in the development of an ensemble model for hand signal classification. As an improvement upon existing models, the developed ensemble model combines the decisions of individual models to produce a single decision output. The model is capable of producing more reliable and accurate classification decisions compared to the individual models it comprises. The model also allows for the comparison of the impact of different redundant scenarios on the reliability of communication. This framework constitutes a substantive addition to the body of knowledge, yielding decisions that are more reliable than those generated by individual models, and that can be used in range of industrial and academic applications in which reliability is a high priority.

7.3 Limitations and Future Research

There are opportunities for future work to improve upon the presented framework. Potential avenues of future work in this domain are summarized below:

- 1) The present research is limited to the case of the classification of hand signal and speech communication between crane operator and signalman. The extension of this framework to the classification of other hand signals used for communication in construction, such as

in bulldozer operation, concrete truck operation, or wayfinding on construction sites, is highly recommended as an avenue of future study to better generalize the model.

- 2) Future research could focus on improving the quantity and quality of data by adopting more advanced methods of data collection. For instance, in the present research, although the video data is collected from multiple positions and angles as shown in Figure 3-3, the crane signalman helmet could be equipped with a 360° camera for more comprehensive video data collection and classification. The use of a 360° camera would allow for a wider field of view and for even subtle motions of the crane signalman's hand to be captured from any position/angle. This could enhance the accuracy and efficiency of data collection.
- 3) A promising potential avenue of future research is the exploration of novel methods to train deep learning-based classification models more effectively. The current process of training the classification models involves a considerable amount of manual effort to label training datasets, as well as significant computing resources and time. There is potential to improve upon this by redesigning the training process. Furthermore, efforts should be made to optimize the model in order to reduce the computation burden, enabling implementation on microcontrollers and other portable devices for on-site field classification. These improvements could enhance the efficiency and accessibility of deep-learning-based crane signalman hand signal classification models in practical settings.
- 4) Although the model presented herein has been tested for a maximum distance of 30 m, more effective means of transmitting hand signal labels over longer distances between crane operator and signalman should be investigated in future work. Further research is needed to explore alternative approaches to enable accurate communication over greater distances to improve safety and efficiency.

5) In this research, the ensemble model was only tested on a test dataset. Future research could explore the real-time application of the ensemble model to improve the performance of deep-learning classification models for the purpose of enhancing the reliability of communication between crane signalman and crane operator. Real-time implementation of the ensemble model would help to verify its efficacy and feasibility for practical use in construction sites. This could pave the way for more widespread adoption of such technologies in the construction industry, benefiting workers and construction enterprises alike.

REFERENCES

References for Chapter 1:

Ahmed, S. 2018. A review on using opportunities of augmented reality and virtual reality in construction project management. *Organization, Technology and Management in Construction: An International Journal*, 10(1), 1839–1852.

Ahuja, V. 2021. Transforming the media and entertainment industry: cases from the social media marketing world. *Journal of Cases on Information Technology (JCIT)*, 23(4), 1–17.

Ali, G. M., Mansoor, A., Liu, S., Olearczyk, J., Bouferguene, A., & Al-Hussein, M. (2021). Decision support for hydraulic crane stabilization using combined loading and crane mat strength analysis. *Automation in Construction*, 131, 103884.

All-West Crane and Rigging Ltd. 2019. An introduction to crane hand signals.

<https://www.allwestcrane.com/blog/2018/04/an-introduction-to-crane-hand-signals/>

(June 13, 2019).

Armírola Garcés, L. P., García Nieto, M. T., and Romero González, G. C. 2020. Digital communication in micro and small business: The case of the cultural sector in the Colombian Department of Bolivar. *Revista de Comunicación de la SEECI*, (52).

Azhar, S. 2011. Building information modeling (BIM): Trends, benefits, risks, and challenges for the AEC industry. *Leadership and Management in Engineering*, 11(3), 241–252.

Beavers, J.E., Asce, F., Moore, J.R., Rinehart, R., and Schriver, W.R. 2006. Crane-related fatalities in the construction industry. *Journal of Construction Engineering and Management.*, 132, 901–910.

- Bernold, L. E., Lorenc, S. J., and Luces, E. 1997. On-line assistance for crane operators. *Journal of Computing in Civil Engineering*, 11(4), 248–259.
- Bust, P. D., Gibb, A. G. F., and Pink, S. 2008. Managing construction health and safety : Migrant workers and communicating safety messages. *Safety Science*, 46, 585–602.
- CCOHS (Canadian Centre for Occupational Health and safety) 2019. OSH answers fact sheets retrieved from https://www.ccohs.ca/oshanswers/safety_haz/materials_handling/signals.html on (june12, 2019)
- Chan, M., Estève, D., Escriba, C., and Campo, E. 2008. A review of smart homes-Present state and future challenges. *Computer Methods and Programs in Biomedicine*, 91(1), 55–81.
- Chen, K., and Xue, F. 2022. The renaissance of augmented reality in construction: history, present status and future directions. *Smart and Sustainable Built Environment*, 11(3), 575–592.
- Chen, Q., García de Soto, B., and Adey, B. T. 2018. Construction automation: Research areas, industry concerns and suggestions for advancement. *Automation in Construction*, 94, 22–38.
- CPWR 2019. Communication in cranes. <https://www.cpwr.com/search-results/?search_txt=crane+communication> (May 2019)
- El-Sayegh, S., Romdhane, L., and Manjikian, S. 2020. A critical review of 3D printing in construction: Benefits, challenges, and risks. *Archives of Civil and Mechanical Engineering*, 20(2), 1–25.
- ENFORM 2013. D8 bulldozer contact with surveyor on ATV. <http://www.energysafe>

tycanada.com/files/safety-alerts/SA05-13-ATV-Bulldozer.pdf, 2013.

Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., and Li, C. 2018a. Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment. *Automation in Construction*, 93, 148–164.

Fang, W., Ding, L., Luo, H., and Love, P.E.D. 2018c. Falls from heights: A computer vision-based approach for safety harness detection. *Automation in Construction*, 91, 53–61. doi:10.1016/j.autcon.2018.02.018.

Fang, Y., and Cho, Y. K. 2016. A framework of lift virtual prototyping (LVP) approach for crane safety planning. *Proceedings of the 33rd International Symposium on Automation and Robotics in Construction*, 291–297.

Guo, F., Tang, J., and Wang, X. 2017. Gesture recognition of traffic police based on static and dynamic descriptor fusion. *Multimedia Tools and Applications*, 8915–8936.

Hagan, P. E., Montgomery, J. F., and O'Reilly, J. T. (Eds.). 2015. Accident prevention manual for business & industry: engineering & technology. *National Safety Council*.

Hu, Y. C., Chiu, Y. J., Hsu, C. S., and Chang, Y. Y. 2015. Identifying key factors for introducing GPS-based fleet management systems to the logistics industry. *Mathematical Problems in Engineering*, 2015.

Hwang, B. G., Ngo, J., and Teo, J. Z. K. 2022. Challenges and strategies for the adoption of smart technologies in the construction industry: The case of Singapore. *Journal of Management in Engineering*, 38(1), 05021014.

IATA 2020. <https://www.iata.org/en/training/courses/aircraft-marshalling-ramp/talp58/en/>

accessed on March 5, 2021

Jaffar, N., Abdul-Tharim, A. H., Mohd-Kamar, I. F., and Lop, N. S. 2011. A literature review of ergonomics risk factors in construction industry. *Procedia Engineering*, 20, 89–97.

Kines, P., Andersen, L. P. S., Spangenberg, S., Mikkelsen, K. L., Dyreborg, J., and Zohar, D. 2010. Improving construction site safety through leader-based verbal safety communication. *Journal of Safety Research.*, 41(5), 399–406.

King, R.A. 2012. Analysis of crane and lifting accidents in north america from 2004 to 2010. (*Doctoral Dissertation, Massachusetts Institute of Technology*)

Mansor, S. A. 2010. The construction sector at the onset of the 10th Malaysia plan. *Proceedings of the 7th Malaysia Construction Sector Review and Outlook Seminar*. Kuala Lumpur, Malaysia.

Matusitz, J., and Breen, G. M. 2007. Telemedicine: its effects on health communication. *Health Communication*, 21(1), 73–83.

Maxim Crane Works 2019. Hand signals you need to know during crane operations. <https://www.cranerental.com/hand-signals-need-know-crane-operations/> (June 7, 2019).

Mishra, M., Lourenço, P. B., and Ramana, G. V. 2022. Structural health monitoring of civil engineering structures by using the internet of things: A review. *Journal of Building Engineering*, 48, 103954.

NCCCO 2019. Why signal person is needed in crane operations? <<https://www.nccco.org/home>> (June 2019)

Neitzel, R. L., Seixas, N. S., and Ren, K. K. 2001. A review of crane safety in the construction industry. *Applied Occupational and Environmental Hygiene.*, 16(12), 1106–1117.

- Occupational Safety and Health Administration (OSHA). 2021. Cranes and derricks in construction. Accessed April 19, 2020. <https://open.alberta.ca/dataset/757fed78-8793-40bb-a920-6f000853172b/resource/9296e033-fd12-40dc-ac86-21e5873d4161/download/4403880-part-6-cranes-hoists-and-lifting-devices.pdf>.
- OCN 2021. The evolution of communication. (<
https://cdn2.hubspot.net/hubfs/214969/00_OCN_Content/OCN_NEU_eBook_Evolution-of-Communication_637-CTA.pdf> Accessed 12 January 2021)
- Patel, K. K., Patel, S. M., and Scholar, P. 2016. Internet of things-IOT: definition, characteristics, architecture, enabling technologies, application & future challenges. *International journal of engineering science and computing*, 6(5).
- Ratta, P., Kaur, A., Sharma, S., Shabaz, M., and Dhiman, G. 2021. Application of blockchain and internet of things in healthcare and medical sector: applications, challenges, and future perspectives. *Journal of Food Quality*, 2021, 1–20.
- Raviv, G., and Shapira, A. 2018. Systematic approach to crane-related near-miss analysis in the construction industry. *International Journal of Construction Management.*, 18(4), 310–320.
- Reakes, K. 2018. Traffic signal worker thrown from bucket in Stamford. <https://dailyvoic.com/connecticut/stamford/news/traffic-signal-worker-thrown-from-bucket-instamford/732557/>.
- Scolari, C. A. 2009. Mapping conversations about new media: the theoretical field of digital communication. *New media & society*, 11(6), 943–964.
- Soltanmohammadlou, N., Sadeghi, S., Hon, C. K. H., and Mokhtarpour-Khanghah, F. 2019.

- Real-time locating systems and safety in construction sites: A literature review. *Safety Science*, 117, 229–242.
- Sriram, N., and Nithiyanandham, M. 2013. A hand gesture recognition based communication system for silent speakers. *Proceedings of the International Conference on Human Computer Interactions, ICHCI 2013*, IEEE.
- Stevenson Crane 2019. Crane signals 101 – Can you hear me now?
<<https://stevensoncrane.com/crane-signals-101-can-hear-now/>> (June 17, 2019).
- Teizer, J., Wolf, M., Golovina, O., Perschewski, M., Propach, M., Neges, M., and König, M. 2017. Internet of Things (IoT) for integrating environmental and localization data in Building Information Modeling (BIM). *Proceedings of the International Symposium on Automation and Robotics in Construction* (Vol. 34). IAARC Publications.
- U.S. Bureau of Labor Statistics. 2017. Fatal occupational injuries involving cranes. Accessed September 13, 2021. <https://www.bls.gov/iif/oshwc/cfoi/cranes-2017.htm>
- Vivaldini, M., Pires, S. R., and Souza, F. B. D. 2012. Improving logistics services through the technology used in fleet management. *JISTEM-Journal of Information Systems and Technology Management*, 9, 541–562.
- Wang, P., Wu, P., Wang, J., Chi, H. L., and Wang, X. 2018. A critical review of the use of virtual reality in construction engineering education and training. *International journal of environmental research and public health*, 15(6), 1204.
- Zagan, I., Gaitan, V. G., Petrariu, A. I., and Brezulianu, A. 2017. Healthcare IoT m-GreenCARDIO remote cardiac monitoring system–concept, theory of operation and implementation. *Advances in Electrical and Computer Engineering*, 17(2), 23–30.

Zhao, Q. 2011. Cause analysis of U.S. crane-related accidents. (*Doctoral dissertation, University of Florida*).

References for Chapter 2:

Baker, H., Hallowell, M.R., and Tixier, A.J.P. 2020. Automatically learning construction injury precursors from text. *Automation in Construction*, 118, 103145.

doi:10.1016/j.autcon.2020.103145.

Bang, S., Baek, F., Park, S., Kim, W., and Kim, H. 2019. An image augmentation method for detecting construction resources using convolutional neural network and UAV images.

Proceedings of the 36th International Symposium on Automation and Robotics in Construction, ISARC 2019, 639–644. doi:10.22260/isarc2019/0085.

Bang, S., Baek, F., Park, S., Kim, W., and Kim, H. 2020. Image augmentation to improve construction resource detection using generative adversarial networks, cut-and-paste, and image transformation techniques. *Automation in Construction*, 115, 103198.

doi:10.1016/j.autcon.2020.103198.

Bangaru, S.S., Wang, C., Busam, S.A., and Aghazadeh, F. 2021. ANN-based automated scaffold builder activity recognition through wearable EMG and IMU sensors. *Automation in*

Construction, 126, 103653. doi:10.1016/j.autcon.2021.103653.

Bayer, A.E., Smart, J.C., and McLaughlin, G.W. 1990. Mapping intellectual structure of a scientific subfield through author cocitations. *Journal of the American Society for*

Information Science, 41(6), 444–452. doi:10.1002/(SICI)1097-4571(199009)41:6<444::AID-ASI12>3.0.CO;2-J.

- Brilakis, I., and Ariyachandra, M. 2020. Automatic detection of railway masts in air-borne LiDAR data. *Proceedings of the 8th Transport Research Arena*. <https://doi.org/10.17863/CAM.48898>
- Cha, Y. J., Choi, W., and Büyüköztürk, O. 2017. Deep learning-based crack damage detection using convolutional neural networks. *Computer-Aided Civil and Infrastructure Engineering*, 32(5), 361–378.
- Cha, Y. J., Choi, W., Suh, G., Mahmoudkhani, S., and Büyüköztürk, O. 2018. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. *Computer-Aided Civil and Infrastructure Engineering*, 33(9), 731–747
- Chen, C. 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 64, 1852–1863. doi:10.1002/asi.
- Chen, C. 2016. *CiteSpace : A practical guide for mapping scientific literature*. Hauppauge, NY, USA: Nove Science Publishers.
- Chen, C. 2017. Science Mapping: A systematic review of the literature. *Journal of Data and Information Science*, 2(2), 1–40. doi:10.1515/jdis-2017-0006.
- Chen, C., and Morris, S. 2003. Visualizing evolving networks: Minimum spanning trees versus Pathfinder networks. *Proceedings of the IEEE Symposium on Information Visualization, INFO VIS*, 2003, 67–74. doi:10.1109/INFVIS.2003.1249010.
- Chen, C., Zhu, Z., and Hammad, A. 2020. Automated excavators activity recognition and productivity analysis from construction site surveillance videos. *Automation in*

Construction, 110, 103045. doi:10.1016/j.autcon.2019.103045.

Chen, S., and Demachi, K. 2021. Towards on-site hazards identification of improper use of personal protective equipment using deep learning-based geometric relationships and hierarchical scene graph. *Automation in Construction*, 125, 103619. doi:10.1016/j.autcon.2021.103619.

Chi, N. W., Lin, K. Y., and Hsieh, S. H. 2013. On effective text classification for supporting job hazard analysis. *Proceedings of the Computing in Civil Engineering* (pp. 613–620).

Chian, E., Fang, W., Goh, Y.M., and Tian, J. 2021. Computer vision approaches for detecting missing barricades. *Automation in Construction*, 131, 103862. doi:10.1016/j.autcon.2021.103862.

Daniels, J., and Thistlethwaite, P. 2017. Measuring scholarly impact. *Being a Scholar in the Digital Era*. Springer International Publishing, Switzerland.

Darko, A., Chan, A.P.C., Adabre, M.A., Edwards, D.J., Hosseini, M.R., and Ameyaw, E.E. 2020. Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities. *Automation in Construction*, 112, 103081. doi:10.1016/j.autcon.2020.103081.

Davis, P., Aziz, F., Newaz, M.T., Sher, W., and Simon, L. 2021. The classification of construction waste material using a deep convolutional neural network. *Automation in Construction*, 122, 103481. doi:10.1016/j.autcon.2020.103481.

Deng, D., Hou, Y., Zhang, J., and Wang, M. 2020. Analysis of construction accident reports based on C-BiLSTM model. *Proceedings of the 13th International Conference on Intelligent*

- Ding, L., Fang, W., Luo, H., Love, P.E.D., Zhong, B., and Ouyang, X. 2018. A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. *Automation in Construction*, 86, 118–124. doi:10.1016/j.autcon.2017.11.002.
- Eck, N.J. Van, and Waltman, L. 2010. Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523–538. doi:10.1007/s11192-009-0146-3.
- Eck, N.J. Van, and Waltman, L. 2016. VOSviewer Manual 1.6.11. *Manual (version 1.6.4)*, 1–28.
- Fan, C., Sun, Y., Xiao, F., Ma, J., Lee, D., Wang, J., and Tseng, Y.C. 2020. Statistical investigations of transfer learning-based methodology for short-term building energy predictions. *Applied Energy*, 262, 114499. doi:10.1016/j.apenergy.2020.114499.
- Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., and Li, C. 2018a. Computer vision aided inspection on falling prevention measures for steeplejacks in an aerial environment. *Automation in Construction*, 93, 148–164. doi:10.1016/j.autcon.2018.05.022.
- Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T.M., and An, W. 2018b. Detecting non-hardhat-use by a deep learning method from far-field surveillance videos. *Automation in Construction*, 85, 1–9. doi:10.1016/j.autcon.2017.09.018.
- Fang, W., Ding, L., Luo, H., and Love, P.E.D. 2018c. Falls from heights: A computer vision-based approach for safety harness detection. *Automation in Construction*, 91, 53–61. doi:10.1016/j.autcon.2018.02.018.
- Fang, W., Ding, L., Zhong, B., Love, P.E.D., and Luo, H. 2018d. Automated detection of

- workers and heavy equipment on construction sites: A convolutional neural network approach. *Advanced Engineering Informatics*, 37, 139–149. doi:10.1016/j.aei.2018.05.003.
- Follini, C., Cheng, A. L., Latorre, G., and Amores, L. F. 2018. Design and development of a novel robotic gripper for automated scaffolding assembly. *Proceedings of the 2018 IEEE Third Ecuador Technical Chapters Meeting (ETCM)*.
- Frid-Adar, M., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. 2018. Synthetic data augmentation using GAN for improved liver lesion classification. *Proceedings of the IEEE 15th International Symposium on Biomedical Imaging*. IEEE. pp. 289–293.
- Girvan, M., and Newman, M.E.J. 2002. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(12), 7821–7826. doi:10.1073/pnas.122653799.
- Guo, Y., Xu, Y., and Li, S. 2020. Dense construction vehicle detection based on orientation-aware feature fusion convolutional neural network. *Automation in Construction*, 112, 103124. doi:10.1016/j.autcon.2020.103124.
- Hazrati Fard, S.M., and Hashemi, S. 2019. Sparse representation using deep learning to classify multi-class complex data. *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, 43(s1), 637–647. Springer International Publishing. doi:10.1007/s40998-018-0154-5.
- Hinton, G.E., and Salakhutdinov, R.R. 2006. Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504–507. doi:10.1126/science.1127647.
- Hong, Y., Hammad, A.W.A., Akbarnezhad, A., and Arashpour, M. 2020. A neural network

- approach to predicting the net costs associated with BIM adoption. *Automation in Construction*, 119, 103306. doi:10.1016/j.autcon.2020.103306.
- Hossain, M.M., Prybutok, V., and Evangelopoulos, N. 2011. Causal latent semantic analysis (clsa): an illustration. *International Business Research*, 4(2), 38–50. doi:10.5539/ibr.v4n2p38.
- Hou, X., Zeng, Y., and Xue, J. 2020. Detecting structural components of building engineering based on deep-learning method. *Journal of Construction Engineering and Management*, 146(2), 04019097. doi:10.1061/(asce)co.1943-7862.0001751.
- Kamal, R., Chemmanam, A.J., Jose, B.A., Mathews, S., and Varghese, E. 2020. Construction safety surveillance using machine learning. *Proceedings of the International Symposium on Networks, Computers and Communications, ISNCC 2020*, doi:10.1109/ISNCC49221.2020.9297198.
- Kang, D., and Cha, Y. J. 2018. Autonomous UAVs for structural health monitoring using deep learning and an ultrasonic beacon system with geo-tagging. *Computer-Aided Civil and Infrastructure Engineering*, 33(10), 885–902.
- Kang, D., Benipal, S. S., Gopal, D. L., and Cha, Y. J. 2020. Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning. *Automation in Construction*, 118, 103291.
- Khan, S., and Yairi, T. 2018. A review on the application of deep learning in system health management. *Mechanical Systems and Signal Processing*, 107, 241–265.
- Kim, J., Hwang, J., Chi, S., and Seo, J.O. 2020. Towards database-free vision-based monitoring

- on construction sites: A deep active learning approach. *Automation in Construction*, 120, 103376. doi:10.1016/j.autcon.2020.103376.
- Kingma, D.P., and Ba, J.L. 2015. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*.
- Kleinberg, J. 2002. Bursty and hierarchical structure in streams. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 91–101. doi:10.1145/775060.775061.
- LeCun, Y., Aurelio, M.', and Ranzato. 2013. Deep learning tutorial. *Tutorials in International Conference on Machine Learning (ICML '13)*.
- LeCun, Y., Bengio, Y., and Hinton, G. 2015. Deep learning. *Nature*, 521(7553), 436–444. doi:10.1038/nature14539.
- Li, R. Y. M., Li, H. C. Y., Tang, B., and Au, W. 2020. Fast AI classification for analyzing construction accidents claims. *Proceedings of the 2020 Artificial Intelligence and Complex Systems Conference*.
- Li, Y., Lu, Y., and Chen, J. 2021. A deep learning approach for real-time rebar counting on the construction site based on YOLOv3 detector. *Automation in Construction*, 124, 103602. doi:10.1016/j.autcon.2021.103602.
- Li, Y., Wei, H., Han, Z., Huang, J., and Wang, W. 2020. Deep learning-based safety helmet detection in engineering management based on convolutional neural networks. *Advances in Civil Engineering*, 2020. doi:10.1155/2020/9703560.
- Lin, Z.H., Chen, A.Y., and Hsieh, S.H. 2021. Temporal image analytics for abnormal

construction activity identification. *Automation in Construction*, 124,103572.

doi:10.1016/j.autcon.2021.103572.

Linton, C.F. 2016. A set of measures of centrality based on betweenness. *American Sociological Association Stable URL* : <http://www.jstor.org/stable/3033543>, 40(1): 35–41. (accessed December 13,2021).

Liu, H., Wang, G., Huang, T., He, P., Skitmore, M., and Luo, X. 2020. Manifesting construction activity scenes via image captioning. *Automation in Construction*, 119,103334.

doi:10.1016/j.autcon.2020.103334.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., and Berg, A.C. 2016. SSD: Single shot multibox detector. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9905, 21–37.

doi:10.1007/978-3-319-46448-0_2.

Lu, R., Brilakis, I., and Middleton, C. R. 2019. Detection of structural components in point clouds of existing RC bridges. *Computer-Aided Civil and Infrastructure Engineering*, 34(3), 191–212.

Luo, H., Wang, M., Wong, P.K.Y., and Cheng, J.C.P. 2020. Full body pose estimation of construction equipment using computer vision and deep learning techniques. *Automation in Construction*, 110, 103016. doi:10.1016/j.autcon.2019.103016.

Luo, X., Li, H., Cao, D., Dai, F., Seo, J., and Lee, S. 2018. Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. *Journal of Computing in Civil Engineering*, 32(3),

04018012. doi:10.1061/(asce)cp.1943-5487.0000756.

- Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., Al-Hussein, M., & Hassan, I. 2022. Keyword identification framework for speech communication on construction sites. *Modular and Offsite Construction (MOC) Summit Proceedings*, 106-113.
- Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., & Al-Hussein, M. 2022. The Effectiveness of data augmentation in construction site-related image classification. *Canadian Society of Civil Engineering Annual Conference* (pp. 247-257). Cham: Springer International Publishing.
- Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., & Al-Hussein, M., Soda (2022). Mobile crane signalman static hand signals classification framework using deep convolution neural network. *Proceedings of the 34th European Modeling & Simulation Symposium, EMSS, 2022*.
- Martinez, P., Al-Hussein, M., and Ahmad, R. 2019. A scientometric analysis and critical review of computer vision applications for construction. *Automation in Construction*, 107,102947. doi:10.1016/j.autcon.2019.102947.
- Mikołajczyk, A., and Grochowski, M. 2018. Data augmentation for improving deep learning in image classification problem. *Proceedings of the International Interdisciplinary PhD Workshop (IIPhDW)*. IEEE, 117–122.
- Mnemyneh, B.E., Abbas, M., and Khoury, H. 2019. Vision-Based framework for intelligent monitoring of hardhat wearing on construction sites. *Journal of Computing in Civil Engineering*, 33(2), 04018066. doi:10.1061/(ASCE)CP.1943-5487.0000813.
- Mok, K.Y., Shen, G.Q., and Yang, J. 2015. Stakeholder management studies in mega construction projects: A review and future directions. *International Journal of Project*

- Management*, 33(2), 446–457. doi:10.1016/j.ijproman.2014.08.007.
- Mongeon, P., and Paul-Hus, A. 2016. The journal coverage of Web of Science and Scopus: a comparative analysis. *Scientometrics*, 106(1), 213–228. doi:10.1007/s11192-015-1765-5.
- Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M., Seliya, N., Wald, R., and Muharemagic, E. 2015. Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2(1), 21 pages. doi:10.1186/s40537-014-0007-7.
- Nath, N.D., and Behzadan, A.H. 2020. Deep convolutional networks for construction object detection under different visual conditions. *Frontiers in Built Environment*, 6, 22 pages. doi:10.3389/fbuil.2020.00097.
- Neuhausen, M., Herbers, P., and König, M. 2020. Using synthetic data to improve and evaluate the tracking performance of construction workers on site. *Applied Sciences*, 10(14), 18 pages. doi:10.3390/app10144948.
- Olanrewaju, O. I., Sandanayake, M., and Babarinde, S. A. 2020. Voice assisted key-in building quantities estimation system. *Journal of Engineering, Project & Production Management*, 10(2).
- Olawumi, T.O., and Chan, D.W.M. 2018. A scientometric review of global research on sustainability and sustainable development. *Journal of Cleaner Production*, 183, 231–250. doi:10.1016/j.jclepro.2018.02.162.
- Oraee, M., Hosseini, M.R., Papadonikolaki, E., Palliyaguru, R., and Arashpour, M. 2017. Collaboration in BIM-based construction networks: A bibliometric-qualitative literature review. *International Journal of Project Management*, 35(7), 1288–1301.

doi:10.1016/j.ijproman.2017.07.001.

- Pan, G., Muresan, M., Yu, R., and Fu, L. 2021. Real-time winter road surface condition monitoring using an improved residual CNN. *Canadian Journal of Civil Engineering*, 48(9), 1215–1222. <https://doi.org/10.1139/cjce-2019-0367>.
- Patel, T., Guo, B. H., and Zou, Y. 2021. A scientometric review of construction progress monitoring studies. *Engineering, Construction and Architectural Management*. <https://doi.org/10.1108/ECAM-10-2020-0799>.
- Pereira, E., Ali, M., Wu, L., and AbouRizk, S. 2020. Distributed simulation–based analytics approach for enhancing safety management systems in industrial construction. *Journal of Construction Engineering and Management*, 146(1), 04019091. doi:10.1061/(asce)co.1943-7862.0001732.
- Perianes-Rodriguez, A., Waltman, L., and van Eck, N.J. 2016. Constructing bibliometric networks: A comparison between full and fractional counting. *Journal of Informetrics*, 10(4), 1178–1195. doi:10.1016/j.joi.2016.10.006.
- Redmon, J., and Farhadi, A. 2018. YOLOv3: An incremental improvement. *Computer Vision and Pattern Recognition*. *arXiv preprint arXiv:1804.02767*.
- Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. 2016. You Only Look Once: unified, real-time object detection. *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 779–788.
- Ren, S., He, K., Girshick, R., and Sun, J. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and*

Machine Intelligence, 39(6), 1137–1149. doi:10.1109/TPAMI.2016.2577031.

Ryu, J., Seo, J., Jebelli, H., and Lee, S. 2019. Automated action recognition using an accelerometer-embedded wristband-type activity tracker. *Journal of Construction Engineering and Management*, 145(1), 04018114. doi:10.1061/(asce)co.1943-7862.0001579.

Scarpiniti, M., Comminiello, D., Uncini, A., and Lee, Y.C. 2021. Deep recurrent neural networks for audio classification in construction sites. *Proceedings of the European Signal Processing Conference*, 2021, 810–814. doi:10.23919/Eusipco47968.2020.9287802.

Sergeev, A., and Del, B.M. 2018. Horovod: fast and easy distributed deep learning in TensorFlow. *arXiv preprint arXiv:1802.05799*.

Sherafat, B., Ahn, C.R., Akhavian, R., Behzadan, A.H., Golparvar-Fard, M., Kim, H., Lee, Y.C., Rashidi, A., and Azar, E.R. 2020. Automated methods for activity recognition of construction workers and equipment: state-of-the-art review. *Journal of Construction Engineering and Management*, 146(6). doi:10.1061/(ASCE)CO.1943-7862.0001843.

Shorten, C., and Khoshgoftaar, T.M. 2019. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1), 1–48.

Siluo, Y., and Qingli, Y. 2017. Are scientometrics, informetrics, and bibliometrics different? *Proceedings of the 16th International Conference on Scientometrics and Informetrics*, 1507–1518.

Sim, J., Kasahara, J.Y.L., Chikushi, S., Nagatani, K., Chiba, T., Chayama, K., Yamashita, A., and Asama, H. 2021. Effects of video filters for learning an action recognition model for

- construction machinery from simulated training data. *Proceedings of the 2021 IEEE/SICE International Symposium on System Integration, SII 2021*: 12–16.
doi:10.1109/IEEECONF49454.2021.9382735.
- Slaton, T., Hernandez, C., and Akhavian, R. 2020. Construction activity recognition with convolutional recurrent networks. *Automation in Construction*, 113,103138.
doi:10.1016/j.autcon.2020.103138.
- Small, H. 1973. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24(4), 265–269. doi:10.1002/asi.4630240406.
- Son, H., and Kim, C. 2021. Integrated worker detection and tracking for the safe operation of construction machinery. *Automation in Construction*, 126, 103670.
doi:10.1016/j.autcon.2021.103670.
- Su, H.N., and Lee, P.C. 2010. Mapping knowledge structure by keyword co-occurrence: A first look at journal papers in Technology Foresight. *Scientometrics*, 85(1), 65–79.
doi:10.1007/s11192-010-0259-8.
- Torok, M. M., Golparvar-Fard, M., and Kochersberger, K. B. 2014. Image-based automated 3D crack detection for post-disaster building assessment. *Journal of Computing in Civil Engineering*, 28(5), A4014004.
- Wang, Z., and Hong, T. 2020. Reinforcement learning for building controls: The opportunities and challenges. *Applied Energy*, 269. doi:10.1016/j.apenergy.2020.115036.
- Wang, Z., Zhang, Y., Mosalam, K.M., Gao, Y., and Huang, S.L. 2022. Deep semantic

- segmentation for visual understanding on construction sites. *Computer-Aided Civil and Infrastructure Engineering*, 37(2), 145–162. doi:10.1111/mice.12701.
- Xiao, B., and Kang, S.C. 2019. Deep learning detection for real-time construction machine checking. *Proceedings of the 36th International Symposium on Automation and Robotics in Construction, ISARC 2019*, 1136–1141. doi:10.22260/isarc2019/0151.
- Xiao, B., and Kang, S.-C. 2021a. Vision-Based method integrating deep learning detection for tracking multiple construction machines. *Journal of Computing in Civil Engineering*, 35(2), 04020071. doi:10.1061/(asce)cp.1943-5487.0000957.
- Xiao, B., and Kang, S.-C. 2021b. Development of an image data set of construction machines for deep learning object detection. *Journal of Computing in Civil Engineering*, 35(2), 05020005. doi:10.1061/(asce)cp.1943-5487.0000945.
- Xuehui, A., Li, Z., Zuguang, L., Chengzhi, W., Pengfei, L., and Zhiwei, L. 2021. Dataset and benchmark for detecting moving objects in construction sites. *Automation in Construction*, 122,103482. doi:10.1016/j.autcon.2020.103482.
- Yang, Z., He, B., Liu, Y., Wang, D., and Zhu, G. 2021. Classification of rock fragments produced by tunnel boring machine using convolutional neural networks. *Automation in Construction*, 125, 103612. doi:10.1016/j.autcon.2021.103612.
- Yu, Y., Li, H., Umer, W., Dong, C., Yang, X., Skitmore, M., and Wong, A.Y.L. 2019a. Automatic biomechanical workload estimation for construction workers by computer vision and smart insoles. *Journal of Computing in Civil Engineering*, 33(3), 04019010. doi:10.1061/(ASCE)CP.1943-5487.0000827.

- Yu, Y., Li, H., Yang, X., Kong, L., Luo, X., and Wong, A.Y.L. 2019b. An automatic and non-invasive physical fatigue assessment method for construction workers. *Automation in Construction*, 103, 1–12. doi:10.1016/j.autcon.2019.02.020.
- Yu, Y., Yang, X., Li, H., Luo, X., Guo, H., and Fang, Q. 2019c. Joint-level vision-based ergonomic assessment tool for construction workers. *Journal of Construction Engineering and Management*, 145(5), 04019025. doi:10.1061/(asce)co.1943-7862.0001647.
- Zhang, M., Zhu, M., and Zhao, X. 2020. Recognition of high-risk scenarios in building construction based on image semantics. *Journal of Computing in Civil Engineering*, 34(4), 04020019. doi:10.1061/(ASCE)CP.1943-5487.0000900.
- Zhang, T., Lee, Y. C., Zhu, Y., and Hernando, J. 2018. A conversation analysis framework using speech recognition and naïve bayes classification for construction process monitoring. *Proceedings of the Construction Research Congress 2018* (pp. 572–580)
- Zheng, Z., Zhang, Z., and Pan, W. 2020. Virtual prototyping- and transfer learning-enabled module detection for modular integrated construction. *Automation in Construction*, 120,103387. doi:10.1016/j.autcon.2020.103387.
- Zhong, B., Xing, X., Love, P., Wang, X., and Luo, H. 2019. Convolutional neural network: Deep learning-based classification of building quality problems. *Advanced Engineering Informatics*, 40, 46–57.
- Zhou, X., Gong, Q., Liu, Y., and Yin, L. 2021a. Automatic segmentation of TBM muck images via a deep-learning approach to estimate the size and shape of rock chips. *Automation in Construction*, 126, 103685. doi:10.1016/j.autcon.2021.103685.

Zhou, Y., Guo, H., Ma, L., Zhang, Z., and Skitmore, M. 2021b. Image-based onsite object recognition for automatic crane lifting tasks. *Automation in Construction*, 123,103527. doi:10.1016/j.autcon.2020.103527.

References for Chapter 3:

- Agarap, A. F. M. 2018. Deep learning using rectified linear units (ReLU). *arXiv preprint arXiv:1803.08375*.
- Bae, J.H., Le, N.T. and Kim, J.T. 2017. Smartphone image receiver architecture for optical camera communication. *Wireless Personal Communications*, 93(4), 1043–1066.
- Bochkovskiy, A., Wang, C., and Liao, H. M. 2020. YOLOv4: optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*.
- Boncellet, C. 2009. Image noise models: The essential guide to image processing, 1st edition, 143–167. Academic Press, Elsevier.
- Bull, D. (2014). Digital picture formats and representation. *Communicating Pictures*, 99–132.
- Chen, Y., Chi, H., Kangm, S., and Hsieh, S. 2011. A smart crane operations assistance system using augmented reality technology. *Proceedings of the 28th International Symposium on Automation and Robotics in Construction*, 643–649.
- Ghiasi, G., Lin, T., and V.Le, Q. 2018. DropBlock: A regularization method for convolutional networks. *Advances in Neural Information Processing Systems*, 1–11.
- Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. 2016. LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10), 2222–2232.
- Hernandez-Garcia, A., and Konig, P. 2018. Further advantages of data augmentation on convolutional neural networks. *Proceedings of the International Conference on Artificial*

- Neural Networks.*, 95–103.
- Hou, X., Zhang, Y., and Hou, J. 2020. Application of YOLO V2 in construction vehicle detection. *Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery.*, 1249–1256.
- Hu, J., Geo, X., Wu, H., and Gao, S. 2019. Detection of workers without the helmets in videos based on YOLO V3. *Proceedings of the 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics.*, 1553–1560.
- Huang, J., Zhou, W., Zhang, Q., Li, H., and Li, W. 2018. Video-based sign language recognition without temporal segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, (Vol. 32, No. 1).
- Huang, Z., Wang, J., Fu, X., Yu, T., Guo, Y., and Wang, R. 2020. DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection. *Information Sciences.*, 522, 241–258.
- Kim, J. A., Sung, J. Y., and Park, S. H. 2020. Comparison of Faster-RCNN, YOLO, and SSD for real-time vehicle type recognition. *Proceedings of the 2020 IEEE International Conference on Consumer Electronics-Asia.*
- Kukkala, V. K., Tunnell, J., Pasricha, S., and Bradley, T. 2018. Advanced driver-assistance systems: A path toward autonomous vehicles. *IEEE Consumer Electronics Magazine*, 7(5), 18–25.
- Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. 2018. Path aggregation network for instance segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 8759–8768.
- Mikołajczyk, A., and Grochowski, M. 2018. Data augmentation for improving deep learning in

- image classification problem. *Proceedings of the International Interdisciplinary PhD Workshop*, 117–122.
- Misra, D. 2019. MISH: A self regularized non-monotonic activation function. *arXiv preprint arXiv,1908.08681*.
- Misra, S., and Wu, Y. 2019. Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. *Machine Learning for Subsurface Characterization*, 289–314.
- Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S., and Kautz, J. 2016. Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4207–4215.
- Nepal, U., and Eslamiat, H. 2022. Comparing YOLOv3, YOLOv4 and YOLOv5 for autonomous landing spot detection in faulty UAVs. *Sensors*, 22(2), 464.
- Occupational Safety and Health Administration (OSHA). 2021. Cranes and derricks in construction. Accessed April 19, 2020. <https://open.alberta.ca/dataset/757fed78-8793-40bb-a920-6f000853172b/resource/9296e033-fd12-40dc-ac86-21e5873d4161/download/4403880-part-6-cranes-hoists-and-lifting-devices.pdf>.
- Okan, K., Ahmet, G., Neslihan, K., and Gerhard, R. 2019. Real-time hand gesture detection and classification using convolutional neural networks. *Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition*, 1–8.
- Oyedotun, O. K., and Khashman, A. 2017. Deep learning in vision-based static hand gesture recognition. *Neural Computing and Applications.*, 28(12), 3941–3951.
- Pickering, C. a., Burnham, K. J., and Richardson, M. J. 2007. A research study of hand gesture

- recognition technologies and applications for human vehicle interaction. *Proceedings of the 3rd Institution of Engineering and Technology Conference on Automotive Electronics*, 1–15.
- Potter, M.C., Wyble, B., Haggmann, C.E. and McCourt, E.S. 2014. Detecting meaning in rsvp at 13 ms per picture. *Attention, Perception, & Psychophysics*, 76(2), 270–279.
- Qasim, A. B., and Pettirsch, A. 2020. Recurrent neural networks for video object detection. *arXiv preprint arXiv:2010.15740*.
- Rahman, E. U., Zhang, Y., Ahmad, S., Ahmad, H. I., and Jobaer, S. 2021. Autonomous vision-based primary distribution systems porcelain insulators inspection using UAVs. *Sensors*, 21(3), 974.
- Ramachandran, P., Zoph, B., and Le, Q. V. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941*.
- Su, H., Qi, W., Yang, C., Sandoval, J., Ferrigno, G., and Momi, E. De. 2020. Deep neural network approach in robot tool dynamics identification for bilateral teleoperation. *IEEE Robotics and Automation Letters*, 5(2), 2943–2949.
- Xu, J., Li, Z., Du, B., Zhang, M., and Liu, J. 2020. Reluplex made more practical: Leaky ReLU. *Proceedings of the 2020 IEEE Symposium on Computers and Communications*, 1–7.
- Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Al-Hussein, M., and Kurach, L. 2020. A deep learning-based framework for an automated defect detection system for sewer pipes. *Automation in Construction*, 109, 102967.
- Zavichi, A., and Behzadan, A. H. 2011. A real time decision support system for enhanced crane operations in construction and manufacturing. *Computing in Civil Engineering*, 586–593.
- Zekavat, P. R., and Bernold, L. 2014. Embedded wireless communication platform addresses

crane safety and efficiency. *Proceedings of the Construction Research Congress 2014.*, 309–318.

Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., and Ren, D. 2019. Distance-IoU Loss : faster and better learning for bounding box regression. *Proceedings of the AAAI Conference on Artificial Intelligence.* 12993–13000.

References for Chapter 4:

Agarap, A. F. 2018. Deep learning using rectified linear units (ReLU). *arXiv preprint arXiv:1803.08375.*

Ahn, C. R., Lee, S., Sun, C., Jebelli, H., Yang, K., and Choi, B. 2019. Wearable sensing technology applications in construction safety and health. *Journal of Construction Engineering and Management*, 145(11), 03119007.

Akhavian, R., and Behzadan, A. H. 2016. Smartphone-based construction workers' activity recognition and classification. *Automation in Construction*, 71(Part 2), 198–209.

Al Mamun, A., Polash, M. S. J. K., and Alamgir, F. M. 2017. Flex sensor based hand glove for deaf and mute people. *International Journal of Computer Networks and Communications Security*, 5(2), 38.

Awolusi, I., Marks, E., and Hallowell, M. 2018. Wearable technology for personalized construction safety monitoring and trending: Review of applicable devices. *Automation in Construction*, 85, 96–106.

Begg, R. K., Palaniswami, M., and Owen, B. 2005. Support vector machines for automated gait classification. *IEEE transactions on Biomedical Engineering*, 52(5), 828–838.

- Chaudhury, S. B., Sengupta, M., and Mukherjee, K. 2014. Vibration monitoring of rotating machines using MEMS accelerometer. *International Journal of Scientific Engineering and Research*, 2(9), 5–11.
- Chuang, W. C., Hwang, W. J., Tai, T. M., Huang, D. R., and Jhang, Y. J. 2019. Continuous finger gesture recognition based on flex sensors. *Sensors (Switzerland)*, 19(18), 1–21.
- Dong, W., Yang, L., and Fortino, G. 2020. Stretchable human machine interface based on smart glove embedded with PDMS-CB strain sensors. *IEEE Sensors Journal*, IEEE, 20(14), 8073–8081.
- Fan, L., Wang, Z., and Wang, H. 2013. Human activity recognition model based on decision tree. *Proceedings of the 2013 International Conference on Advanced Cloud and Big Data* (pp. 64–68). IEEE.
- Ghiasi, M. M., Zendejboudi, S., and Mohsenipour, A. A. 2020. Decision tree-based diagnosis of coronary artery disease: CART model. *Computer methods and programs in biomedicine*, 192, 105400.
- Ghobakhloo, M. (2018). The future of manufacturing industry: a strategic roadmap toward Industry 4.0. *Journal of Manufacturing Technology Management*, 29(6), 910–936.
- Guerrero-Ibáñez, J., Zeadally, S., and Contreras-Castillo, J. (2018). Sensor technologies for intelligent transportation systems. *Sensors*, 18(4), 1212.
- Gupta, H. P., Chudgar, H. S., Mukherjee, S., Dutta, T., and Sharma, K. 2016. A continuous hand gestures recognition technique for human-machine interaction using accelerometer and gyroscope sensors. *IEEE Sensors Journal*, IEEE, 16(16), 6425–6432.

- Gurbuz, S. Z., Gurbuz, A. C., Malaia, E. A., Griffin, D. J., Crawford, C. S., Rahman, M. M., ... and Mdrafi, R. 2020. American Sign Language recognition using RF sensing. *IEEE Sensors Journal*, 21(3), 3763–3775.
- Hemingway, E. G., and O'Reilly, O. M. 2018. Perspectives on Euler angle singularities, gimbal lock, and the orthogonality of applied forces and applied moments. *Multibody System Dynamics*, 44(1), 31–56.
- Hwang, S., Seo, J., Jebelli, H., and Lee, S. 2016. Feasibility analysis of heart rate monitoring of construction workers using a photoplethysmography (PPG) sensor embedded in a wristband-type activity tracker. *Automation in Construction*, 71, 372–381.
- Jani, A. B., Kotak, N. A., and Roy, A. K. 2018. Sensor based hand gesture recognition system for english alphabets used in sign language of deaf-mute people. *Proceedings of IEEE Sensors*, 17–20.
- Jebelli, H., Choi, B., and Lee, S. 2019. Application of wearable biosensors to construction sites. I: assessing workers' stress. *Journal of Construction Engineering and Management*, 145(12), 04019079.
- Jo, B. W., Lee, Y. S., Khan, R. M. A., Kim, J. H., and Kim, D. K. 2019. Robust Construction Safety System (RCSS) for collision accidents prevention on construction sites. *Sensors*, 19(4), 932.
- Kaghyan, S., and Sarukhanyan, H. 2012. Activity recognition using k-nearest neighbor algorithm on smartphone with tri-axial accelerometer. *International Journal of Informatics Models and Analysis (IJIMA)*, ITHEA International Scientific Society, Bulgaria, 1, 146–156.

- Kanan, R., Elhassan, O., and Bensalem, R. 2018. An IoT-based autonomous system for workers' safety in construction sites with real-time alarming, monitoring, and positioning strategies. *Automation in Construction*, 88, 73–86.
- Kim, Jayoung, Alan S. Campbell, Berta Esteban-Fernández de Ávila, and Joseph Wang. 2019. Wearable biosensors for healthcare monitoring. *Nature biotechnology* 37(4): 389–406.
- Kim, M., Cho, J., Lee, S., and Jung, Y. 2019. Imu sensor-based hand gesture recognition for human-machine interfaces. *Sensors (Switzerland)*, 19(18), 1–13.
- Lai, X., Yang, T., Wang, Z., and Chen, P. 2019. IoT implementation of Kalman Filter to improve accuracy of air quality monitoring and prediction. *Applied Sciences (Switzerland)*, 9(9).
- Mandong, A., and Munir, U. 2018. Smartphone based activity recognition using k-nearest neighbor algorithm. *Proceedings of the International Conference on Engineering Technologies, Konya, Turkey* (pp. 26–28).
- Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., & Al-Hussein, M. 2023. A deep-learning classification framework for reducing communication errors in dynamic hand signaling for crane operation. *Journal of Construction Engineering and Management*, 149(2), 04022167.
- Marks, E. D., and Teizer, J. 2013. Method for testing proximity detection and alert technology for safe construction equipment operation. *Construction Management and Economics*, 31(6), 636–646.
- McGinnis, R. S., Cain, S. M., Davidson, S. P., Vitali, R. V., McLean, S. G., and Perkins, N. C. 2014. Validation of complementary filter based IMU data fusion for tracking torso angle and rifle orientation. *Proceedings of the ASME International Mechanical Engineering*

- Congress and Exposition* (Vol. 46469, p. V003T03A052). American Society of Mechanical Engineers.
- Minaie, A., Sanati-Mehrizy, A., Sanati-Mehrizy, P., and Sanati-Mehrizy, R. (2013, June). Application of wireless sensor networks in health care system. *Proceedings of the 2013 ASEE Annual Conference and Exposition* (pp. 23.200.1- 23.200.12).
- Minh, V. T., Moezzi, R., and Katushin, N. 2019. Haptic smart glove for augmented and virtual reality. *Sensor Letters*, 17(5), 358–364.
- Nath, N. D., Akhavian, R., and Behzadan, A. H. 2017. Ergonomic analysis of construction worker's body postures using wearable mobile sensors. *Applied Ergonomics*, 62, 107–117.
- Nonnarit, O., and Barreto, A. 2016. Gyroscope drift correction algorithm for inertial measurement unit used in hand motion tracking. *Proceedings of the 15th IEEE SENSORS Conference*.
- O'Flynn, B., Saez-Torres, J., Tedesco, S., Downes, B., Connolly, J., Condell, J., and Curran, K. 2015. Novel smart glove technology as a biomechanical monitoring tool. *Sensors and Transducers*, 193(10), 23–32.
- Oztemel, E., and Gursev, S. (2020). Literature review of Industry 4.0 and related technologies. *Journal of Intelligent Manufacturing*, 31, 127–182.
- Park, J., Cho, Y. K., and Timalina, S. K. 2016. Direction aware Bluetooth low energy based proximity detection system for construction work zone safety. *Proceedings of the 33rd International Symposium on Automation and Robotics in Construction* (pp. 76–82).
- Pathak, V., Mongia, S., and Chitranshi, G. 2016. A framework for hand gesture recognition

- based on fusion of flex, contact and accelerometer sensor. *Proceedings of the 3rd International Conference on Image Information Processing, ICIIP 2015*, 312–319.
- Qamar, S., Abdelrehman, A. M., Elshafie, H. E., and Mohiuddin, K. (2018). Sensor-based IoT industrial healthcare systems. *Int. J. Sci. Eng. Sci.*, 11(2), 29–34.
- Qureshi, K. N., and Abdullah, A. H. (2013). A survey on intelligent transportation systems. *Middle-East Journal of Scientific Research*, 15(5), 629–642.
- Raheja, Jagdish Lal, Radhey Shyam, Umesh Kumar, and P. Bhanu Prasad. 2010. Real-time robotic hand control using hand gestures. *Proceedings of the 2nd International Conference on Machine Learning and Computing*, pp. 12–16. IEEE.
- Roelofs, C., Sprague-Martinez, L., Brunette, M., and Azaroff, L. 2011. A qualitative investigation of Hispanic construction worker perspectives on factors impacting worksite safety and risk. *Environmental Health*, 10, 1–9.
- Rosero-Montalvo, P. D., Godoy-Trujillo, P., Flores-Bosmediano, E., Carrascal-Garcia, J., Otero-Potosi, S., Benitez-Pereira, H., and Peluffo-Ordenez, D. H. 2018. Sign language recognition based on intelligent glove using machine learning techniques. *Proceedings of the 3rd IEEE Ecuador Technical Chapters Meeting*, 5–9.
- Saggio, G. 2012. Mechanical model of flex sensors used to sense finger movements. *Sensors and Actuators, A: Physical*, 185, 53–58.
- Sathiyarayanan, Mithileysh, Syed Azharuddin, Santhosh Kumar, and Gibran Khan. 2014. Gesture controlled robot for military purpose. *International Journal for Technological Research in Engineering* 1(11): 1300–1303.

- Shibuya, N., Nukala, B. T., Rodriguez, A. I., Tsay, J., Nguyen, T. Q., Zupancic, S., and Lie, D. Y. 2015. A real-time fall detection system using a wearable gait analysis sensor and a Support Vector Machine (SVM) classifier. *Proceedings of the 8th International Conference on Mobile Computing and Ubiquitous Networking (ICMU)* (pp. 66–67). IEEE.
- Sriram, N., and Nithiyandham, M. 2013. A hand gesture recognition based communication system for silent speakers. *Proceedings of the 2013 International Conference on Human Computer Interactions*, IEEE, (Adx1 335), 2–6.
- Taneja, S., and Sukhija, M. 2020. Technical paper gift of voice to mute. *Hand Gestures Converted to Text and Voice*, 8(4), 64–69.
- Tewelde, G. S. (2012, May). Sensor and network technology for intelligent transportation systems. *Proceedings of the 2012 IEEE International Conference on Electro/Information Technology* (pp. 1–7). IEEE.
- Tu, D., Bein, D., and Gofman, M. 2020. Designing a unity game using the haptic feedback gloves, VMG 30 Plus. *Proceedings of the 17th International Conference on Information Technology–New Generations (ITNG 2020)* (pp. 393–400). Springer International Publishing.
- Valenti, R. G., Dryanovski, I., and Xiao, J. 2015. Keeping a good attitude: A quaternion-based orientation filter for IMUs and MARGs. *Sensors*, 15(8), 19302–19330.
- Zhang, F. 1997. Quaternions and matrices of quaternions. *Linear Algebra and Its Applications*, 251, 21–57.
- Zhang, X., Chen, X., Li, Y., Lantz, V., Wang, K., and Yang, J. 2011. A framework for hand

gesture recognition based on accelerometer and EMG sensors. *IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans*, IEEE, 41(6), 1064–1076.

References for Chapter 5:

- Biadsy, F. 2011. Automatic dialect and accent recognition and its application to speech recognition. *Doctoral dissertation, Columbia University*, New York City, NY, USA.
- Bocklet, T., Maier, A., Bauer, J. G., Burkhardt, F., and Noth, E. 2008. Age and gender recognition for telephone applications based on GMM supervectors and support vector machines. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1605–1608.
- Bust, P. D., Gibb, A. G., and Pink, S. 2008. Managing construction health and safety: Migrant workers and communicating safety messages. *Safety Science*, 46(4), 585–602.
- Carbonell, M. L. B., Carpio, J. M. R., Medina, J. C. C., Perote, J. P., Tamayo, T. J. J., and Mappatao, G. P. 2020. Development of a stand-alone and scalable weather monitoring system using two-way VHF radios. *Indonesian Journal of Electrical Engineering and Computer Science*, 20(1), 475–484.
- Chelba, C., Hazen, T. J., and Saraclar, M. 2008. Retrieval and browsing of spoken content. *IEEE Signal Processing Magazine*, 25(3), 39–49.
- Cheng, C. F., Rashidi, A., Davenport, M. A., and Anderson, D. V. 2017. Activity analysis of construction equipment using audio signals and support vector machines. *Automation in Construction*, 81, 240–253.

- Davis, S., and Mermelstein, P. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), 357–366.
- De Pinto, M. G., Polignano, M., Lops, P., and Semeraro, G. 2020. Emotions understanding model from spoken language using deep neural networks and mel-frequency cepstral coefficients. *Proceedings of the 2020 IEEE Conference on Evolving and Adaptive Intelligent Systems*.
- Fragopanagos, N., and Taylor, J. G. 2005. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4), 389–405.
- Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... and Kingsbury, B. 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6), 82–97.
- Huang, X., Acero, A., Hon, H. W., and Reddy, R. (2001). Spoken language processing: a guide to theory, algorithm, and system development. *Prentice Hall PTR, Upper Saddle River, NJ, USA*.
- Kannan, A., Datta, A., Sainath, T. N., Weinstein, E., Ramabhadran, B., Wu, Y., and Lee, S. 2019. Large-scale multilingual speech recognition with a streaming end-to-end model. *arXiv preprint arXiv:1909.05330*.
- Kwon, N., Park, M., Lee, H. S., Ahn, J., and Shin, M. 2016. Construction noise management using active noise control techniques. *Journal of Construction Engineering and Management*, 142(7), 04016014.

- Latif, S., Cuayáhuitl, H., Pervez, F., Shamshad, F., Ali, H. S., and Cambria, E. 2021. A survey on deep reinforcement learning for audio-based applications. *arXiv preprint arXiv:2101.00240*.
- Lei, C., Hu, B., Wang, D., Zhang, S., and Chen, Z. 2019. A preliminary study on data augmentation of deep learning for image classification. *Proceedings of the 11th Asia-Pacific Symposium on Internetware*.
- Logan, B. 2000. Mel frequency cepstral coefficients for music modeling. *Proceedings of the International Symposium on Music Information Retrieval*.
- López-Espejo, I., Tan, Z. H., Hansen, J., and Jensen, J. 2021. Deep spoken keyword spotting: An overview. *IEEE Access*.
- Mahmood, A., and Utku, K. Ö. S. E. 2021. Speech recognition based on convolutional neural networks and MFCC algorithm. *Advances in Artificial Intelligence Research, 1(1)*, 6–12.
- Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., & Al-Hussein, M. 2022. Scientometric analysis and critical review on the application of deep learning in the construction industry. *Canadian Journal of Civil Engineering, 50(4)*, 253-269.
- Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., and Al-Hussein, M. 2020. Conceptual framework for safety improvement in mobile cranes. *Proceedings of the Construction Research Congress 2020: Computer Applications* (pp. 964–971). Reston, VA: American Society of Civil Engineers.
- Meftah, A. H., Alotaibi, Y. A., and Selouani, S. A. 2018. Evaluation of an Arabic speech corpus of emotions: A perceptual and statistical analysis. *IEEE Access, 6*, 72845–72861.

- Michaely, A. H., Zhang, X., Simko, G., Parada, C., and Aleksic, P. 2017. Keyword spotting for Google assistant using contextual speech recognition. *Proceedings of the 2017 IEEE Automatic Speech Recognition and Understanding Workshop*, 272–278.
- Mixkit. 2022. Construction sound effects. Retrieved March 3, 2022, from <https://mixkit.co/free-sound-effects/construction/>
- Momeni, L., Afouras, T., Stafylakis, T., Albanie, S., and Zisserman, A. 2020. Seeing wake words: Audio-visual keyword spotting. *arXiv preprint arXiv:2009.01225*.
- Otter, D. W., Medina, J. R., and Kalita, J. K. 2020. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2), 604–624.
- Petrushin, V. A. 2000. Emotion recognition in speech signal: experimental study, development, and application. *Proceedings of the Sixth International Conference on Spoken Language Processing*.
- Schultz, T., and Waibel, A. 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35(1–2), 31–51.
- Shin, S., and Issa, R. R. 2021. BIMASR: Framework for voice-based BIM information retrieval. *Journal of Construction Engineering and Management*, 147(10), 04021124.
- Stevenson Crane. 2019. *Crane signals 101 – Can you hear me now?* Retrieved June 17, 2019, from <https://stevensoncrane.com/crane-signals-101-can-hear-now/>

- Strehl, A. L., Li, L., Wiewiora, E., Langford, J., and Littman, M. L. 2006. PAC model-free reinforcement learning. *Proceedings of the 23rd International Conference on Machine Learning*, 881–888.
- Vogt, T., and André, E. 2006. Improving automatic emotion recognition from speech via gender differentiation. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*.
- Werchniak, A., Chicote, R. B., Mishchenko, Y., Droppo, J., Condal, J., Liu, P., and Shah, A. 2021. Exploring the application of synthetic audio in training keyword spotters. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 7993–7996.
- Zapsplat 2022. Construction site sound effects. Retrieved March 2, 2022, from <https://www.zapsplat.com/sound-effect-category/construction-site/>
- Zhang, T., Lee, Y. C., Scarpiniti, M., and Uncini, A. 2018b. A supervised machine learning-based sound identification for construction activity monitoring and performance evaluation. *Proceedings of the Construction Research Congress*, 358–366.
- Zhang, T., Lee, Y. C., Zhu, Y., and Hernando, J. 2018a. A conversation analysis framework using speech recognition and naïve bayes classification for construction process monitoring. *Proceedings of the Construction Research Congress*, 572–580.
- Zhou, G., Hansen, J. H., and Kaiser, J. F. 2001. Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing*, 9(3), 201–216

References for Chapter 6:

- Alrehili, A., and Albalawi, K. 2019. Sentiment analysis of customer reviews using ensemble

- method. *Proceedings of the 2019 International Conference on Computer and Information Sciences* (pp. 1–6). IEEE.
- Atallah, R., and Al-Mousa, A. 2019. Heart disease detection using machine learning majority voting ensemble method. *Proceedings of the 2nd International Conference on New Trends in Computing Sciences* (pp. 1–6). IEEE.
- Barkhordari, M. S., Armaghani, D. J., Mohammed, A. S., and Ulrikh, D. V. 2022. Data-driven compressive strength prediction of fly ash concrete using ensemble learner algorithms. *Buildings*, 12(2), 132.
- Cao, Y., Ashuri, B., and Baek, M. 2018. Prediction of unit price bids of resurfacing highway projects through ensemble machine learning. *Journal of Computing in Civil Engineering*, 32(5), 04018043.
- Choi, J. Y., and Lee, B. 2019. Ensemble of deep convolutional neural networks with Gabor face representations for face recognition. *IEEE Transactions on Image Processing*, 29, 3270–3281.
- George, M. R., Nalluri, M. R., and Anand, K. B. 2022. Application of ensemble machine learning for construction safety risk assessment. *Journal of the Institution of Engineers (India): Series A*, 103(4), 989–1003.
- Hu, J., Dong, F., Qiu, Y., Xi, L., Majdi, A., and Ali, E. 2022. Ensembles of neural network with stochastic optimization algorithms in predicting concrete tensile strength. *STEEL AND COMPOSITE STRUCTURES*, 45(2), 205–218.
- Irvine, N., Nugent, C., Zhang, S., Wang, H., and Ng, W. W. 2019. Neural network ensembles for

- sensor-based human activity recognition within smart environments. *Sensors*, 20(1), 216.
- Kumari, S., Kumar, D., and Mittal, M. 2021. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering*, 2, 40–46.
- Mehanović, D., Mašetić, Z., and Kečo, D. 2020. Prediction of heart diseases using majority voting ensemble method. *Proceedings of the International Conference on Medical and Biological Engineering, 16–18 May 2019, Banja Luka, Bosnia and Herzegovina* (pp. 491–498). Springer International Publishing.
- Milošević, I., Kovačević, M., and Petronijević, P. 2021. Estimating residual value of heavy construction equipment using ensemble learning. *Journal of Construction Engineering and Management*, 147(7), 04021073.
- Mukherjee, D., Mondal, R., Singh, P. K., Sarkar, R., and Bhattacharjee, D. 2020. EnsemConvNet: a deep learning approach for human activity recognition using smartphone sensors for healthcare applications. *Multimedia Tools and Applications*, 79, 31663–31690.
- Neloy, M. A. I., Nahar, N., Hossain, M. S., and Andersson, K. 2022. A weighted average ensemble technique to predict heart disease. *Proceedings of the Third International Conference on Trends in Computational and Cognitive Engineering: TCCE 2021* (pp. 17–29). Singapore: Springer Nature Singapore.
- Raza, K. 2019. Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule. *U-Healthcare Monitoring Systems* (pp. 179–196). Academic Press.
- Renda, A., Barsacchi, M., Bechini, A., and Marcelloni, F. 2019. Comparing ensemble strategies

for deep learning: An application to facial expression recognition. *Expert Systems with Applications*, 136, 1–11.

Wei, C., Chen, L. L., Song, Z. Z., Lou, X. G., and Li, D. D. 2020. EEG-based emotion recognition using simple recurrent units network and ensemble learning. *Biomedical Signal Processing and Control*, 58, 101756.

Zhang, F., Fleyeh, H., Wang, X., and Lu, M. 2019. Construction site accident analysis using text mining and natural language processing techniques. *Automation in Construction*, 99, 238–248.

APPENDIX 1:

MOBILE CRANE SIGNALMAN STATIC HAND SIGNAL CLASSIFICATION FRAMEWORK USING DEEP CONVOLUTIONAL NEURAL NETWORK⁵

Introduction

The communication between the crane operator and the signalman relies on hand signals and two-way radio communication systems. The crane industry has been using universal hand signals for decades to give direction to the crane operator for safe crane operation (Everett and Slocum 1993). Hand signals are the fastest and most reliable way to communicate a message when the crane operator has a direct line of sight with the signalman and the operator then operates the crane based on the direction given by the signalman. However, when the signalman is far away from the crane operator, the crane operator will not be able to clearly distinguish the signalman's hand signals; while at other times the crane operator's line of sight will be obstructed due to construction site congestion. These limitations make this method ineffective and unsafe (Everett and Slocum 1993; Shapira et al. 2008). To overcome the limitations, the crane industry has used two signalmen at the same time such that the second signalman copies the main signalman's signals and transmits these to the operator; however, this process is less productive and cannot be 100% reliable due to the potential for the signalman to misinterpret the signals or miscommunicate them to the crane operator, which may lead to a disastrous accident (Fang and Cho 2016). Another means of communication in the crane industry is the use of a two-way radio communication system.

⁵ A version of this work has been published as follows: Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., Al-Hussein, M., and Soda (2022) "Mobile crane signalman static hand signals classification framework using deep convolution neural network." *Proceedings of the 34th European Modeling and Simulation Symposium*, Rome, Italy, Sep. 19–21, 2022, ISSN 2724-0029.

Typically, this two-way radio communication system is used on the site when the crane operator's direct line of sight/vision is blocked by an obstacle or when the signalman is far away from the operator and he cannot see the signalman clearly; however, a two-way radio communication system needs to be maintained on a dedicated channel that is available at all times (Zekavat et al.

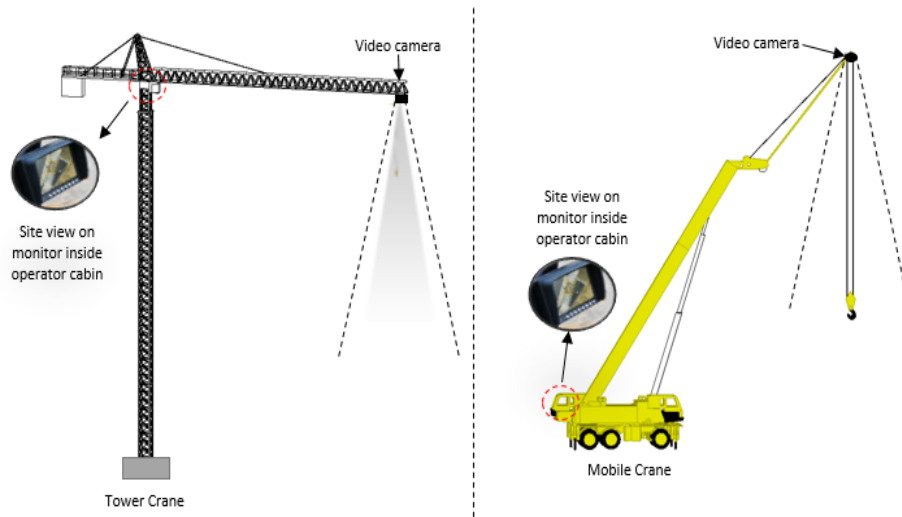


Figure 1. Camera-based vision system for tower and mobile cranes

2014). This system can be viewed as an alternative to the hand signalling system under certain circumstances and, potentially, as an extra layer of safety. A two-way radio communication system cannot be 100% reliable when the construction site is noisy, such as when there is drilling on site and the operator cannot hear the signalman clearly (Zavichi and Behzadan 2011; Mansoor et al. 2020). Second, while communicating using two-way radios, one hand needs to be used to push the talk button to send the voice message while the other hand must be used for signalling, which may cause a miscommunication error and can be dangerous (Zekavat et al. 2014). Another drawback of the two-way radio system is when a problem occurs with the dedicated channel being used by the operator and signalman because the whole crane operation is halted until the problem is resolved, which causes delays, productivity losses, and safety risks (Zavichi and Behzadan 2011). To overcome the aforementioned limitations, information technology can provide another layer of safety by making the communication between signalman and crane operator more efficient and

more accurate by using deep convolutional neural network (DCNN). The objectives of the framework are to develop a DCNN model, to train the model with an image dataset of crane signalman hand signals, to achieve a high level of accuracy in classifying images of crane signalman hand signals in training, validation, and test datasets, and to further validate the model for real-time crane signalman hand signal classification. The proposed framework is capable of classifying the crane signalman hand signals in real time. The benefit of using this framework is that it assists communication between the crane operator and the signalman.

The present study is organized as follows: Section 2 Introduces related work in context of improving the communication between crane operator and signalman. Section 3 describes the methodology and architecture of the proposed DCNN model. Sections 4 and 5 present the model optimization techniques and data preparations. Section 6 evaluates the Implementation and model performance and describe the model real-time classification results of the proposed model. Section 7 contain the conclusion, limitations of the present study and related future work.

Related work

To improve safety and communication in the context of crane operation, researchers have done a significant amount of work. Camera-based vision systems have been developed that enable the operator to monitor the construction site while operating the crane. Researchers have also developed sensors to detect hazards and dangerous situations on the construction site. These technologies are quite beneficial in terms of improving safety and communication on construction sites, but due to the limitations of these technologies, they are not yet widely employed in the crane industry. These technologies are discussed below in detail.

Camera-based systems

In crane industry, in particular, can benefit from improved communication and safety systems; therefore, researchers have developed camera-based systems as shown in Figure 1. (Shapira et al. 2008) developed a video monitoring system that not only improves safety in the construction site but also increases productivity by 11 to 26%. The noted results were based on 2400 time complete delivery of payload from picking to dropping the load to their allotted locations. The system consists of a video camera mounted to the top of the tower crane to show a live video feed of the site that focuses on the signalman (Shapira et al. 2008). A monitor in the crane cabin allows the operator to see the live video of the signalman as well as the site. According to (Rosenfeld 1995), video cameras affixed to the cranes can be useful both in terms of efficiency and safety improvements as the live site vision enables the operator to make judgments on site without hesitation.

The drawback of this technology is that the camera shows only 2-dimensional images without any perception of depth. The operator has difficulty accurately determining the distance of the load to the ground, which can lead to serious accidents on the site (Shapira et al. 2008). CRANIUM is another camera-based technology, developed by Everett (Everett and Slocum 1993), that has been found to improve both safety and cost-effectiveness because it eliminates the need for the second signalman (who is responsible for copying the hand signals of the first signalman at the lifting point and communicating these to the crane operator) (Everett and Slocum 1993). In this system, a camera is fixed to the top of the boom and a monitor is placed in the operator's cabin. From the monitor, the crane operator can see the loading area as well as the signalman (Everett and Slocum 1993). Stoneridge-Orlaco and HoistCam are camera and monitor manufacturers providing camera-based solutions for live video feed for tower, telescopic and crawler cranes to improve the communication, safety, and efficiency of crane operation on construction sites. The disadvantage

of these camera-based systems is that sometimes while the crane is in operation, a part of the crane, such as a hook or sling, moves in front of the camera which obstructs the operator's view and the operator is unable to see the signalman or the target area. This drawback can slow down the operation and increase the safety risks for the workers on the site (Everett and Slocum 1993).

Sensor-based systems

Another means to accomplish improved safety and communication in the crane industry is the application of sensor-based systems. (Li et al. 2013) introduced RFIDs to track construction workers on site with the help of a GPS tracking system. This system is used to limit the movement of the workers because only authorized workers were allowed in the danger zone during any crane lift operation. By automatically detecting unauthorized workers, the system sends an alert to warn authorized management personnel. This system was approved by the management and staff on the site (Li et al. 2013). (Fang and Cho 2016) developed a sensor-based framework that was capable of providing real-time safety assistance for mobile crane lift operations on a construction site. The sensors were affixed to a 70-ton telescopic crane and were responsible for detecting nearby objects and warning the crane operator of any dangers. This framework was implemented on the site and the responses from the working crew were recorded. Most of the crew was satisfied with the developed framework and responded that the framework can be helpful during blind lifts by responding early before the crane comes into contact with any object (Fang and Cho 2016). To improve the two-way radio communication system between crane operator and signalman, (Zekavat et al. 2014) developed a camera-based vision system with a wireless microphone to monitor the blind lifts. In this system, a laptop was placed above the eye level of the operator in the crane operator cabin, where the operator have to look up the laptop to make decision while moving the load. This system improved the communication and visibility of the site during crane

operation (Zekavat et al. 2014). On the other hand, the limitations of sensor-based technologies are the accuracy of the system, and the measurement error and setup error of the system makes these systems less trustworthy because a small margin of error can cause serious accidents on the construction site (Everett and Slocum 1993). With respect to implementing RFIDs for the purpose of worker tracking, the workers responded that they have privacy concerns, in particular, that their productivity will be judged based on their movements; this system was noted to have accuracy issues as well (Li et al. 2013).

Methodology and CNN architecture

In recent years, rapid improvements in computation power have led to the development of deep-learning algorithms for image classification, notably, convolutional neural networks (CNNs). A CNN extracts the features from the input image pixels and classifies the images with high accuracy

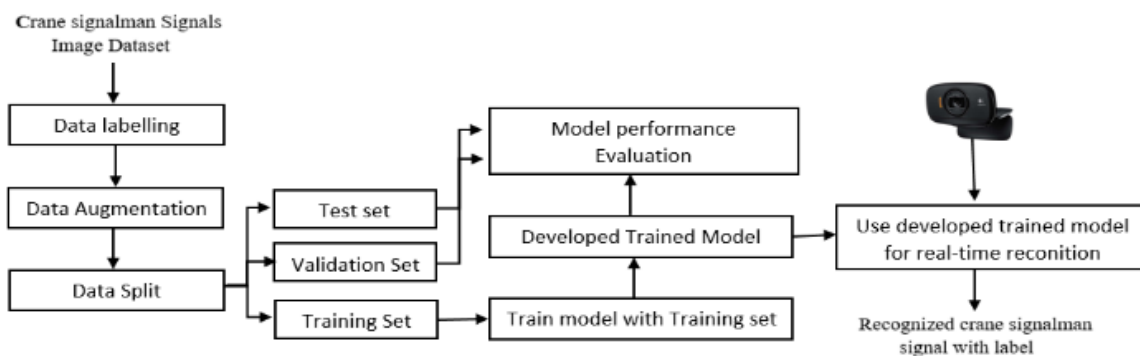


Figure 2. Overview of the deep learning based framework

and generalization capabilities. In the civil engineering domain, CNNs have been widely used for the classification and detection of equipment on construction sites, and construction workers wearing hardhat (Arabi et al. 2019; Gu et al. 2019; Hu et al. 2019; Wang et al. 2020; Wu et al. 2019), of defects in sewer pipes (Yin et al. 2020), and of defects in concrete surfaces (Cha et al. 2017), for example.

In the present study, a DCNN approach is developed to classify crane signalman static hand signals in real time. The developed DCNN is trained, validated, and tested using a dataset of 8,133 augmented and non-augmented images that are collected by individuals from the research team. An overview of the developed deep learning-based framework is given in Figure 2. The crane signalman hand signals dataset containing all 18 hand signals necessary for crane operation serves as the original data source for the signalman hand signals. The crane signalman hand signals are based on the occupational health and safety (OHS) crane hand signals (Occupational Safety and Health Administration, 2009), as shown in Figure 3.

In order to improve communication, the crane signalman hand signals must be classified in real time. To accomplish this, a deep learning approach is developed and is consists of two main parts. First data is collected and preprocessed and the second is the development of the DCNN model that is discussed in detail in Sections 3.1 and 3.2

Data collection and preprocessing

Since there is no accessible dataset available for crane signalman hand signals, the dataset needed for this work is collected by taking 6,507 images of individuals from the research team doing all 18 crane signalman hand signals in all possible angles (0° to 360°), and positions (right, left, front, and back sides of signalman). The camera is set up at multiple distances (5 m to 20 m) to record the image dataset in different environments (sunny and cloudy). While creating the dataset, data balancing is taken into account, which means the portion of images in the dataset showing each one of the 18 hand signals should be 5.5% of the whole dataset. However, in the collected dataset, 7.51% of images depict the emergency stop hand signal, which is the maximum portion, and the minimum portion of images, 3.64%, depict the travel hand signal as shown in Table 1. Therefore, the amount of data provided to train the model is sufficient for the model to learn the features from

all hand signals in the dataset. The number of samples collected for each hand signal can be seen in Table 1.

All images have a resolution of 1280×720 pixels. The DCNN can be trained using images of any resolution; however, a higher resolution means more features will be extracted from images, which increases the computational complexity and the processing time. To reduce the processing time and computational load, all images are scaled down to a 280×280 pixel resolution for further processing.

Table 1. Image samples collected for each hand signal

Standard Hand signals	Number of sample images	Percentage (%)
Hoist	486	7.47
Lower	411	6.32
Use main hoist	231	3.55
Use whipline	333	5.12
Boom up	477	7.33
Boom down	453	6.96
Move slowly	417	6.41
Swing	243	3.73
Boom down and raise the load	393	6.04
Boom up and lower the load	459	7.05
Stop	423	6.50
Emergency stop	489	7.51

Travel	237	3.64
Dog everything	285	4.38
Travel both tracks	246	3.78
Travel one track	288	4.43
Telescope out	333	5.12
Telescope in	303	4.66



















 <p>HOIST With upper arm extended to the side, forearm and index finger pointing straight up, hand and finger make small circles.</p>	 <p>LOWER With arm and index finger pointing down, hand and finger make small circles.</p>	 <p>USE MAIN HOIST A hand taps on top of the head. Then regular signal is given to indicate desired action.</p>	 <p>USE WHIPLINE (Auxiliary Hoist) With arm bent at elbow and forearm vertical, elbow is tapped with other hand. Then regular signal is used to indicate desired action.</p>	 <p>TRAVEL With all fingers pointing up, arm is extended horizontally out and back to make a pushing motion in the direction of travel.</p>	 <p>DOG EVERYTHING Hands held together at waist level.</p>
 <p>BOOM UP With arm extended horizontally to the side, thumb points up with other fingers closed.</p>	 <p>BOOM DOWN With arm extended horizontally to the side, thumb points down with other fingers closed.</p>	 <p>MOVE SLOWLY A hand is placed in front of the hand that is giving the action signal. (Hoist slowly shown in example.)</p>	 <p>SWING With arm extended horizontally, index finger points in direction that boom is to swing.</p>	 <p>TELESCOPE OUT (TELESCOPING BOOMS) With hands to the front at waist level, thumbs point outward with other fingers closed.</p>	 <p>TELESCOPE IN (TELESCOPING BOOMS) With hands to the front at waist level, thumbs point at each other with other fingers closed.</p>
 <p>BOOM DOWN AND RAISE THE LOAD With arm extended horizontally to the side and thumb pointing down, fingers open and close while load movement is desired.</p>	 <p>BOOM UP AND LOWER THE LOAD With arm extended horizontally to the side and thumb pointing up, fingers open and close while load movement is desired.</p>	 <p>STOP With arm extended horizontally to the side, palm down, arm is swung back and forth.</p>	 <p>EMERGENCY STOP With both arms extended horizontally to the side, palms down, arms are swung back and forth.</p>	 <p>TRAVEL (BOTH TRACKS) Rotate fists around each other in front of body; direction of rotation towards body indicates travel forward; rotation away from body indicates travel backward. (For crawler cranes only)</p>	 <p>TRAVEL (ONE TRACK) Indicate track to be locked by raising fist on that side. Rotate other fist in front of body in direction that other track is to travel. (For crawler cranes only)</p>

Figure 3. Crane signalman hand signals (Occupational Safety and Health Administration., 2009).

Data augmentation

Data augmentation is mandatory to better generalize the dataset, and it will help to increase the number of images in the dataset, which decreases the chances of model overfitting. There are several techniques that can be used to increase the size of a dataset and generalize the dataset, such as the cropping, rotating or flipping of the images. In the present work, the intensity transformation technique is used, which is an adjustment of the contrast and brightness of the images to generalize the image dataset for different scenarios such as low and high lighting conditions. The data augmentation increased the size of dataset by 25%.

Convolutional neural network-based deep-learning model

The architecture of the developed deep convolutional neural network

CNNs are constructed by artificial neurons (i.e., mathematical functions that receive one or many inputs and sum them to produce an output) with weight, biases, and activation functions, which are responsible for transforming input images into a single output value. According to LeCun and Bengio (1995), CNNs use spatial decomposition of input images in multiple stages. Spatial decomposition is achieved through convolution and pooling layers. A DCNN is composed of four main features: convolution layer that is responsible to transform images through various filters to

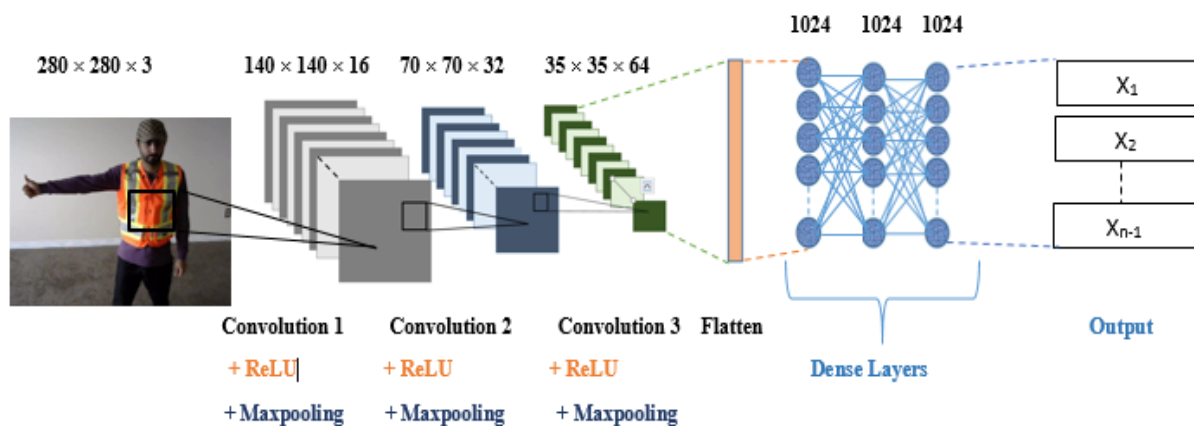


Figure 4. Architecture of the developed DCNN

extract the features from the input, activation that adds non-linearity to the output neurons, pooling or sub-sampling responsible for reducing dimensions of the feature map, and classification responsible for transferring output into a classification score.

The DCNN model developed for the proposed approach consists of convolution layers, pooling layers, dense layers, and an output layer. The input image in the DCNN has three colour channels, red, green and blue (RGB), and it can be viewed as three 2-dimensional matrices arranged over each other having a pixel value in the range of 0 to 255 and is passed through convolution, pooling and dense layers to achieve an output vector. The architecture of the DCNN is shown in Figure 4.

The weights in the network are adjusted and optimized through the process of backpropagation. The process of backpropagation is achieved through epochs/ iterations such that the CNN can correctly classify all images in the dataset. The architecture of the developed DCNN model consists of layers that are discussed in the following sections.

Convolutional layers

Convolution layers are considered the building blocks of CNNs. The convolution operation is responsible for extracting features from the input image. The input image is convolved into a number of kernels. Kernels are the matrices used to store the weights of convolution operations.

Developed DCNN model architecture has three convolution layers. In the first convolution layer, the images are of size $280 \times 280 \times 3$ (width, height, colour channels) and each input image is convolved into 16 different kernels. Each kernel has a size of $3 \times 3 \times 3$ (width, height, colour channels). After the first convolution, each output has a size of 280×280 and 16 channels, so the resulting output will become $280 \times 280 \times 16$. The resulting output is passed through the activation function and then subsampled to size $140 \times 140 \times 16$ using max-pooling layers.

Second convolution layer takes the output of the max-pooling layer of size $140 \times 140 \times 16$, which is further convolved with 32 different kernels, each kernel with a size of $3 \times 3 \times 16$. This will result in 32 output channels of size 140×140 . The resulting output is passed through the activation function and then subsampled to size $70 \times 70 \times 32$ using max-pooling layers.

Third and final convolution layer takes the output of max-pooling after the second convolution layer of size $70 \times 70 \times 32$ and is then convolved with 64 different kernels, each kernel with the size of $3 \times 3 \times 32$. This will result in 64 output channels of size 70×70 . After adding biases to 64 channels, the resulting output is passed through the activation function and then subsampled to size $35 \times 35 \times 64$ using max-pooling layers.

The variation in the number of kernels is used in the architecture to obtain the highest accuracy in the model. Finally, a DCNN architecture with 16, 32, and 64 kernels in the first, second, and third convolution layers, respectively, was found to achieve the highest accuracy on the validation dataset.

The developed network used a stride value of 2 that helps the kernel to move two matrix pixels at a time. This parameter affects the dimensions of output and reduces the chances of model overfitting. In the model, padding is used to assist the kernel to move uniformly over the matrix and its edges to obtain all the desired information in the image.

Activation function

An activation function is used in the CNN to add non-linearity to the output neurons. Adding an activation function is essential, otherwise, the DCNN would compute linear combinations of linear functions and the model would not be able to learn complicated or non-linear functions (Nair and Hinton., 2010).

Activation function used in the developed DCNN is rectified linear unit (ReLU) and softmax. The ReLU activation function is an identity line where $y = x$ for all positive lines and 0 for all negative values (Nair and Hinton, 2010). The mathematical equation for ReLU is given in Eq. (1).

$$f(x) = \max(0, x) \quad (1)$$

For the activation of the output layer, the softmax activation function is used. The softmax function is generally used for multiclass classification. It squashes the output of each unit to be between 0 and 1 and returns the probabilities of the input being in a particular class. Mathematically, it can be written as shown in Eq. (2).

$$P(y_i | x_i) = \frac{e^{f_{yi}}}{\sum_j e^{f_{yj}}} \quad (2)$$

where y_i = correct label of image x_i , and f_{yi} = predicted score.

Max-pooling layer

Convolution layer along with the activation function is followed by the pooling or subsampling layer. The purpose of adding the pooling layer is to reduce the dimensions of the feature map and retain the important information (LeCun and Bengio, 1995). This layer also reduces computation and helps in reducing the overfitting of the model. In max-pooling, a kernel of size $n \times n$ is moved across the matrix and for each position; the maximum value is taken.

In the developed DCNN, each convolution layer is followed by a max-pooling layer to reduce the dimension by a factor of 2. In the first convolution layer, max-pooling reduces the output channel from 280×280 to 140×140 . In the second convolution layer, it further reduces the output channel from 140×140 to 70×70 , and at the final convolution layer, the output channel is reduced from 70×70 to 35×35 .

Dense/fully connected layers

Output of max-pooling layers is the input to the dense layer. In the dense layer, all input and output are connected to all the neurons in each layer, while neurons within a single layer share no connection. In a CNN, dense layers are used to create the final non-linear combination of features and to predict the output layer. In the present study, three dense layers with 1,024 neurons in each layer are used. These layers are selected based on the classification performance of the validation set. A different number of dense layers is used and the accuracy of each is recorded. While recording the accuracy, it was noted that three dense layers with 1,024 neurons in each layer increase the average classification accuracy by 3%.

Output layer

The final layer of the CNN architecture is the output layer, which is responsible for transferring the output into a classification score. The softmax function is used for the activation of the output layer. The softmax function takes as input the predicted class labels and outputs a probability score. The softmax function is expressed as Eq. (2). The probability scores of the outputs must have a sum of 1. The probability score indicates class prediction. The largest probability score for any hand signal is considered to belong to the correct class.

Model optimization

When the model achieves higher accuracy in the training dataset than the validation dataset, the model is considered to be overfitted. To prevent overfitting in the model, data augmentation is performed for the dataset (Chatfield et al. 2014), which includes brightness change, contrast change, image resizing, and image rescaling of images. Another method used to mitigate the overfitting in the model is dropout regularization. The dropout technique was proposed by

(Srivastava et al. 2014). This technique activates the neurons with a certain probability and is implemented during the training stage. As the network is trained, neurons get randomly deactivated with respect to their weights and it will lead to better generalization of predictive capabilities (Srivastava et al. 2014). In the developed DCNN, a drop rate of 0.2 is chosen. A drop rate of 0.2 is selected based on multiple trails on the model, which gives higher accuracy during the validation and testing of the dataset.

Preparation of training, validation and test datasets

The dataset contains 8,133 augmented and non-augmented images with 18 different crane signalman hand signals. The images were collected with individuals from the research team. The dataset was randomly split into three different datasets: a training set, a validation set, and a test set. The training set includes 70% of the images, while the validation and test sets each include 15%. The reason for choosing a larger sample size for the training set is to get more features from the training dataset, which leads to better accuracy in the validation and test datasets.

Model performance evaluation

The DCNN model is developed using Python with Keras and TensorFlow API, which is an open-source computer software library for dataflow. The training set is converged into 160 epochs (iterations) which were completed in 8 hours on windows operating system with Intel corei7 processor and GeForce RTX 3050ti graphics card. The number of epochs, 160, was chosen to achieve maximum accuracy in the validation dataset. As shown in Figure 5, the accuracies of the training and validation dataset increase when the number of epochs in the model is increased while the value of loss decreases with the increase of the number of epochs.

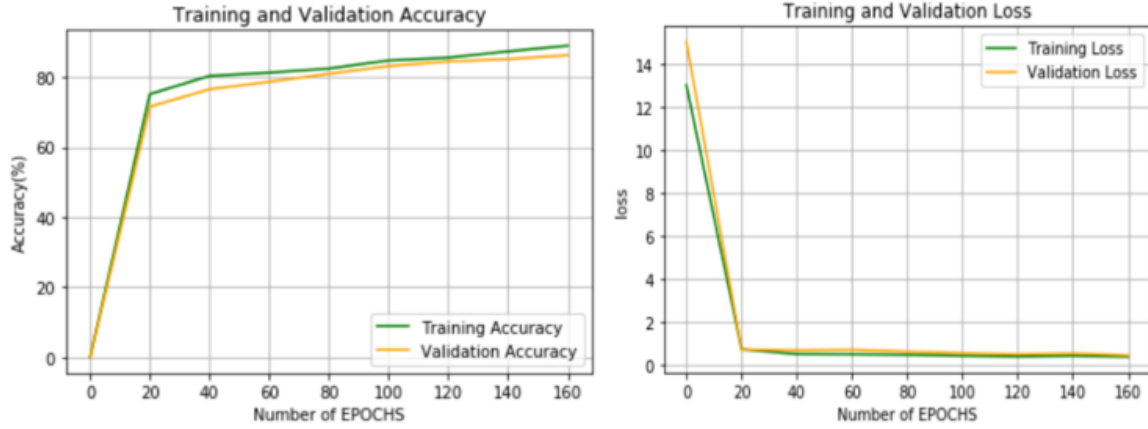


Figure 5. Training and validation accuracy and loss

The average training and validation accuracies achieved by the developed model were 89.1% and 84.6%, respectively, and the loss calculated for training and validation was 0.38 and 0.44, respectively. The model calculates the accuracy using Eq. (3) and the cross entropy loss using Eq. (4).

$$Accuracy = \frac{\text{Correctly classified hand signal}}{\text{Total number of hand signal shown}} \times 100 \quad (3)$$

$$\text{cross entropy loss} = \frac{-1}{N} \times \sum_{x=1}^N \sum_{y=1}^M Z_{xy} \times \log(p_{xy}) \quad (4)$$

where N is the number of samples and M is the number of classes; Z_{xy} represents whether sample x belongs to class y or not and p_{xy} shows the probability of sample x belonging to class y . The loss has no upper limit and it exists in the range $[0, \infty]$. The value of loss nearer to 0 indicates higher accuracy and vice versa.

Furthermore, a confusion matrix is used as a metric to evaluate the performance of the developed model on the test dataset. On the basis of the results obtained in the confusion matrix in the test dataset, the precision, recall, and F1 score are calculated using Equations 5, 6, and 7.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

where TP is the number of true positives (the detected hand signal belongs to the class that is shown); FN is the number of false negatives which represent that the detected hand signal is not of the same class as actually shown, and FP is the number of false positives that represent that the detected hand signal is of different class and shown hand signal is of a different class. The model achieved an average precision for all crane signalman hand signals of 81.5%, and the average value recorded for the recall was 81.8%.

The other metric used to measure the performance of the model is the F1 score, which captures the properties of both precision and recall and combines them into a single unit. The reason for using the F1 score is that the model cannot be judged only on the basis of good results in either precision or recall. The F1 score for the developed model was recorded as 81.7%. Typically, the F1 score falls in a range from 0% to 100%, where 0% is poor performance and 100% is the best performance of the model.

Real-time classification

The model is further validated by deploying the developed model such that it is used to classify the crane signalman hand signals in real time using a live stream. The developed DCNN model is validated in real time using live stream by showing all the crane signalman hand signals a total of 802 times. The model was capable of correctly recognizing the hand signals 706 times with an average accuracy of 87.9%. The deep-learning model was capable of correctly recognizing 42

hand signals out of 45 with an accuracy of 93.3%, 46 hand signals out of 48 with an accuracy of 95.8%, and 51 hand signals out of 52 with an accuracy of 98.1% for the labels hoist, stop, and emergency stop, respectively. The correctly and incorrectly recognized hand signals along with the accuracy can be seen in Table 2. The accuracy is measured using Eq. (3).

The proposed developed framework with deep-learning model can be implemented in cranes using a camera and a screen/ head-up display. The camera is used to record the live stream of the signalman hand signals, which are transmitted to the screen placed inside the crane operator cabin. The developed DCNN model shows the results (i.e., the classified labels of detected signalman hand signals) on the screen. The operator inside the cabin can see the signalman hand signals and their classified labels on the screen. This developed framework assists the crane operator to take the decision about the movement of load more confidently and efficiently. In this manner, the framework can be used as an improvement to the current state of practice for communication between crane operator and signalman and serve as another layer of communication and safety in crane industry.

Table 2. Accuracy of real-time crane signalman hand signal classification

Standard Hand signals	Number of times crane signalman hand signal shown in camera	Number of times crane signalman hand signal is correctly classified	Number of times crane signalman hand signal is incorrectly classified	Accuracy (%)
Hoist	45	42	3	93.3
Lower	42	39	5	88.1
Use main hoist	49	39	10	79.6
Use whipline	46	40	6	87.0
Boom up	48	42	6	87.5
Boom down	45	39	6	86.7
Move slowly	48	44	4	91.4
Swing	47	38	9	80.9
Boom down and raise the load	40	35	5	87.5
Boom up and lower the load	39	34	5	87.2
Stop	48	46	2	95.8
Emergency stop	52	51	1	98.1
Travel	41	33	8	80.5

Dog everything	49	44	5	89.8
Travel both tracks	36	31	5	86.1
Travel one track	32	28	4	87.5
Telescope out	48	42	6	87.5
Telescope in	47	41	6	87.2

Conclusions, limitations and future work

This developed framework presented in this paper used a DCNN model to classify the crane signalman hand signals in real time, which will help to improve communication between crane operator and signalman. The DCNN model is trained using images of 18 hand signals that are used by the signalman to communicate instructions to the crane operator. The collected images are resized, rescaled, and the contrast and brightness of the images are adjusted to increase the number of images in the dataset, which leads to a better generalization of images in the dataset and decreases the chances of model overfitting. The 8,133 images of the 18 crane signalman hand signals are passed through the DCNN to train the model to recognize the hand signals correctly. The architecture of the proposed DCNN model consists of 3 convolution layers, 3 max-pooling layers, three dense layers, and an output layer. The developed model achieved an accuracy of 89.1% and 84.6% in training and validation, respectively. The precision, recall and F1 score achieved by the model were 81.5%, 81.8%, and 81.7%, respectively, for the test set. The model is further validated for real-time hand signal classification, where accuracy of 87.9% is recorded. The F1 score was greater than 80% which means the model is performing well and the average accuracy of 87.9% in real-time crane signalman hand signal classification makes the model acceptable. However, during real-time classification, some misjudgments are made by the DCNN

model while classifying the signalman hand signals, but the crane operator is not relying solely on the classification label because the screen/ head-up display inside the operator cabin is showing the signalman hand signals in real time along with the classification label. The crane operator can still take the correct action by looking at the signalman hand signal.

In terms of future research, there is room for improvement. For example, a larger dataset of crane signalman hand signals images can be used to train the model, which would lead to an improvement in the performance of the developed DCNN model. Another technique is the use of transfer learning with a pre-trained model in place of using the current model, which can reduce the computational time and increase accuracy in terms of correctly classifying the crane signalman hand signals. Data fusion techniques can also be used to classify dynamic hand signals that will lead to an improved the accuracy of the deep-learning models. The development of such model is necessary in the crane industry as they help in the construction safety improvement and with well-achieved high accuracy; they assist towards automation of the process and development of automated cranes.

References

Arabi, S., Haghghat, A., and Sharma, A. 2019. "A deep learning based solution for construction equipment detection: from development to deployment." <http://arxiv.org/abs/1904.09021>.

Cha, Y. J., Choi, W., and Büyüköztürk, O. 2017. "Deep Learning-Based Crack Damage Detection Using Convolutional Neural Networks." *Computer-Aided Civil and Infrastructure Engineering*, 32(5), 361–378.

Chatfield, K., Simonyan, K., Vedaldi, A., and Zisserman, A. 2014. "Return of the devil in the details: Delving deep into convolutional nets." *Proceedings of the British Machine Vision Conference 2014*, 1–11.

Everett, B. J. G., and Slocum, A. H. 1993. "Device for Improving Crane." *Journal of Construction Engineering and Management*, 119(1), 23–39.

Fang, Y., and Cho, Y. K. 2016. "A framework of lift virtual prototyping (LVP) approach for crane safety planning." *Proceedings of the 33rd International Symposium on Automation and Robotics in Construction*, 291–297.

Gu, Y., Xu, S., Wang, Y., and Shi, L. 2019. "An advanced deep learning approach for safety helmet wearing detection." *Proceedings of the 2019 IEEE International Congress on Cybermatics: 12th IEEE International Conference on Internet of Things, 15th IEEE International Conference on Green Computing and Communications, 12th IEEE International Conference on Cyber, Physical and Social Computing, and IEEE Smart Data*, 669–674.

Hu, J., Geo, X., Wu, H., and Gao, S. 2019. "Detection of Workers Without the Helments in Videos Based on YOLO V3." *Proceedings of the 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 1553–1560.

LeCun, Y., and Bengio, Y. 1995. "Convolutional Networks for Images, Speech, and Time-Series." *Handb. brain Theory Neural Netw.* 3361 (10).

Li, H., Chan, G., and Skitmore, M. 2013. "Integrating real time positioning systems to improve blind lifting and loading crane operations." *Construction Management and Economics*, 31(6), 596–605.

- Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., and Al-Hussein, M. 2020. Conceptual Framework for Safety Improvement in Mobile Cranes. *Proceedings of the Construction Research Congress 2020: Computer Applications* (pp. 964–971). Reston, VA: American Society of Civil Engineers.
- Nair, V., and Hinton, G. E. 2010. Rectified linear units improve restricted boltzmann machines. *Proceedings of the International Conference on Machine Learning*, 807–814.
- Occupational Safety and Health Administration (OSHA). 2021, Cranes and Derricks in Construction (1926.1408), <https://open.alberta.ca/dataset/757fed78-8793-40bb-a920-6f000853172b/resource/9296e033-fd12-40dc-ac86-21e5873d4161/download/4403880-part-6-cranes-hoists-and-lifting-devices.pdf> (2009) (accessed December 19, 2021)
- Rosenfeld, Y. 1995. “Automation of existing cranes: from concept to prototype.” *Automation in Construction*, 4(2), 125–138.
- Shapira, A., Rosenfeld, Y., and Mizrahi, I. 2008. “Vision system for tower cranes.” *Journal of Construction Engineering and Management*, 134(5), 320–332.
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958
- Wang, L., Xie, L., Yang, P., Deng, Q., Du, S., and Xu, L. 2020. “Hardhat-wearing detection based on a lightweight convolutional neural network with multi-scale features and a top-down module.” *Sensors (Switzerland)*, 20(7), 3–7.

Wu, J., Cai, N., Chen, W., Wang, H., and Wang, G. 2019. "Automatic detection of hardhats worn by construction personnel: A deep learning approach and benchmark dataset." *Automation in Construction*, 106, 102894.

Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Al-Hussein, M., and Kurach, L. 2020. "A deep learning-based framework for an automated defect detection system for sewer pipes." *Automation in Construction*, 109, 102967.

Zavichi, A., and Behzadan, A. H. 2011. "A Real Time Decision Support System for Enhanced Crane Operations in Construction and Manufacturing." *Computing in Civil Engineering*, 194–201.

Zekavat, P. R., Moon, S., and Bernold, L. E. 2014. "Holonc construction management: Unified framework for ICT-supported process control." *Journal of Management in Engineering*, 31(1), 1–15.

APPENDIX 2:

THE EFFECTIVENESS OF DATA AUGMENTATION IN CONSTRUCTION SITE-RELATED IMAGE CLASSIFICATION⁶

INTRODUCTION

In recent times as the construction industry is moving from traditional construction operations to modular construction techniques, the construction sites are getting more overwhelmed with on-site resources such as construction machinery, workers, etc., that make the construction operation less productive and unsafe for workers. The implementation of recent technological advancements can be useful to make the construction operation's working conditions more smooth, productive, and safe.

The adoption of technological advancement in the construction industry is moving at a very slow pace compared to other industries such as automotive, healthcare, manufacturing industry, etc. However, researchers in the construction field have leveraged technological advancements such as deep-learning techniques, IOT, and Big Data to solve the construction site-related problems. Deep learning has been studied and implemented by researchers in the construction industry to assess construction worker's ergonomics (Yu et al. 2019), on-site construction machinery, and their activity detection (Xiao and Kang 2021a), detection and classification of multiple objects such as construction hardhat (Fang et al. 2018) and waste materials (Davis et al. 2021).

⁶ A version of this work has been accepted for publication as follows: Mansoor, A., Liu, S., Ali, G. M., Bouferguene, A., Al-Hussein, M. "Effectiveness of data augmentation in construction site-related image classification." Accepted (Apr., 2022) for publication in *Proceedings of the Canadian Society for Civil Engineering Annual Conference*, Whistler, BC, Canada, May 25–28, 2022.

Meanwhile, the construction industry is facing multiple challenges in adopting the deep-learning techniques in day-to-day construction operations for example issues on data privacy, high cost of computational machines to process the data, requiring expert engineers with the knowledge of technology, etc. The most significant challenge is the lack of availability of construction site-related image datasets. Where deep-learning models require significant amount of data for training purposes from multiple backgrounds, variable weather and lighting conditions, etc., to improve the accuracy of the deep-learning model. The case of implementing deep learning in construction sites has multiple challenges. First, the amount of open-source construction site related data is very limited in the industry; second, construction site is dynamic in nature where the workers, machinery and other background work conditions are changing every minute, and finally, the construction sites are congested which makes deep-learning model less effective to detect and classify objects. The weather and lighting conditions of the construction site also affect the accuracy of deep-learning models.

To overcome the issue of small dataset and different lighting conditions, data augmentation can be used. Data augmentation is a technique that is used to apply realistic transformation to the dataset (Hernández-García and König 2018). This technique can assist to reduce the chances of model overfitting, solve data imbalance and data biasness issues (Mikołajczyk and Grochowski 2018; Shorten and Khoshgoftaar 2019). The data of mentation will also help to increase the size of the image dataset by duplicating the samples in the dataset. In this research, we will be using You Only Look Once (YOLOv4) state-of-the-art deep-learning model to classify construction site-related images of construction equipment and construction site workers hand signals. The YOLOv4 deep-learning model is capable, when trained, of looking at an image and finding the subset of object class, enclosing it with the bounding boxes, and identifying its class (Bochkovski

et al. 2020). The reason for using the YOLOv4 model is its superior performance over other state-of-the-art deep-learning models such as YOLOv3, Faster Recurrent-Convolutional neural network (Faster R-CNN), and Single Shot Detector (SSD) (Bochkovskiy et al. 2020).

The present work is organized as follows: Section 2 describes the related work in the context of implementing data augmentation in image classification. Section 3 defines the methodology and image data selection. Section 4 and Section 5 introduce the different augmentation techniques used in research and the architecture of the state-of-the-art YOLOv4 model. Section 6 presents the experimental setup for the research. Section 6 and Section 7 contain the results and the conclusion of the study.

RELATED WORK

The implementation of data augmentation techniques is very important to improve the performance of deep-learning models. For example, Walawalkar et al. (2020) introduced an attentive CutMix data augmentation method to enhance the performance of deep learning image classification. The experiment was conducted on CIFAR10/100 image dataset, and they used various CNN architectures such as ResNet, DenseNet, and EfficientNet. The results indicated that the developed approach increased the classification accuracy by 1.5% compared to the traditional data augmentation technique and by 3.04% compared to baseline non-augmented techniques. Perez and Wang (2017) used traditional transformations such as zoom out, zoom in, and shaded, Generative Adversarial Networks (GANs) such as Cezanne and Enhance, and augmentation techniques on ImageNet and MNIST datasets. The smallNet CNN model is trained on the augmented and non-augmented datasets. The model trained on the augmented dataset showed a 6% increase in performance from the non-augmented dataset. Mikołajczyk and Grochowski (2018) trained VGG 16 model on skin melanomas diagnosis, histopathological images and breast

magnetic resonance imaging (MRI) datasets. They used GAN augmentation techniques and shear, reflection, rotation, etc., augmentation techniques and observed that implementation of data augmentation techniques improved the performance of the VGG 16 model. Gu et al. (2019) used VGG-16 architecture with the custom CNN model on CIFAR-10 image datasets. Rotation, width and height shift, flip, etc., augmentations are used to train the model. The result showed a 2.1% improvement in the accuracy of the model when trained with the augmented dataset. Lei et al. (2019) implemented rotation, solarize, invert, shear, colour balance, etc., augmentation techniques on CIFAR-10, MNIST and Fashion-MNIST datasets. The ResNet and LeNet-5 models are trained on augmented, and non-augmented image datasets. It is noted that the model trained on the augmented image dataset showed better accuracy in the classification of the test dataset.

The augmentation techniques and deep-learning models discussed above showed great results but the construction site has its challenges. For example, the low visibility due to dusty conditions in the construction site, congestion of construction sites, full of the obstacle, changes in outdoor lighting condition, dynamic nature and business of construction site make the construction site related image classification a challenging task. Therefore, we propose three different augmentation techniques (Gamma transformation, Gaussian blur, and salt-and-pepper noise) to mimic the construction site scenario changes and as well as to improve the classification performance of deep-learning models in construction site related images classification.

METHODOLOGY

We propose three different data augmentation techniques in which five different experiments are conducted. Initially, the YOLOv4 model is trained on the original image dataset without any augmentation techniques implemented. Second, we applied gamma transformation to the original image dataset where lighting conditions in the images are adjusted. Third, we applied Gaussian

blur to the original image dataset. Fourth, we tested the model by adding salt-and-pepper noise augmentation technique to the original image dataset; and finally, we combined all the above augmentation techniques to test the performance of the model.

Dataset Selection

For this work, we use Alberta Construction Image Dataset (ACID) image dataset (Xiao and Kang 2021b) and the construction site worker's hand signal image dataset. ACID is a construction machine detection dataset developed at the University of Alberta (Xiao and Kang 2021b). ACID was developed as a source to assist the use and development of deep-learning applications in the construction automation field. The dataset contains images collected from construction sites all over the world (Xiao and Kang 2021b). The construction site worker's hand signals image dataset is collected manually in 720×720 pixels in indoor and outdoor environments. The dataset has static and dynamic backgrounds as well as sunny and cloudy weather conditions. For the original image dataset, we used 1,000 images of construction vehicles from the ACID dataset and 1,000 images of construction site workers' hand signal image datasets. Further data augmentation techniques are applied to the original image dataset.

DATA AUGMENTATION TECHNIQUES

Data augmentation is a method to increase the size of the collected dataset. Using data augmentation techniques, the dataset can be diversified by applying a random but realistic transformation. The implementation of data augmentation can improve the performance of the deep-learning models (Hernandez-Garcia and Konig 2018), reduce the likelihood of model overfitting, as well as better, generalize the dataset for model training purposes (Mikołajczyk and Grochowski 2018). The following data augmentation techniques are used for the present work.

Gamma Transformation

Gamma transformation is a technique to control the intensity of light in the images. With gamma transformation, we can adjust the contrast and brightness of the original image (Bull 2014). This technique can bring more originality to the dataset and help the model to perform better in any lighting condition. For the current dataset, the range of gamma values in which to adjust the brightness and contrast is set to between -35 and $+35$, where a set of -35 will result in darker images and $+35$ will result in lighter images. This range is chosen to make the conditions more realistic and representative of weather conditions ranging from sunny/clear to cloudy/overcast.

Gaussian blur

Gaussian blur is used to blur the images in the dataset; it has been widely used for multiple purposes such as reducing the detail in the images. This technique smooths out the randomness in an image based on a chosen blur radius. Each pixel in the frame will adopt a new value based on the weighted average of its surrounding pixels (Misra and Wu 2020), where more weight is given to pixels in closer proximity. The amount of blur in the frame is measured in pixels (px), where a higher value of px means more blur and a lower value means less blur will appear in the frame. To add blur to the dataset, a blur value between 1.15 to 3.25 px is chosen for the present work. Adding blur to the images in the dataset ensures that the model will be trained for a scenario where the camera may lose focus, thereby allowing the model to achieve better accuracy.

Salt-and-pepper noise

The “salt-and-pepper” noise data augmentation technique is used for image degradation. In this technique, some pixels of the frame are kept very noisy. The effect is likened to sprinkling salt and pepper on the frame (Boncellet 2009), hence the name. The intensity of salt-and-pepper noise is measured in percentage, where 0% means there is no noise in the frame. For the current work, 10% to 40% salt-and-pepper noise is applied. This helps the model to be trained for unintentional and

unwanted changes in the scenes. It also helps the model to correctly classify hand signals captured from low-quality cameras.

The reason for using these augmentation techniques is to better generalize the training dataset and to mimic the construction site environment, which is typically subject to changing weather, dusty conditions, among other dynamic and unpredictable conditions. After applying these transformations, the data augmentation techniques randomly choose each image, apply the transformation, and store the image as a different image. The example of discussed augmentation is given in Figure 1.

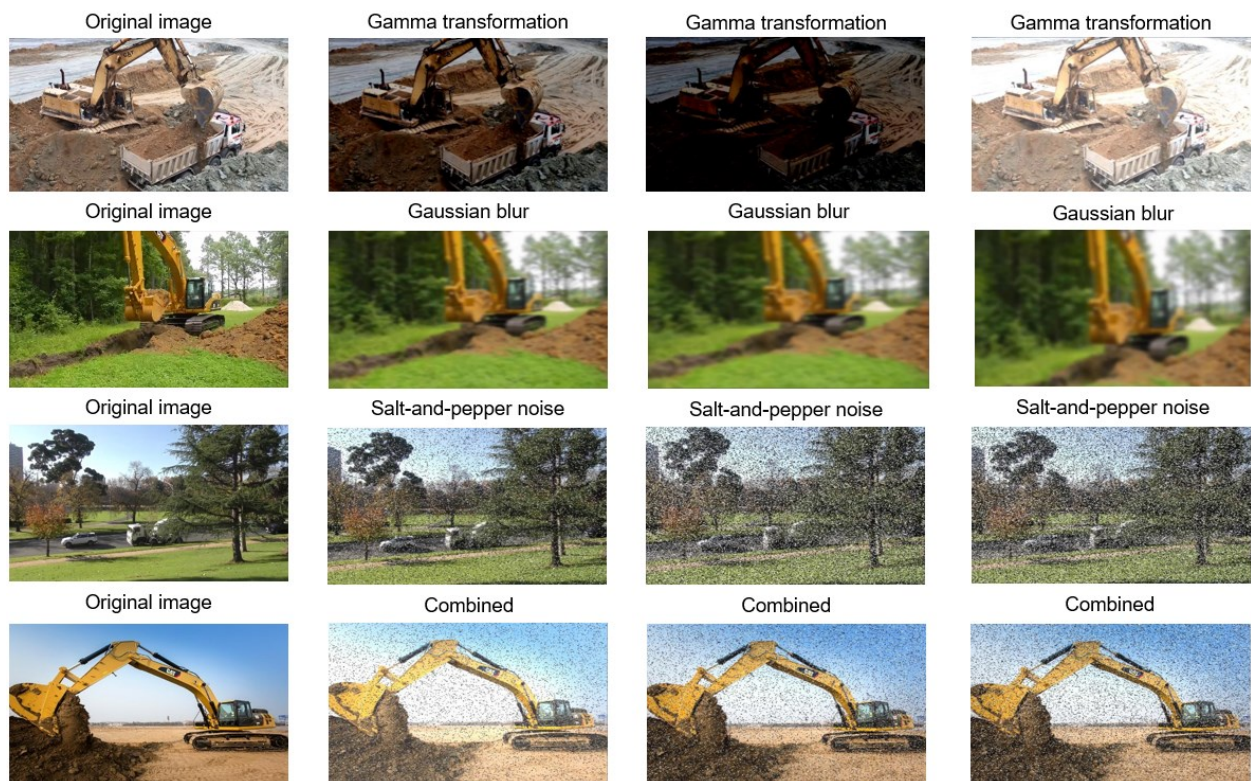


Figure 1: Example of different augmentation techniques.

STATE-OF-THE-ART YOLOv4 MODEL ARCHITECTURE

YOLOv4 is a deep learning-based object detection and classification model capable of inspecting an image to find the subset of object class, enclosing it within bounding boxes, and identifying its class when trained (Bochkovskiy et al. 2020). YOLOv4 is the updated version of the YOLO algorithm series, which has been used in many applications in various fields, such as in the classification of safety helmets (Hu et al. 2019), construction vehicles (Hou et al. 2021), defects in sewer pipes (Yin et al. 2020), etc. The YOLOv4 model is selected for the present work based on its superior accuracy and speed compared to other deep-learning models, such as YOLOv3, Faster R-CNN, SSD, and RetinaNet (Bochkovskiy et al. 2020). The YOLOv4 model architecture consists of three parts—the backbone, neck, and head—as shown in Figure 2.

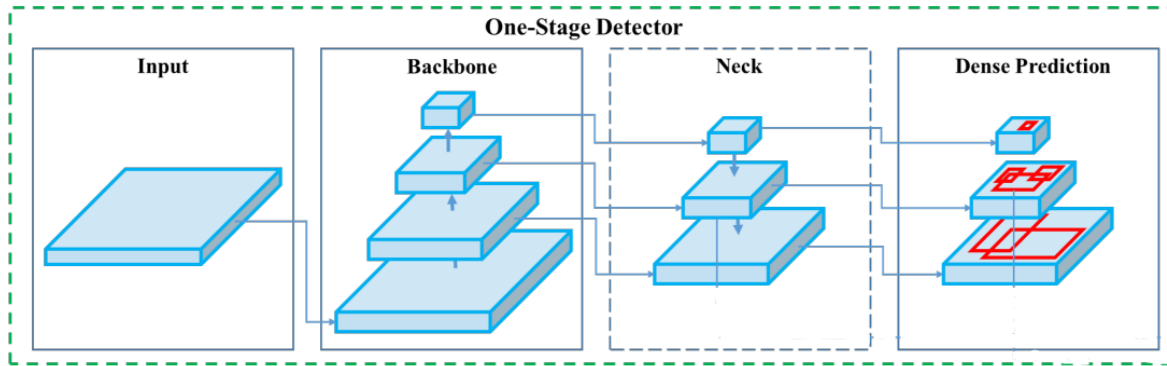


Figure 2: Architecture of state-of-the-art YOLOv4 model (Bochkovskiy et al. 2020).

The backbone of YOLOv4 is composed of cross-stage partial connection CSPDarknet53. CSPDarknet53 is a CNN with residual connections that are responsible for extracting features from the image dataset (Bochkovskiy et al. 2020). The neck of YOLOv4 architecture is composed of spatial pyramid pooling (SPP) block (Huang et al. 2020) and Path aggregation network (PANet) (Liu et al. 2018) is responsible to collect the feature maps from different stages in the network. The YOLOv4 network architecture used the same YOLOv3 head, aiming to predict objects in multiscale. The details about the architecture of YOLOv4 are given in (Bochkovskiy et al. 2020)

EXPERIMENTAL SETUP

The YOLOv4 model has trained on both (ACID and Workers hand signals) the datasets. The images in the datasets are of varying resolutions. To feed the images in the YOLOv4 model the resolution of the images is scaled down to 416×416 pixels. The reason for scaling down the resolution of the images is to reduce the computational expense of the system. Initially, the model is trained on the original image datasets. In the original dataset, the images were used without any enhancement or augmentation to training the YOLOv4 model. After training the model the training accuracy of the model is recorded. Furthermore, the model is evaluated on the test dataset. The test dataset is the images that are not used while training the YOLOv4 model. Second, the model is trained on the images with gamma transformation. The detail about gamma transformation is discussed in Section 4.1. Here the original images and some images with gamma transformation are used to train the model. The accuracy of the model is evaluated based on training and test dataset. Third, the original dataset is used with Gaussian blur data augmentation. The detail about Gaussian blur is explained in Section 4.2. The YOLOv4 model is trained on the original images along with Gaussian blur images and the performance of the model is evaluated on training and test datasets. Further salt-and-pepper noise data augmentation technique is implemented to the original image dataset. The original images and some images with salt-and-pepper noise are used to train the YOLOv4 model. The accuracy of the model is recorded on the training and test dataset. Finally, all the augmented techniques are combined and implemented on original images. Here each augmented images have gamma transformation, Gaussian blur and salt-and-pepper noise. The model is trained on the combined augmented and original dataset. The model is evaluated on training and test dataset accuracies. Note that the same test dataset images are used to evaluate all

the models while the models are trained on different datasets. As well as the same architecture of the YOLOv4 state-of-the-art model is used in this work without any changes or improvement.

RESULTS AND DISCUSSION

A total of 5 experiments are conducted where the YOLOv4 model is trained on the non-augmented dataset, augmented with gamma transformation, augmented with Gaussian blur, augmented with salt-and-pepper noise and finally, the model is trained with combined augmented images. The model is trained on 6500 epochs/iteration. Based on the experimental results as shown in Figure 3, the state-of-the-art YOLOv4 model achieved the best training accuracy of 92.4% on a non-augmented dataset where the model achieved an accuracy of 86.8% in the test dataset. When random transformations, such as gamma transformation, are implemented on the original image dataset, a training accuracy of 90.7% is recorded, while, when the model is used to evaluate the test dataset, the model achieves an accuracy of 87.2%. When the YOLOv4 model is trained on, the images with Gaussian blur the model achieved an accuracy of 91.3% and 88.9% in training and test datasets. The YOLOv4 model is also trained on the images with salt-and-pepper noisy images; the model achieved a training accuracy of 91.8% and test accuracy of 87.9%. Furthermore, by combining all the augmentation techniques and the YOLOv4 model is trained on the augmented combined dataset, the model achieved an accuracy of 92.7% in training. When the model is deployed for the test dataset, the model achieved an accuracy of 90.8% in the test dataset.

The results showed that the YOLOv4 model achieved the highest accuracy in the augmented combined dataset in training as well as in the test dataset. For this case, we will be evaluating the performance of the models on the test dataset. The test dataset used in this work contains the same images while the models are trained on different datasets. The augmented combined dataset trained model achieved an accuracy of 90.8% in the test dataset, which is 4% better than the model trained

with the non-augmented dataset. The combined augmented dataset trained model also performs better than the model trained on augmented with gamma transformation, Gaussian blur, and salt-and-pepper noise by 3.6%, 1.9%, and 2.9%, respectively. It is also noted that all models trained with an augmented dataset perform better than the model trained on a non-augmented dataset when evaluated on the test dataset.

These results clearly indicate that the data augmentation techniques implemented lead to an improvement in the classification results of the state-of-the-art YOLOv4 deep-learning model. Data augmentation is an effective technique to increase the size of the dataset, better generalize the dataset and help improve the classification accuracy of deep-learning models. It also seems that the augmentation techniques chosen in the study to train the model perform well to mimic the construction site changing scenarios. However, the implementation of data augmentation techniques to the image dataset is a time-consuming task. The applicability of data augmentation for video analysis of construction sites is under study for future work.

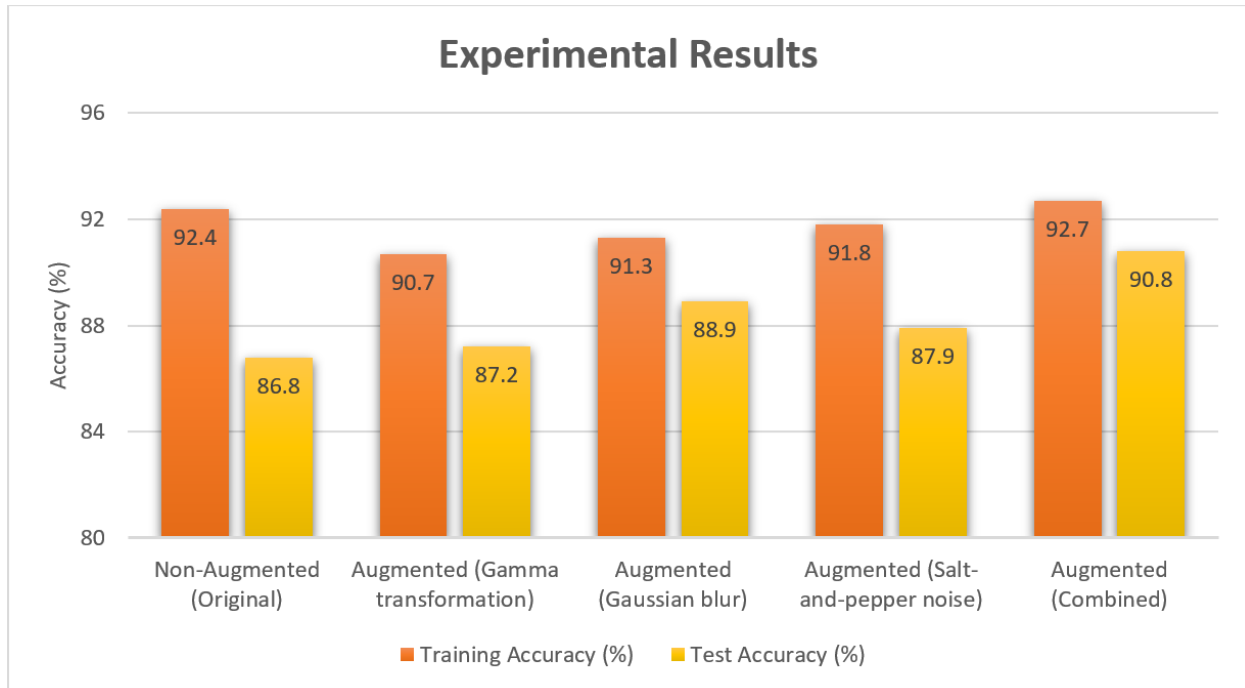


Figure 3: Comparison of YOLOv4 model accuracy between original and different augmented image datasets.

CONCLUSIONS

In this work, three different data augmentation techniques (gamma transformation, Gaussian blur, and salt-and-pepper noise) are used to evaluate the performance of the state-of-the-art YOLOv4 model with augmented and non-augmented construction site image datasets (ACID and construction site worker's hand signals). A total of five experiments are conducted. Initially, the state-of-the-art YOLOv4 model is trained on the original image dataset without any augmentation and evaluated based on the accuracy of the training and test dataset. Second, data augmentation is applied to the dataset one by one using gamma transformation, Gaussian blur, and salt-and-pepper noise. Finally, all the data augmentation techniques are combined and the model is trained on a combined dataset and the performance of the model is evaluated. Based on the experiments conducted the highest test dataset accuracy is achieved by the state-of-the-art YOLOv4 model

when trained on the combined dataset with 90.8% which is 4% more than the model trained on the non-augmented dataset. In addition, the model performed better when trained on the augmented dataset from the model trained on the non-augmented dataset. The result is convincing that the data augmentation technique is an effective way to mimic the construction site scenario changes. Data augmentation also helps to better generalize the dataset as well as increase the size of the image dataset. The future work will include applying data augmentation techniques in videos and real-time classification in dynamic construction sites.

REFERENCES

- Bochkovski, A., Wang, C., and Liao, H. M. 2020. "YOLOv4: Optimal Speed and Accuracy of Object Detection." *arXiv preprint arXiv:2004.10934*.
- Boncellet, C. 2009. Image noise models. *The Essential Guide to Image Processing*, A. Bovik (Ed.), pp. 143–167, Academic Press.
- Bull, D. 2014. Digital picture formats and representations. *Communicating pictures*, 99–132.
- Davis, P., Aziz, F., Newaz, M. T., Sher, W., and Simon, L. 2021. "The classification of construction waste material using a deep convolutional neural network." *Automation in Construction*, 122, 103481.
- Fang, Q., Li, H., Luo, X., Ding, L., Luo, H., Rose, T. M., and An, W. 2018. "Detecting non-hardhat-use by a deep learning method from far-field surveillance videos." *Automation in Construction*, 85, 1–9.
- Gu, S., Pednekar, M., and Slater, R. 2019. "Improve Image Classification Using Data Augmentation and Neural Networks." *SMU Data Science Review*, 2(2), 1–43.

- Hernandez-Garcia, A., and König, P. 2018. “Further Advantages of Data Augmentation on Convolutional Neural Networks.” *Springer Nature Switzerland*, 95–103.
- Hernández-García, A., and König, P. 2018. “Further advantages of data augmentation on convolutional neural networks.” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 95–103.
- Hou, X., Zhang, Y., and Hou, J. 2021. *Application of YOLO V2 in Construction Vehicle Detection. Lecture Notes on Data Engineering and Communications Technologies*, Springer International Publishing.
- Hu, J., Geo, X., Wu, H., and Gao, S. 2019. “Detection of Workers Without the Helments in Videos Based on YOLO V3.” *Proceedings of the 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, 1553–1560.
- Huang, Z., Wang, J., Fu, X., Yu, T., Guo, Y., and Wang, R. 2020. “DC-SPP-YOLO: Dense connection and spatial pyramid pooling based YOLO for object detection.” *Information Sciences*, 522, 241–258.
- Lei, C., Hu, B., Wang, D., Zhang, S., and Chen, Z. 2019. A preliminary study on data augmentation of deep learning for image classification. *Proceedings of the 11th Asia-Pacific Symposium on Internetware* (pp. 1–6).
- Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. 2018. “Path Aggregation Network for Instance Segmentation.” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 8759–8768.

- Mikołajczyk, A., and Grochowski, M. 2018. “Data augmentation for improving deep learning in image classification problem.” *International Interdisciplinary PhD Workshop (IIPhDW)*, IEEE, 117–122.
- Misra, S., and Wu, Y. 2020. Machine learning assisted segmentation of scanning electron microscopy images of organic-rich shales with feature extraction and feature ranking. *Machine Learning for Subsurface Characterization*, 289.
- Perez, L., and Wang, J. 2017. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.
- Shorten, C., and Khoshgoftaar, T. M. 2019. “A survey on Image Data Augmentation for Deep Learning.” *Journal of Big Data*, Springer International Publishing, 6(1), 1–48.
- Walawalkar, D., Shen, Z., Liu, Z., and Savvides, M. 2020. “Attentive Cutmix: An Enhanced Data Augmentation Approach for Deep Learning Based Image Classification.” *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 3642–3646.
- Xiao, B., and Kang, S.-C. 2021a. “Vision-Based Method Integrating Deep Learning Detection for Tracking Multiple Construction Machines.” *Journal of Computing in Civil Engineering*, 35(2), 04020071.
- Xiao, B., and Kang, S.-C. 2021b. “Development of an Image Data Set of Construction Machines for Deep Learning Object Detection.” *Journal of Computing in Civil Engineering*, 35(2), 05020005.

Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Al-Hussein, M., and Kurach, L. 2020. “A deep learning-based framework for an automated defect detection system for sewer pipes.”

Automation in Construction, 109, 102967.

Yu, Y., Li, H., Umer, W., Dong, C., Yang, X., Skitmore, M., and Wong, A. Y. L. 2019.

“Automatic Biomechanical Workload Estimation for Construction Workers by Computer Vision and Smart Insoles.” *Journal of Computing in Civil Engineering*, 33(3), 1–13.