DATA-DRIVEN STRATEGIES FOR SYSTEMIC RISK MITIGATION AND RESILIENCE MANAGEMENT OF INFRASTRUCTURE PROJECTS

DATA-DRIVEN STRATEGIES FOR SYSTEMIC RISK MITIGATION AND RESILIENCE MANAGEMENT OF INFRASTRUCTURE PROJECTS

By

Ahmed W. Gondia

B.Sc., M.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy

McMaster University

Doctor of Philosophy (2021)	McMaster University
(Civil Engineering)	Hamilton, Ontario
TITLE:	Data-driven strategies for systemic risk mitigation and resilience management of infrastructure projects
AUTHOR:	Ahmed Gondia
	B.Sc., M.Sc. (German University in Cairo)
SUPERVISORS:	Dr. Wael El-Dakhakhni
	Dr. Mohamed Ezzeldin
NUMBER OF PAGES:	xvii, 319

ABSTRACT

Public infrastructure systems are crucial components of modern urban communities as they play major roles in elevating countries' socio-economics. However, the inherent complexity and systemic interdependence of infrastructure construction/renewal projects have left sites hindered with multiple forms of performance disruptions (e.g., schedule delays, cost overruns, workplace injuries) that result in long-term consequences such as claims, disputes, and stakeholder dissatisfactions. The evolution of advanced data-driven tools (e.g., machine learning and complex network analytics) can play a pivotal role in driving improvements in the management strategies of complex projects due to such tools' usefulness in applications related to interdependent systems. In this respect, the research presented in this dissertation is aimed at developing data-driven strategies geared towards a resilience-based approach to managing complex infrastructure projects. Such strategies can support project managers and stakeholders with data-informed decision-making to mitigate the impacts of systemic interdependence-induced risks at different levels of their projects. Specifically, the developed data-driven resilience-based strategies can empower decision-makers with the ability to: i) predict potential performance disruptions based on real-time and dynamic project conditions such that proactive response/mitigation strategies and/or contingencies can be deployed ahead of time; and ii) develop adaptive solutions against potential interdependence-induced cascade project disruptions such that rapid restoration of the most important set of performance targets can be restored. It is important to note that data-driven strategies and other analytics-based approaches are not proposed herein to replace but rather to complement the expertise and sensible judgment of project managers and the capabilities of available analysis tools. Specifically, the enriched predictive and analytical insights together with the proactive and rapid adaptation capabilities facilitated by the developed strategies can empower the new paradigm of resilience-guided management of complex dynamic infrastructure projects.

DEDICATIONS

In memory of my beloved grandfather

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my supervisors, Dr. Wael El-Dakhakhni and Dr. Mohamed Ezzeldin, whose knowledge and innovative insight inspired this dissertation. Without their continuous support and guidance, I would not have been able to accomplish all that I have. I am especially grateful for their mentorship which has helped me become a better version of myself and shaped the person I am today.

Special thanks are due to my supervisory committee members, Dr. Michael Tait and Dr. Elkafi Hassini, for their constructive feedback and valuable input which significantly impacted the quality of the dissertation. Thank you for your time, effort and advice during every meeting.

There are no words that could describe my gratitude to my wife for her patience, understanding and encouragement throughout the toughest and the best of times. She has happily agreed to join me on this journey away from home and supported me beyond her fair share. Hana, I hope someday I can bring you as much joy as you have brought me over the years.

I would like to thank my parents, Fathia and Wael, who taught me the value of hard work by their own example and encouraged my education for as long as I can remember. Over the years, they have equipped me with the tools to overcome what life throws my way. Anything I have achieved and will achieve in my life is a result of their tireless efforts and unconditional love.

Finally, I would like to dedicate this work to my late grandfather who was my role model. Throughout my life, he inspired me with his work ethic and helped me get through the hardest of times with his endless support and kind heart. He was the first person I eagerly rushed to with news of my PhD application acceptance, which had been one of his dreams even though it would mean I would be away from home and the family. Although he never saw the end of this adventure, I truly hope I have made him proud.

CO-AUTHORSHIP

This thesis has been prepared in accordance with the regulations for a sandwich thesis format or as a compilation of research papers stipulated by the Faculty of Graduate Studies at McMaster University. This thesis presents research and analytical work carried out solely by Ahmed Gondia. Advice and guidance were provided for the whole thesis by the academic supervisors Dr. Wael El-Dakhakhni and Dr. Mohamed Ezzeldin. For Chapter 2, Dr. Ahmad Siam provided comments and edits, and Dr. Ayman Nassar assisted in collecting the dataset used in the study. Information presented from outside sources, which has been used towards analysis or discussion, has been cited where appropriate; all other materials are the sole work of the author. This thesis consists of the following manuscripts in the following chapters:

Chapter 2

Gondia, A., Siam, A., El-Dakhakhni, W., and Nassar, A. H. (2020). "Machine Learning Algorithms for Construction Projects Delay Risk Prediction." *Journal of Construction Engineering and Management*, 146(1), 04019085.

Chapter 3

Gondia, A., Ezzeldin, M., and El-Dakhakhni, W. (2021). "Machine Learning-based Decision Support Framework for Construction Injury Severity Prediction and Risk Mitigation." Submitted for publication and under review as of November 2021.

Chapter 4

Gondia, A., Ezzeldin, M., and El-Dakhakhni, W. (2021). "Machine Learning-based Construction Site Risk Models." Submitted for publication and under review as of November 2021.

Chapter 5

Gondia, A., Ezzeldin, M., and El-Dakhakhni, W. (2021). "Dynamic Networks for Resilience-driven Management of Infrastructure Projects." Submitted for publication and under review as of November 2021.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1:

INTRODUCTION

1.1 BACKGROUND AND MOTIVATION

1.1.1 Industry Challenges

Public infrastructure systems (e.g., power, water/wastewater, telecommunication, and transportation) function as arteries of modern urban communities as they provide vital services to meet societal, economic, and political needs (Di Maddaloni and Davis 2018). Construction projects of such systems typically: 1) have long schedules and large budgets; 2) involve work scopes with high degrees of technical complexity and uncertainty; 3) require substantial resources and diverse specialized expertise; 4) need the collaboration of a multidisciplinary workforce each with their own constraints and uncertainties; and 5) spread spatially over a large geographical area of dynamic work environments (Sun and Zhang 2011; Flyvbjerg 2014; Luo et al. 2016). The success of such projects is typically measured by key performance indicators (**KPI**) such as schedule delay, cost overrun, workplace injuries, quality deficits, resource disruptions, etc. However, the above-listed characteristics set construction apart from many other industries and pose significant challenges to the management of its projects. As a result, projects' inability to meet basic performance targets has been globally recognized (Yeo 1995; Han et al. 2009;

Cantarelli et al. 2012; McKinsey Global Institute 2017).

For example, 3,632 projects in Canada were reported as delayed at the time of writing, as shown in Figure 1.1 (Construct Connect 2021). It should be appreciated that it is neither a federal nor a provincial requirement on construction organizations to report delayed projects to the above-cited platform, and thus the actual numbers for each jurisdiction may be more than what is shown in the figure. Nonetheless, even if the presented data are considered as the lower bounds of delayed projects, the problem is still clear. In addition, the Canadian construction industry in 2019 accounted for the third highest number of injuries (28,111) and the highest number of fatalities (204) in the workplace among all other major industries, as shown in Figures 1.2a and 1.3a, respectively (AWCBC 2021; Statistics Canada 2021). It can also be seen through Figures 1.2c and 1.3c that the Canadian construction industry was constantly ranked within the top three industries with the highest injury and fatality rates, respectively, over the past two decades.

Such alarming KPI examples, as well as other challenges faced by the industry, can be attributed to the industry's inherent complexity and systemic risks which stem from the interdependence of many of its components (Jarkas 2017). Within the current research program, construction systemic risk is looked at from two levels: ***intra-KPI*** and ***inter-KPI systemic (interdependence-induced) risks***.

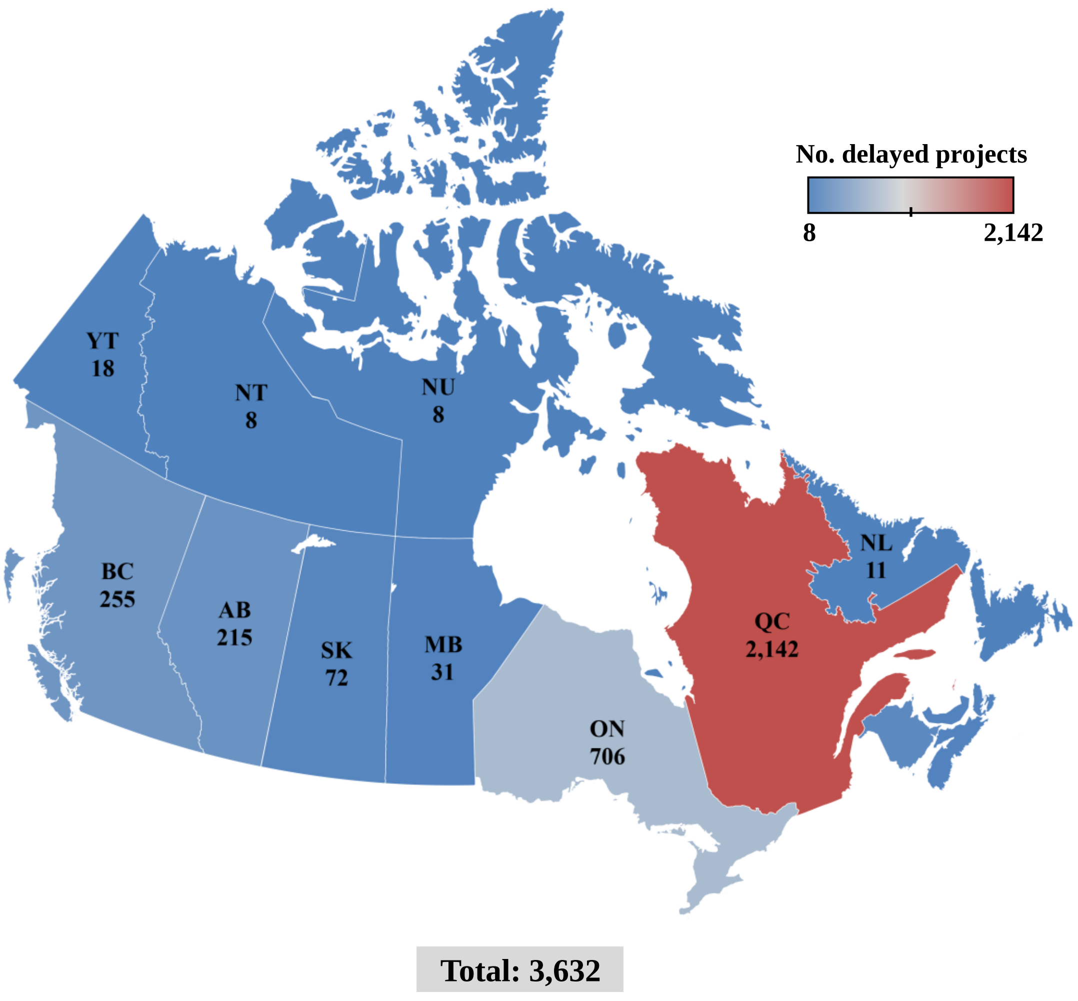


Figure 1.1: Number of delayed construction projects in Canada per province as of September 2021 (based on data from Construct Connect, 2021)

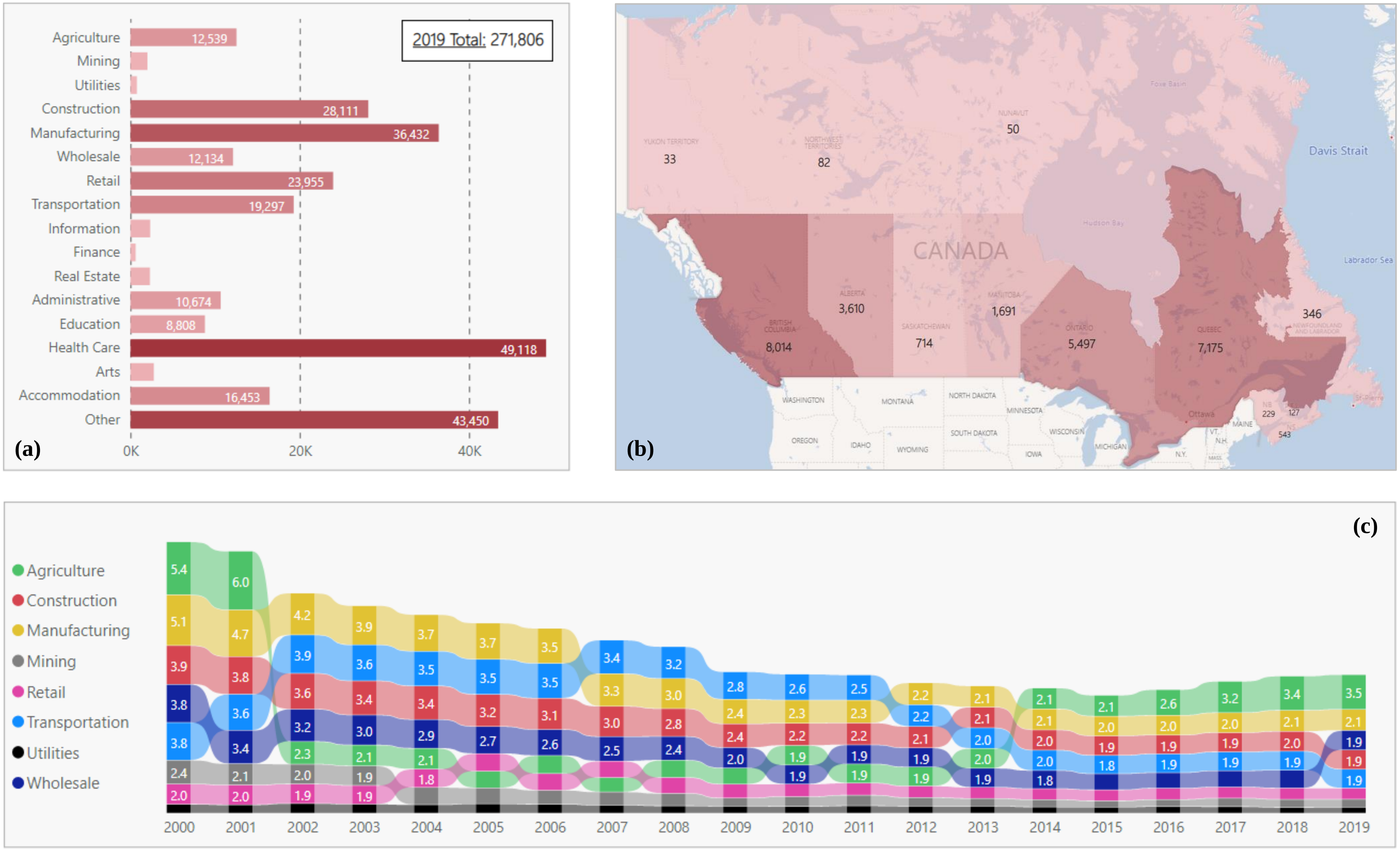


Figure 1.2: Lost-time injuries in Canada: a) all industry counts, 2019; b) construction counts by province, 2019; and c) major industry rates (per 100 workers), 2000-2019 (based on data from AWCBC, 2021 and Statistics Canada, 2021)

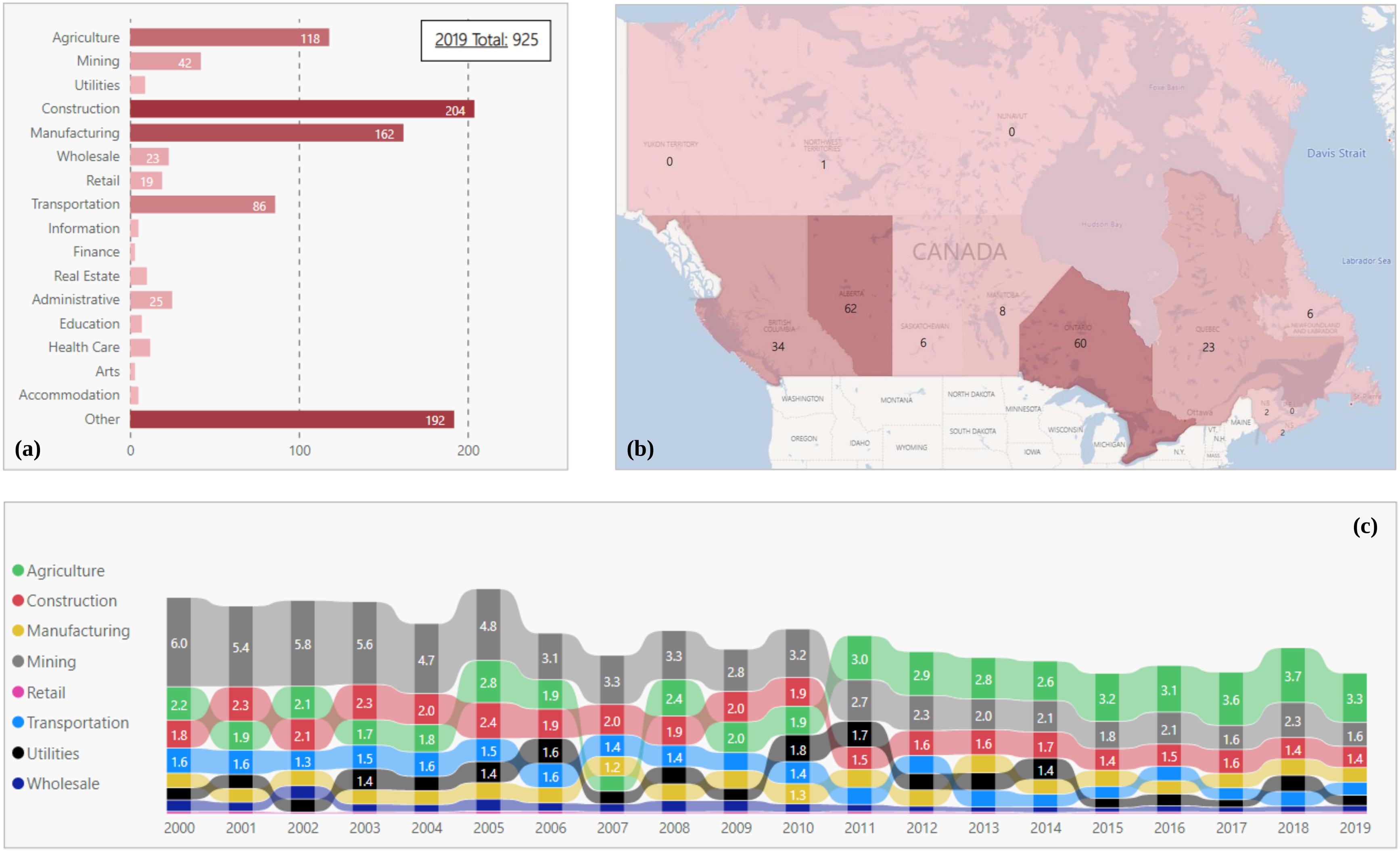


Figure 1.3: Workplace fatalities in Canada: a) all industry counts, 2019; b) construction counts by province, 2019; and c) major industry rates (per 10,000 workers), 2000-2019 (based on data from AWCBC, 2021 and Statistics Canada, 2021)

1.1.2 Intra-KPI Systemic Risks

Disruption to one KPI can be attributed to the combined and interdependent effects (as opposed to the independent effects alone) of multiple underlying factors affecting such KPI. For example, schedule delay can be initiated by a i) *change request* from the project owner which would require the introduction of ii) *new construction techniques/technologies* and subsequently iii) *re-design*. This situation may thus induce iv) delays in *preparing and approving design documents* by the consultant, v) challenges with *new equipment acquisition and site mobilization* and vi) re-work due to *contractor inexperience* with such new techniques. While any of such delay factors can occur on its own within a project and influence the schedule independently without triggering other factors, multiple factors can also materialize due to their interdependent nature as in the above example (Eriksson et al. 2017). Similarly, workplace injuries result from one or more injury precursors such as those related to the i) *surrounding worksite*, ii) *work means/methods*, iii) *exposure to hazards* and iv) *environmental conditions*. Such precursors are also interdependent where, for example, the type of hazard exposure is typically influenced by changes to worksite environment and/or weather conditions (Feng et al. 2014).

1.1.3 Inter-KPI Systemic Risks

Project performance targets, and thus KPIs, are also highly interconnected with one another, as illustrated in Figure 1.4. Specifically, disruption to one KPI

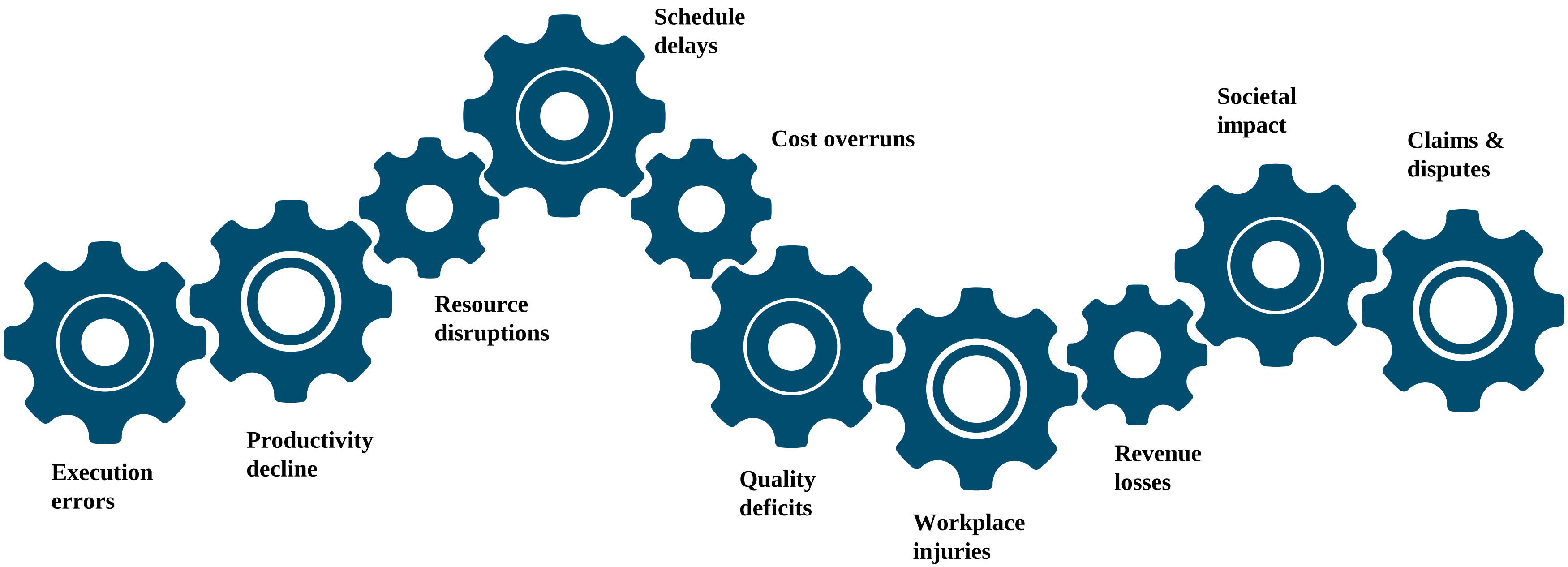


Figure 1.4: Inter-KPI interdependence-induced cascade potential

can potentially lead to interdependence-induced cascade disruptions extending to other KPIs (Serrador and Turner 2015). For instance, execution errors typically require multiple reworks, thus hindering productivity and disrupting resource allocations within a project, and subsequently incurring schedule delays and cost overruns (Larsen et al. 2015). In an attempt to adapt, project managers typically resort to accelerating progress through compressing schedules and/or crashing tasks, which may also have a negative impact on the finished work quality and the safety of the workers, which are catalysts for further schedule and cost complications (Nepal et al. 2006; Love at al. 2016). Such disruptions may ultimately result in delayed infrastructure project completion and operation-readiness, leaving governments and other stakeholders with lost revenues on project capital, and subsequently provoking negative societal perceptions and public controversies (Ndekugri et al. 2008; Di Maddaloni and Davis 2018). Such consequences, in turn, spark tensions between project stakeholders, where unresolved conflicts give rise to legal claims and disputes which have become increasingly common in infrastructure projects (Yates and Epstein 2006; Mehany et al. 2018).

1.1.4 Resilience-guided Project Management

Broadly speaking, the resilience of a system in the face of disruptive events denotes the ability to: 1) absorb the impacts of such disruptions through prior identification of systemic vulnerabilities and proactive preparedness; 2) adapt to such events by mobilizing risk management strategies aimed at preserving continued performance; and 3) rapidly recover from such events and restore the pre-disruption performance state through fully operationalizing the developed response strategies (Barker et al. 2013; Hernandez-Fajardo and Dueñas-Osorio 2013; Wilkinson et al. 2016; Hariri-Ardebili 2018). In that respect, the motivation behind the current research program is to adopt a resilience-guided approach to manage complex projects with the aim of enhancing projects' ability to mitigate systemic risks at the intra- and inter-KPI levels. The program scope is presented in two stages as shown in Figure 1.5a. The first stage initially looks at individual KPIs separately/in isolation as interdependent systems of inducing factors that influence such KPI. The first stage further works toward predicting potential KPI disruptions in projects such that proactive response/mitigation strategies and/or contingencies can be deployed ahead of time. It should be noted that this thesis' focus is on schedule delays and workplace injuries as KPIs. The second stage considers the interconnectedness between different KPIs and develops adaptive solutions against the resulting potential interdependence-induced cascade disruptions, thus facilitating a rapid restoration of the most important set of project performance targets. These two stages would thus improve the overall project resilience under both levels of interdependence-induced vulnerability.

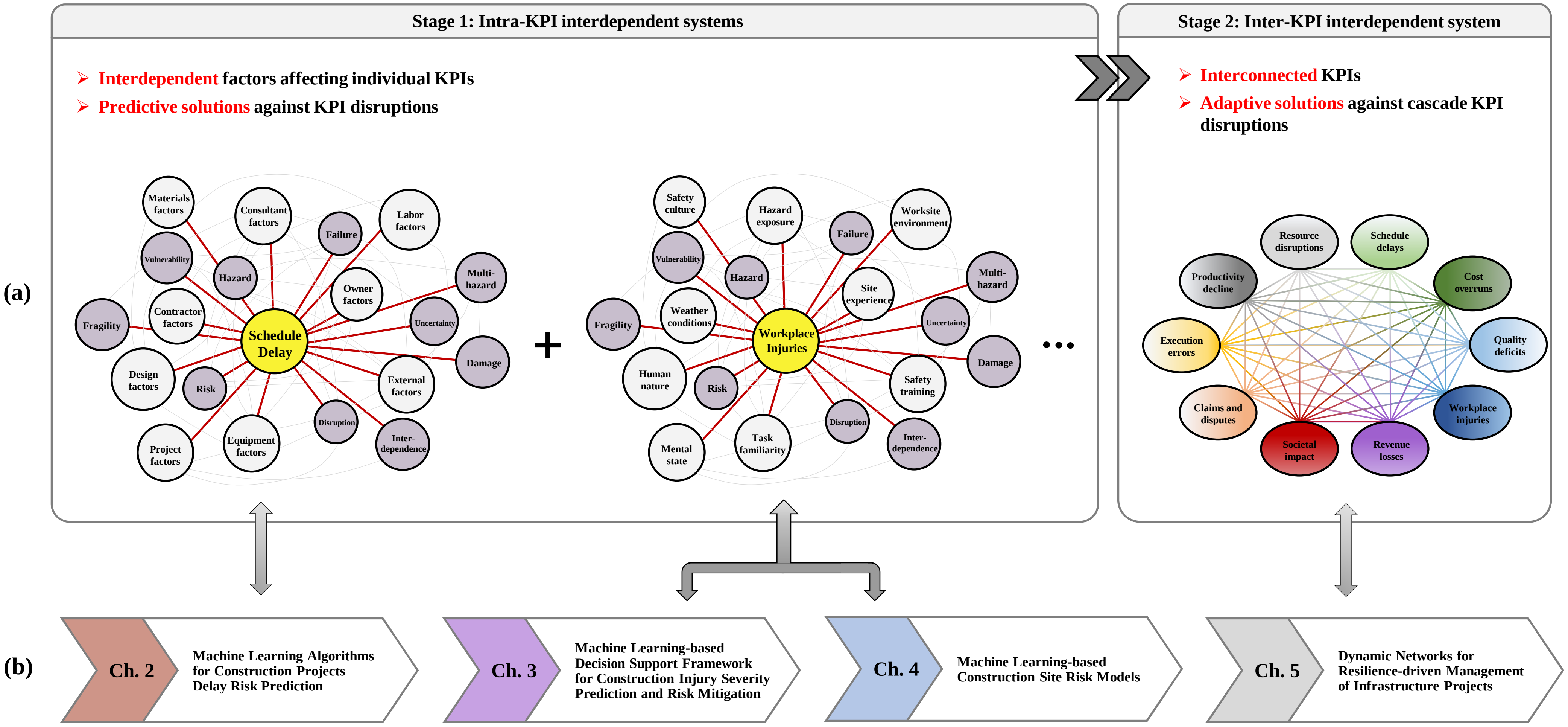


Figure 1.5: Multi-stage research program: a) scope and b) organization

1.1.5 Data-driven Approaches

Data analytics offers a suite of tools that can support actionable decision-making for construction stakeholders based on insights mainly generated from historical/previous project data. Following many other industries, construction is undergoing a technological shift driven by the fourth industrial revolution (Industry 4.0) which is aimed at digitalizing and automating the industry for improved productivity (García de Soto et al. 2019). This technological shift has also seen the construction industry experience rapid growth in the amount of data generated and collected throughout projects' day-to-day operations (Yan et al. 2020). Such digital data collection growth, in turn, has the potential to promote rapid adoption and application of data-driven strategies to solve construction-related problems and facilitate construction stakeholders in making proper decisions towards the better performance of construction projects. However, leveraging analytics tools that are capable of handling interdependent systems, such as those described earlier, is key. In this respect, machine learning (**ML**) is a known effective tool to model and predict outcomes of complex interdependent systems by learning the inherent relationships between inputs (e.g., delay factors or injury precursors) and outputs (e.g., schedule delay or injury incidences) enshrined within such systems' historical data while maintaining good generalization error (Bilal et al. 2016). The key advantage of ML methods is attributed to their capability of avoiding any prespecified model assumptions, unlike in statistical-based methods for instance, which is an important consideration when dealing with interdependent systems whose behaviors are largely complex and unknown a priori (Aggarwal 2016). Other key tools include those related to system simulation and complex dynamic network theory (**CDNT**) due to their ability to model, analyze and facilitate understanding of dynamic interdependencies in complex systems (such as those between project stakeholders or between KPIs) both over space and time and subsequently adapt their behaviors to mitigate possible systemic risk (Gong et al. 2017; Fu et al. 2018).

1.1.6 Related Work and Research Gaps

The past decade has seen significant advances in ML applications within construction project management research, such as those applied to cost estimation (Chao and Chien 2009; Cheng et al. 2009; Son et al. 2012; Ahiaga-Dagbui and Smith 2014; Williams and Gong 2014), quality performance assessment (Shi 2009; Naji et al. 2018; Fan 2020), project dispute classification and resolution (Mahfouz and Kandil 2012; Chou and Lin 2013; Chou et al. 2013), contractor performance classification and prequalification (Elazouni 2006; Kong and Yaman 2014), labor productivity assessment (Desai and Joshi 2010; Heravi and Eslamdoost 2015), and classification of heavy construction equipment (Fan et al. 2008; Gong et al. 2011; Rashid et al. 2019). Accurately predicting project schedules continues to represent a challenge for both researchers and project managers. Construction project schedule estimation has been studied extensively using statistical analysis-based methods such as multiple linear regression, bivariate correlation, factor analyses and Monte Carlo simulation (Chan 2001; Chan and Chan 2004; Rezaie et al. 2007; Hammad et al. 2008; Abu Hammad et al. 2010; Dursun and Stoy 2011; Kokkaew and Wipulanusat 2014; Avlijaš 2019; Tokdemir et al. 2019). Although such research directions introduce models for delay risk simulation, the underlying statistical assumptions together with the complex interdependent nature of such systems can limit the accurate prediction of project duration. In this respect, there is a general lack of research geared towards ML-based approaches; examples include works by Petruvesa et al. (2013), Sobhani and Madadi (2015), Wauters and Vanhoucke (2016) and Cheng et al. (2019). However, such works consider experimental/training datasets that either i) constitute general project attributes as model inputs (e.g., project type, contract type, contract amount, total floor area, number of contractors; ii) are computer-generated to produce fictitious progress executions; or iii) are heavily based on subjective data sources as from expert consultations/surveys. As such, approaches based on delay risk factors collected from objective/actual previous project datasets may prove a viable and more meaningful alternative.

In construction safety research, ML has been applied to predict the likelihood of incident types (Tixier 2016; Gerassis et al. 2017; Kang and Ryu 2019; Ayhan and Tokdemir 2020) and incident risk levels (Zhou et al. 2017; Poh et al. 2018; Sakhakarmi et al. 2019), and to assess construction safety climate scores across projects (Patel et al. 2015; Abubakar et al. 2018; Makki and Mosly 2021). While these studies have focused on employing black-box types of ML (e.g., random forests, support vector machines, artificial neural networks), there has been little discussion about glass-box/transparent ML types which can not only enable quantitative predictions but also support qualitative judgement through interpretable insights that can deepen managers understanding of the cause-and-effect relationships that exist between injury precursors incidences. What is also not yet clear in construction safety prediction research is to what extent glass-box models are comparable to black-box counterparts in terms of predictive accuracy—a discussion which can facilitate rationale for interpretability/performance trade-off and model selection criteria. Furthermore, little is known about the potential of integrated ML approaches (e.g., multiple model ensembles) within the construction research area. Such approaches combine the learning strengths of individual models and compensate for their weaknesses by diminishing the impact of a single source of error—which contributes to better model robustness and stability (Sagi and Rokach 2018). As individual models are better than others in dealing with particular facets within complex systems, an ensemble approach should, in theory, bear high potential in a complex domain such as site safety risk prediction.

Over the past two decades, CDNT has elicited increasing attention in construction project management research as a response to the emerging perspective of viewing projects as network-based organizations (Zheng et al. 2016). As such, CDNT has been mainly used to study stakeholders' (nodes) patterns of interactions and interrelationships (links) during project-based collaborations. An example of links in the construction literature is the frequency of communications between the project development teams (Mead 2001; Chinowsky et al. 2011; Dogan et al. 2015; Castillo et al. 2018) and the channels of information exchange/knowledge flow that exist when project-related issues arise (Thorpe and Mead 2001; Doloi et al. 2016; Schröpfer et al. 2017; Xue et al., 2018). However, there has been little attention on studying the interdependence of technical crews such as on-site contractor groups that also have a significant direct impact on project performance, and how such interdependencies dynamically evolve over time. CDNT has also been employed to study stakeholder collaborations at the intra-organizational cross-project level (Pryke 2004; Chinowsky et al. 2018; Di Marco et al. 2010; Pauget and Wald 2013; Li et al. 2019) and at the wider inter-organizational scale (Park et al. 2011; Ruan et al. 2012; Solis et al. 2013; Lu et al. 2020). Such approaches have been useful in understanding important stakeholders and the interdependencies they induce on their local or global networks; however, investigating the effects of such interdependence-induced vulnerabilities on overall project performance is still needed.

1.2 RESEARCH OBJECTIVES

The main goal of this work is to develop data-driven strategies that can support project managers and stakeholders with resilience-guided approaches to complex infrastructure project management and empower their ability to mitigate systemic risks at both the intra-KPI and inter-KPI levels of their projects. In fulfillment of the stated goal, the following specific objectives were envisioned:

- Build on previous research to identify key factors affecting KPIs, such as delay factors or injury precursors, and identify or collect objective data from previous projects that describe qualitative or quantitative assessments of such factors with regards to influencing relevant KPIs.
- Develop ML-based approaches/frameworks to predict potential disruptions to different KPIs, based on the existence of sets of factors, such that proactive response/mitigation strategies and/or contingencies can be deployed ahead of time.
- Develop CDNT-based approaches/frameworks to model the interconnectedness between project stakeholders and between different KPIs, and facilitate adaptive solutions against potential interdependence-induced cascade in KPI disruptions, such that rapid restoration of the most important set of project performance targets can be achieved in cases of occurrence of disruptive events.

1.3 ORGANIZATION OF THE DISSERTATION

This dissertation comprises six chapters (see Figure 1.5b):

- Chapter 1 presents the motivation and objectives of the dissertation as well as background information pertaining to the research program.
- Chapter 2 discusses the development of a ML-based tool that learns from previous construction project data to facilitate predictions of future project durations based on these projects' risk factors. Key delay factors are identified from the literature and adapted by industry professionals, and subsequently, a relevant historical project dataset is compiled and explored to uncover key interdependencies between such factors. ML-based schedule delay predictive models are then developed and validated, such that they can be utilized by project managers to facilitate accurate forecasts of future and ongoing projects schedules, thus supporting their proactive project risk management strategies.
- Chapter 3 discusses the development of a ML-based construction safety decision support framework that enables quantitative construction injury prediction, while also supporting qualitative safety judgment and interpretation. Both glass- and black-box type ML models are developed and validated to predict injury risk levels across different construction sites based on the existence of key injury factors. Guidance is then provided on selecting between glass-box model interpretability and black-box model possible higher

performance based on different site characteristics. An application using an injury cases dataset is provided to demonstrate the framework utility and other key managerial insights. Using the framework, safety managers can evaluate different sites' safety risk levels and classify them to potentially high-risk or low-risk zones for ultimately formulating and disseminating better-informed and targeted prevention strategies.

- Chapter 4 discusses the development of site risk models that can generate predictions of injury outcomes and safety risk leading indicators across different site zones and over project lifecycles, thus empowering a proactive and real-time approach to construction safety management. In this respect, ensemble ML algorithms, which drive the generation of predictions within these models, are trained and validated to learn from previous injury precursors and outcomes. A demonstration application is then considered, where the ensemble algorithms are employed to develop a risk model of a construction site which generates forward-looking forecasts of safety risk leading indicators, such as injuries' financial implications and body parts most likely affected, across various zones and over different timeframes which can support key proactive and zone-specific injury-preventive decision making.
- Chapter 5 discusses the development of a CDNT-based resilience management framework that can enhance projects' ability to rapidly overcome possible cascades in KPI disruptions. The chapter focuses on: 1)

modelling the complex dynamic interdependencies of project contractors using CDNT approaches and correlating such networks with KPI variations; 2) analyzing the behaviors of these networks and proactively assessing possible cascades in KPI disruptions; and 3) facilitating dynamically adaptive (self-organized) network recovery responses to support project managers with proactive and objective response plans (e.g., re-coordinated schedules) to ensure project resilience against such KPI disruptions. A large-scale infrastructure project is used to demonstrate the implementation of the developed framework which revealed several insights that further both the comprehensive and granular understandings of the project dynamics in terms of key contractor influences, their interdependence-induced vulnerabilities and challenging/critical work packages and months.

- Chapter 6 presents the dissertation summary, major conclusions and recommendations for future research.

It should be noted that although Chapters 2 through 5 represent standalone manuscripts that are already published or submitted as journal articles, these chapters collectively describe a cohesive research program as outlined in this introductory chapter of the dissertation. Nonetheless, some overlap might exist for the completeness of the standalone chapters/manuscripts.

1.4 REFERENCES

Abu Hammad, A. A., Ali, S. M. A., Sweis, G. J., & Sweis, R. J. (2010). Statistical analysis on the cost and duration of public building projects. Journal of Management in Engineering, 26(2), 105-112.

Abubakar, A. M., Karadal, H., Bayighomog, S. W., & Merdan, E. (2018). Workplace injuries, safety climate and behaviors: application of an artificial neural network. International journal of occupational safety and ergonomics.

Aggarwal, C. C. (2016). Data mining: the textbook, 285-426. Springer.

Ahiaga-Dagbui, D. D., & Smith, S. D. (2014). Dealing with construction cost overruns using data mining. Construction management and economics, 32(7-8), 682-694.

Avlijaš, G. (2019). Examining the value of Monte Carlo Simulation for project time management. Management: Journal of Sustainable Business and Management Solutions in Emerging Economies, 24(1), 11-23.

Ayhan, B. U., & Tokdemir, O. B. (2020). Accident analysis for construction safety using latent class clustering and artificial neural networks. Journal of Construction Engineering and Management, 146(3), 04019114.

Barker, K., Ramirez-Marquez, J. E., & Rocco, C. M. (2013). Resilience-based network component importance measures. Reliability Engineering & System Safety, 117, 89-97.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H.A., Alaka, H.A., & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced Engineering Informatics, 30(3), 500-521.

Cantarelli, C. C., van Wee, B., Molin, E. J., & Flyvbjerg, B. (2012). Different cost performance: different determinants?: The case of cost overruns in Dutch transport infrastructure projects. Transport Policy, 22, 88-95.

Castillo, T., Alarcón, L. F., & Pellicer Armiñana, E. (2018). Influence of organizational characteristics on construction project performance using corporate social networks. Journal of Management in Engineering, 34(4), 1-9.

Chan, A. P. (2001). Time–cost relationship of public sector projects in Malaysia. International journal of project management, 19(4), 223-229.

Chan, A. P., & Chan, D. W. (2004). Developing a benchmark model for project construction time performance in Hong Kong. Building and environment, 39(3), 339-349.

Chao, L. C., & Chien, C. F. (2009). Estimating project S-curves using polynomial function and neural networks. Journal of Construction Engineering and Management, 135(3), 169-177.

Cheng, M. Y., Chang, Y. H., & Korir, D. (2019). Novel approach to estimating schedule to completion in construction projects using sequence and nonsequence learning. Journal of Construction Engineering and Management, 145(11), 04019072.

Cheng, M. Y., Tsai, H. C., & Liu, C. L. (2009). Artificial intelligence approaches to achieve strategic control over project cash flows. Automation in construction, 18(4), 386-393.

Chinowsky, P., Diekmann, J., & Galotti, V. (2008). Social network model of construction. Journal of construction engineering and management, 134(10), 804-812.

Chinowsky, P., Taylor, J. E., & Di Marco, M. (2011). Project network interdependency alignment: New approach to assessing project effectiveness. Journal of Management in Engineering, 27(3), 170-178.

Chou, J. S., & Lin, C. (2013). Predicting disputes in public-private partnership projects: Classification and ensemble models. Journal of Computing in Civil Engineering, 27(1), 51-60.

Chou, J. S., Tsai, C. F., & Lu, Y. H. (2013). Project dispute prediction by hybrid machine learning techniques. Journal of Civil Engineering and Management, 19(4), 505-517.

Construct Connect (2021). Delayed Project Reports. Retrieved August 3, 2021, from https://www.constructconnect.com/delayed-projects-report

Desai, V. S., & Joshi, S. (2010). Application of decision tree technique to analyze construction project data. Proceedings of International Conference on Information Systems, Technology and Management, 304-313. Springer Berlin Heidelberg.

Di Maddaloni, F., & Davis, K. (2018). Project manager's perception of the local communities' stakeholder in megaprojects. An empirical investigation in the UK. International Journal of Project Management, 36(3), 542-565.

Di Marco, M. K., Taylor, J. E., & Alin, P. (2010). Emergence and role of cultural boundary spanners in global engineering project networks. Journal of management in engineering, 26(3), 123-132.

Dogan, S. Z., Arditi, D., Gunhan, S., & Erbasaranoglu, B. (2015). Assessing coordination performance based on centrality in an e-mail communication network. Journal of Management in Engineering, 31(3), 04014047.

Doloi, H. K., Loganathan, S., Kalidindi, S. N., & Mahalingam, A. (2016). Assessment of stakeholders' management practice in infrastructure projects—an Indian case project. In Construction Research Congress 2016, 1465-1474.

Dursun, O., & Stoy, C. (2011). Time–cost relationship of building projects: statistical adequacy of categorization with respect to project location. Construction Management and Economics, 29(1), 97-106.

Elazouni, A. M. (2006). Classifying construction contractors using unsupervised-learning neural networks. Journal of construction engineering and management, 132(12), 1242-1253.

Eriksson, P. E., Larsson, J., & Pesämaa, O. (2017). Managing complex projects in the infrastructure sector—A structural equation model for flexibility-focused project management. International journal of project management, 35(8), 1512-1523.

Fan, Ching-Lung. "Defect risk assessment using a hybrid machine learning method." Journal of Construction Engineering and Management 146, no. 9 (2020): 04020102.

Fan, H., AbouRizk, S., Kim, H., & Zaïane, O. (2008). Assessing residual value of heavy construction equipment using predictive data mining model. Journal of Computing in Civil Engineering, 22(3), 181-191.

Feng, Y., Teo, E. A. L., Ling, F. Y. Y., & Low, S. P. (2014). Exploring the interactive effects of safety investments, safety culture and project hazard on safety performance: An empirical analysis. International Journal of Project Management, 32(6), 932-943.

Flyvbjerg, B. (2014). What you should know about megaprojects and why: An overview. Project management journal, 45(2), 6-19.

Fu, T., Lyu, Y., Liu, H., Peng, R., Zhang, X., Ye, M., & Tan, W. (2018). DNA-based dynamic reaction networks. Trends in biochemical sciences, 43(7), 547-560.

García de Soto, B., Agustí-Juan, I., Joss, S., & Hunhevicz, J. (2019). Implications of Construction 4.0 to the workforce and organizational structures. International Journal of Construction Management, 1-13.

Gerassis, S., Martín, J. E., García, J. T., Saavedra, A., & Taboada, J. (2017). Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. Journal of construction engineering and management, 143(2), 04016093.

Gong, J., Caldas, C. H., & Gordon, C. (2011). Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. Advanced Engineering Informatics, 25(4), 771-782.

Gong, M., Cai, Q., Ma, L., Wang, S., & Lei, Y. (2017). Computational Intelligence for Network Structure Analytics. Springer Singapore. ISBN 978-9811045578

Hammad, A. A. A., Ali, S. A., Sweis, G. J., & Bashir, A. (2008). Prediction model for construction cost and duration in Jordan. Jordan Journal of Civil Engineering, 2(3), 250-266.

Han, S. H., Yun, S., Kim, H., Kwak, Y. H., Park, H. K., & Lee, S. H. (2009). Analyzing schedule delay of mega project: Lessons learned from Korea train express. IEEE Transactions on Engineering Management, 56(2), 243-256.

Hariri-Ardebili, M. A. (2018). Risk, Reliability, Resilience (R3) and beyond in dam engineering: A state-of-the-art review. International journal of disaster risk reduction, 31, 806-831.

Hashem M. Mehany, M. S., Bashettiyavar, G., Esmaeili, B., & Gad, G. (2018). Claims and project performance between traditional and alternative project delivery methods. Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 10(3), 04518017.

Heravi, G., & Eslamdoost, E. (2015). Applying artificial neural networks for measuring and predicting construction-labor productivity. Journal of Construction Engineering and Management, 141(10), 04015032.

Hernandez-Fajardo, I., & Dueñas-Osorio, L. (2013). Probabilistic study of cascading failures in complex interdependent lifeline systems. Reliability Engineering & System Safety, 111, 260-272.

Jarkas, A. M. (2017). Contractors' perspective of construction project complexity: definitions, principles, and relevant contributors. Journal of Professional Issues in Engineering Education and Practice, 143(4), 04017007.

Kang, K., & Ryu, H. (2019). Predicting types of occupational accidents at construction sites in Korea using random forest model. Safety Science, 120, 226-236.

Kog, F., & Yaman, H. (2014). A meta classification and analysis of contractor selection and prequalification. Procedia Engineering, 85, 302-310.

Kokkaew, N., & Wipulanusat, W. (2014). Completion delay risk management: A dynamic risk insurance approach. KSCE Journal of Civil Engineering, 18(6), 1599-1608.

Larsen, J. K., Shen, G. Q., Lindhard, S. M., & Brunoe, T. D. (2015). Factors affecting schedule delay, cost overrun, and quality level in public construction projects. Journal of Management in Engineering, 32(1), 04015032.

Li, X., Li, H., Cao, D., Tang, Y., Luo, X., & Wang, G. (2019). Modeling dynamics of project-based collaborative networks for BIM implementation in the construction industry: empirical study in Hong Kong. Journal of Construction Engineering and Management, 145(12), 05019013.

Love, P. E., Teo, P., Morrison, J., & Grove, M. (2016). Quality and safety in construction: Creating a no-harm environment. Journal of Construction Engineering and Management, 142(8), 05016006.

Lu, W., Xu, J., & Söderlund, J. (2020). Exploring the effects of building information modeling on projects: Longitudinal social network analysis. Journal of Construction Engineering and Management, 146(5), 04020037.

Luo, L., He, Q., Xie, J., Yang, D., & Wu, G. (2016). Investigating the relationship between project complexity and success in complex construction projects. Journal of Management in Engineering, 33(2), 04016036.

Mahfouz, T., & Kandil, A. (2012). Litigation outcome prediction of differing site condition disputes through machine learning models. Journal of Computing in Civil Engineering, 26(3), 298-308.

Makki, A. A., & Mosly, I. (2021). Predicting the Safety Climate in Construction Sites of Saudi Arabia: A Bootstrapped Multiple Ordinal Logistic Regression Modeling Approach. Applied Sciences, 11(4), 1474.

McKinsey Global Institute (2017). Bridging infrastructure gaps: Has the world made progress?. McKinsey & Company. Retrieved April 19, 2020, from https://www.mckinsey.com/business-functions/operations/our-insights/bridging-infrastructure-gaps-has-the-world-made-progress#.

Mead, S. P. (2001). Using social network analysis to visualize project teams. Project Management Journal, 32(4), 32-38.

Naji, H., Ibrahim, A. M., & Hassan, Z. (2018). Evaluation of Legislation Adequacy in Managing Time and Quality Performance in Iraqi Construction Projects-a Bayesian Decision Tree Approach. Civil Engineering Journal, 4(5), 993-1005.

Ndekugri, I., Braimah, N., & Gameson, R. (2008). Delay analysis within construction contracting organizations. Journal of construction engineering and management, 134(9), 692-700.

Nepal, M. P., Park, M., & Son, B. (2006). Effects of schedule pressure on construction performance. Journal of Construction Engineering and Management, 132(2), 182-188.

Park, H., Han, S. H., Rojas, E. M., Son, J., & Jung, W. (2011). Social network analysis of collaborative ventures for overseas construction projects. Journal of construction engineering and management, 137(5), 344-355.

Patel, D. A., & Jha, K. N. (2015). Neural network approach for safety climate prediction. Journal of management in engineering, 31(6), 05014027.

Pauget, B., & Wald, A. (2013). Relational competence in complex temporary organizations: The case of a French hospital construction project network. International Journal of Project Management, 31(2), 200-211.

Petruvesa, S., Zileska, V., & Zujo, V. (2013). Predicting construction project duration with support vector machine. International Journal of research in Engineering and Technology, 11(2), 12-24.

Poh, C. Q., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. Automation in construction, 93, 375-386.

Pryke, S. D. (2004). Analysing construction project coalitions: exploring the application of social network analysis. Construction management and economics, 22(8), 787-797.

Rashid, K., Li, X., Deng, J., Xie, Y., Wang, Y., & Chen, S. (2019). Experimental and analytical study on the flexural performance of CFRP-strengthened RC beams at various pre-stressing levels. Composite Structures, 227, 111323.

Rezaie, K., Amalnik, M. S., Gereie, A., Ostadi, B., & Shakhseniaee, M. (2007). Using extended Monte Carlo simulation method for the improvement of risk management: Consideration of relationships between uncertainties. Applied Mathematics and Computation, 190(2), 1492-1501.

Ruan, X., Ochieng, E. G., Price, A. D., & Egbu, C. O. (2012). Knowledge integration process in construction projects: a social network analysis approach to compare competitive and collaborative working. Construction management and economics, 30(1), 5-19.

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(4), e1249.

Sakhakarmi, S., Park, J., & Cho, C. (2019). Enhanced machine learning classification accuracy for scaffolding safety using increased features. Journal of construction engineering and management, 145(2), 04018133.

Schröpfer, V. L. M., Tah, J., & Kurul, E. (2017). Mapping the knowledge flow in sustainable construction project teams using social network analysis. Engineering, Construction and Architectural Management.

Serrador, P., & Turner, R. (2015). The relationship between project success and project efficiency. Project management journal, 46(1), 30-39.

Shi, H. (2009, October). Application of Unascertained Method and Neural Networks to Quality Assessment of Construction Project. In 2009 Second International Conference on Intelligent Computation Technology and Automation (Vol. 1, 52-55). IEEE.

Sobhani, F., & Madadi, T. (2015). Studying the suitability of different data mining methods for delay analysis in construction projects. Journal of applied research on industrial engineering, 2(1), 15-33.

Solis, F., Sinfield, J. V., & Abraham, D. M. (2013). Hybrid approach to the study of inter-organization high performance teams. Journal of construction engineering and management, 139(4), 379-392.

Son, H., Kim, C., & Kim, C. (2012). Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables. Automation in Construction, 27, 60-66.

Statistics Canada (2021). Labour force characteristics by industry annual (x1,000). Retrieved August 4, 2021 from https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1410002301

Sun, J., & Zhang, P. (2011). Owner organization design for mega industrial construction projects. International Journal of Project Management, 29(7), 828-833.

The Association of Workers' Compensation Boards of Canada (AWCBC) (2021). National Work Injury, Disease and Fatality Statistics. Retrieved August 4, 2021 from https://awcbc.org/en/statistics/

Thorpe, T., & Mead, S. (2001). Project-specific web sites: Friend or foe?. Journal of construction engineering and management, 127(5), 406-413.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Application of machine learning to construction injury prediction. Automation in construction, 69, 102-114.

Tokdemir, O. B., Erol, H., & Dikmen, I. (2019). Delay risk assessment of repetitive construction projects using line-of-balance scheduling and Monte Carlo simulation. Journal of Construction Engineering and Management, 145(2), 04018132.

Wauters, M., & Vanhoucke, M. (2016). A comparative study of Artificial Intelligence methods for project duration forecasting. Expert systems with applications, 46, 249-261.

Wilkinson, S., Chang-Richards, A. Y., Sapeciay, Z., & Costello, S. B. (2016). Improving construction sector resilience. International Journal of Disaster Resilience in the Built Environment, 7(2), 173-185.

Williams, T. P., & Gong, J. (2014). Predicting construction cost overruns using text mining, numerical data and ensemble classifiers. Automation in Construction, 43, 23-29.

Xue, X., Zhang, R., Wang, L., Fan, H., Yang, R. J., & Dai, J. (2018). Collaborative innovation in construction project: A social network perspective. KSCE Journal of Civil Engineering, 22(2), 417-427.

Yan, H., Yang, N., Peng, Y., & Ren, Y. (2020). Data mining in the construction industry: Present status, opportunities, and future trends. Automation in Construction, 119, 103331.

Yates, J. K., & Epstein, A. (2006). Avoiding and minimizing construction delay claim disputes in relational contracting. Journal of Professional Issues in Engineering Education and Practice, 132(2), 168-179.

Yeo, K. T. (1995). Planning and learning in major infrastructure development: systems perspectives. International Journal of Project Management, 13(5), 287-293.

Zheng, X., Le, Y., Chan, A. P., Hu, Y., & Li, Y. (2016). Review of the application of social network analysis (SNA) in construction project management research. International journal of project management, 34(7), 1214-1225.

Zhou, Y., Su, W., Ding, L., Luo, H., & Love, P. E. (2017). Predicting safety risks in deep foundation pits in subway infrastructure projects: support vector machine approach. Journal of Computing in Civil Engineering, 31(5), 04017052.

Chapter 2:

MACHINE LEARNING ALGORITHMS FOR CONSTRUCTION PROJECTS DELAY RISK PREDICTION

ABSTRACT

Projects delays are among the most pressing challenges faced by the construction sector attributed to the sector's complexity and its inherent delay risk sources' interdependence. Machine learning offers an ideal set of techniques capable of tackling such complex systems, however adopting such techniques within the construction sector remains at an early stage. The aim of this study is to identify and develop machine learning models in order to facilitate accurate project delay risk analysis and prediction using objective data sources. As such, relevant delay risk sources and factors were first identified, and a multivariate dataset of previous projects' time performance and delay-inducing risk sources was then compiled. Subsequently, the complexity and interdependence of the system were uncovered through an exploratory data analysis. Accordingly, two suitable machine learning models, utilizing decision tree and naïve Bayesian classification algorithms, were identified and trained using the dataset for predicting project delay extents. Finally, the predictive performances of both models were evaluated through cross validation tests and the models were further compared using machine-learning-relevant performance indices. The evaluation

results indicated that the naïve Bayesian model provides better predictive performance for the dataset examined. Ultimately, the work presented herein harnesses the power of machine learning to facilitate evidence-based decision making, while inherent risk factors are active, interdependent and dynamic, thus empowering proactive project risk management strategies.

2.1 INTRODUCTION

Construction project delay is a global phenomenon (Assaf and Al-Hejji, 2006; Sambasivan and Soon, 2006). Its occurrence is mainly attributed to the interdependent inherent risk factors and uncertainties associated with the complex and dynamic nature of construction processes. Such factors might for example be related to stakeholder(s) incompetence, poor communication, inadequate estimation of employed resources, contractual deviations, or even municipal constraints (Assaf and Al-Hejji, 2006; Aziz, 2013). As a result, project delays, quantified through time overrun (**TO**), can negatively impact the project and its stakeholders in multiple ways including: a) claims, disputes and arbitration; b) cost overruns and loss of revenues; c) disruption of work and loss of productivity; or d) contract termination and, possibly, total project abandonment (Aibinu and Jagboro, 2002; Majid, 2006).

In order to provide accurate estimates of project durations, construction firms typically adopt standard quantitative delay risk analysis tools. For example, Monte Carlo analysis can be used for investigating the complete extent of risk associated with scheduled work items to estimate more reasonable project completion dates (Rezaie et al., 2007; Sadeghi et al., 2010; Kokkaew and Wipulanusat, 2014). Similar to most probabilistic modelling tools, Monte Carlo analysis is data-intensive and requires estimates of work item duration ranges and relevant probability distributions. These estimates can be acquired either *objectively*, through historical data of similar projects, or, more commonly in the absence of such data, *subjectively* based on expert opinion and judgement. The latter approach, however, poses several key limitations related to: 1) using imprecise and/or ambiguous data sources that would typically add another layer of uncertainty to the analysis; 2) overlooking the complex and interdependent nature of inherent risk factors in construction projects which might significantly influence original predictions; and 3) using data that is rarely updated to represent the actual project progressions since the subjective data collected are relevant only at the time of apprehension (Ferson, 1996; Guyonnet et al., 2003; Goldstein, 2006; Tixier et al., 2017). To overcome these limitations, it is critical to base construction schedule estimates and planning decisions on knowledge extracted from objective and factual data.

In recent times, the construction sector has experienced an explosive growth in the amount of such objective data generated on a daily basis and stored from the various disciplines throughout the project or the facility lifespan (Bilal et al., 2016). Such provision of data creates an opportunity for extracting useful corporate knowledge and promising solutions to the prevailing project delay dilemma. However, because of the interdependence of construction-related data, the adoption of effective data analytics tools is key. In this respect, the potential of machine learning (**ML**) techniques and algorithms in analyzing voluminous, complex and interdependent datasets of varying structures for deriving useful insights cannot be overemphasized (Kim et al., 2008; Bilal et al., 2016). ML algorithms, in their different forms, have been widely employed in different fields over the past two decades. Nonetheless, ML remains a new prospect within the construction sector despite its highly regarded advantageous potential. A literature survey by the authors of the present study showed that only a limited number of studies have focused on applications of ML techniques within construction research in general, and to an even much lesser extent on delay risk analysis. A non-comprehensive list of ML techniques applied in construction-related disciplines includes artificial neural networks (Elazouni, 2006; Chao and Chien, 2009; Heravi and Eslamdoost, 2015), decision trees (Desai and Joshi, 2010; Chi et al., 2012; Chou and Lin, 2013), naïve Bayesian models (Jiang and Mahadevan, 2008; Gong et al., 2011; Gerassis et al., 2016), and support vector machines (Cheng and Wu, 2009; Lam et al., 2009; Huang and Tserng, 2018).

The *goal* of the present study is to identify and develop an efficient predictive data analytics tool to analyze and learn from objective delay risk sources based on previous construction project data. Achieving this aim will ultimately facilitate more accurate predictions of future project durations based on these projects' inherent and expected risk levels, thus supporting a proactive project risk management strategy.

In fulfillment of the stated research aim, the present study is focused on achieving two key objectives (see Figure 2.1). The *first objective* is to identify relevant delay risk sources and factors extracted from literature and adapted by the industry, and subsequently compile and understand a relevant historical construction project delay risk dataset, as similar data is currently not available in open literature to the best of the authors' knowledge. Within the process of data collection, constraints pertaining to data characteristics, project types, and analysis limitations were set. These constraints and limitations ensured the consistency and homogeneity of the input data to the predictive data analytics tool in order to realize meaningful results. Afterwards, and from various unstructured historical construction project data formats, data pertaining to different types of delay risk sources, along with their level of contribution to project delay, were extracted and subsequently pre-processed to constitute a structured, consistent and multivariate dataset ideally suited for predictive analytics. Subsequently, an exploratory data analysis was conducted to explore the dataset properties and the complex interdependencies between the risk sources were uncovered.

Based on the complexity and interdependence of construction delay risk sources, a ML approach was deemed the most appropriate to tackle such a challenging system of interacting variables, and the rationale behind this selection was reported. As such, the *second objective* of the present study is to identify, develop and validate appropriate ML algorithms to analyze the compiled previous

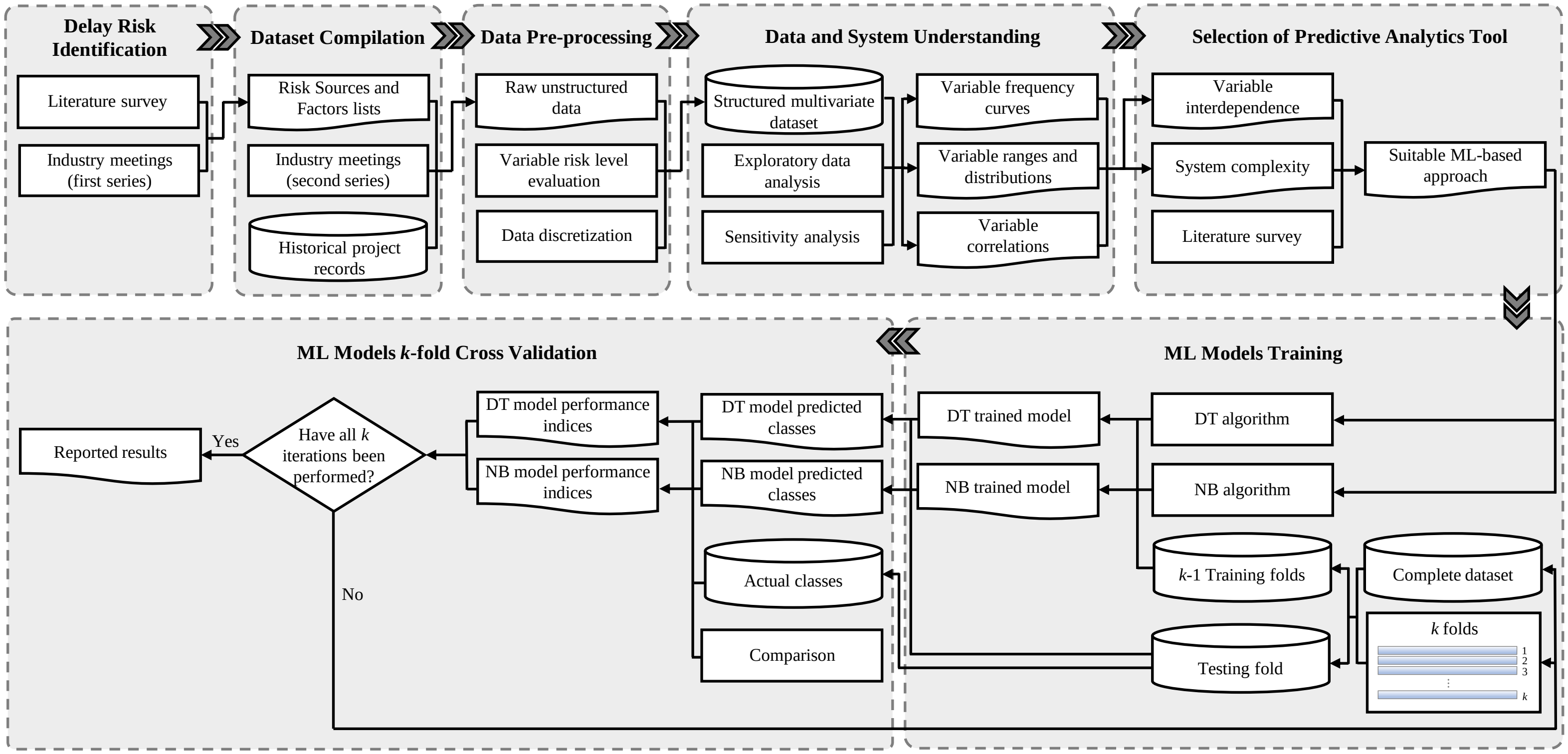

Delay Risk Identification
Literature survey
Industry meetings (first series)
Dataset Compilation
Risk Sources and Factors lists
Industry meetings (second series)
Historical project records
Data Pre-processing
Raw unstructured data
Variable risk level evaluation
Data discretization
Data and System Understanding
Structured multivariate dataset
Exploratory data analysis
Sensitivity analysis
Variable frequency curves
Variable ranges and distributions
Variable correlations
Selection of Predictive Analytics Tool
Variable interdependence
System complexity
Literature survey
Suitable ML-based approach
ML Models Training
DT algorithm
NB algorithm
DT trained model
NB trained model
k-1 Training folds
Complete dataset
k folds
1
2
3
k
Testing fold
ML Models k-fold Cross Validation
DT model predicted classes
NB model predicted classes
Actual classes
Comparison
DT model performance indices
NB model performance indices
Have all k iterations been performed?
Yes
No
Reported results

Figure 2.1: Analysis methodology and objectives

project dataset. Appreciating the dataset size and properties and prior to conducting the ML-based analysis, a review of previous ML applications was conducted in order to identify the ML technique and algorithms best suited to the dataset considered. Subsequently, the study focuses on applying a supervised learning classification technique through two ML algorithms: decision tree and naïve Bayesian classification algorithms, that were found to be ideal considering the compiled dataset properties. These two algorithms were employed to analyze the degrees of variabilities of the influencing risk sources (the independent variables) and their effect on the extents of TO (the dependent variable) described as class labels, in order to generate two TO predictive models/classifiers. Finally, validation of the two generated models' predictive performances was conducted, in terms of both training and testing, through comparisons of predicted and actual class outcomes; and the two models were evaluated and compared using confusion matrices and multiple performance indices.

Figure 2.1 shows the methodology adopted to attain the study's objectives, and thus its aim, described above. The following sections explain the different steps outlined in Figure 2.1 followed by concluding remarks, inferred findings and future recommendations to reach the research long-term goals.

2.2 DELAY RISK FACTOR AND SOURCE IDENTIFICATION

Prior to identifying the type of data to be collected, a literature survey was first conducted in order to identify the most common risk factors influencing building construction project delay and to group such factors into different source categories. This literature survey was conducted in three tiers.

First, a broad search was carried out to identify articles containing the terms: ("delay" OR "time delay" OR "time overrun" OR "delay risk" OR "delay factor" OR "delay source" OR "delay cause") AND ("construction" OR "construction project"). This search was conducted through two main sources: 1) academic literature databases including the *American Society of Civil Engineers Library*, *Elsevier Science Direct Digital Library*, *Springer*, *Taylor & Francis Online* and *Emerald Insight*; and 2) academic literature search engines as *Web of Science*, *EBSCOhost* and *Google Scholar*. The temporal range of the search was set to cover the period from 1980 to 2018, since construction project delay and the factors influencing it have been attracting increased attention for the past three to four decades. By the end of the first tier of literature survey, a total of 83 articles were identified.

In the second tier, the search was narrowed down, whereas the titles, abstracts and keywords of the articles identified from the first tier were reviewed in order to select and retain those articles of relevance to the research scope for a full review. Specifically, the following criteria were set for selecting an article for a full review: 1) peer-reviewed articles published in refereed journals of project management, construction management, built environment or construction economics; and 2) articles focusing on the causes of delays in building construction projects and/or the quantitative assessment of these causes on influencing project delays. This screening process resulted in the selection of 34 relevant research articles from the following journals: *Construction Management and Economics*, *International Journal of Construction Management*, *Journal of Construction Engineering and Management*, *Journal of Management in Engineering*, *International Journal of Project Management*, *Automation in Construction* and *Construction Economics and Building*.

The third tier involved an in-depth review of the 34 selected articles from the second tier in order to examine previous studies' identifications of individual delay risk factors and their assemblies into main delay risk sources. Different authors focused on selected risk sources within their articles and identified lists of individual risk factors within such sources. After reviewing the 34 articles, the results of ten articles were used for risk factor and source identification as the former were repeatedly cited by the rest of the 34 articles. Beyond these ten articles, no distinct risk factors were identified within any risk source. By the end of the three-tier literature survey, nine delay risk sources were established and subsequently adopted in the present study.

These delay risk sources relate to: 1) owner; 2) consultant; 3) contractor; 4) design; 5) labor; 6) material; 7) equipment; 8) project; and 9) external aspects, and cover all possible sources of delay-inducing risk factors. Table 2.1 shows the different subsets of the ten articles on which all the nine risk sources, and their constituting factors, were based.

Inevitably, variations existed in the individual delay risk factor lists identified by the reviewed articles within the different risk source categories. This was attributed to dissimilarities within the different articles in terms of construction environments, geographical conditions, political situations, construction methods, resource availabilities and stakeholder engagements. As such, a *first series* of meetings were held with 15 construction experts to confirm the relevance of the identified risk factors to the construction sector and modify them as necessary. Based on the literature survey and the expert meetings, 59 delay risk factors were identified by the present study and a full listing of these factors within their source categories is presented in Table 2.2.

2.3 DATASET COMPILATION AND PRE-PROCESSING

The dataset used to develop the proposed predictive analytics tool included data from 51 construction projects, from 28 firms, that experienced varying degrees of time delay. Each data record in the dataset represents a specific project

Table 2.1: Previous studies from which risk sources and comprising risk factors were identified

Risk source	Relevant study
1. Owner	Arditi et al., 1985; Chan & Kumaraswamy, 1997; Mezher & Tawil, 1998; Al Momani, 2000; Odeh & Battaineh, 2002; Assaf & Al-Hejji, 2006; Sambasivan & Soon, 2006; Fugar & Agyakwah-Baah, 2010
2. Consultant	Al Momani, 2000; Odeh & Battaineh, 2002; Assaf & Al-Hejji, 2006; Aziz, 2013
3. Contractor	Arditi et al., 1985; Chan & Kumaraswamy, 1997; Mezher & Tawil, 1998; Sambasivan & Soon, 2006; Fugar & Agyakwah-Baah, 2010
4. Design	Arditi et al., 1985; Chan & Kumaraswamy, 1997; Assaf & Al-Hejji, 2006; Aziz, 2013
5. Labor	Al Momani, 2000; Assaf & Al-Hejji, 2006; Sambasivan & Soon, 2006
6. Materials	Mansfield et al., 1994; Al Momani, 2000; Sambasivan & Soon, 2006; Fugar & Agyakwah-Baah, 2010
7. Equipment	Mansfield et al., 1994; Assaf & Al-Hejji, 2006; Sambasivan & Soon, 2006
8. Project	Mansfield et al., 1994; Chan & Kumaraswamy, 1997; Fugar & Agyakwah-Baah, 2010; Aziz, 2013
9. External	Chan & Kumaraswamy, 1997; Al Momani, 2000; Assaf & Al-Hejji, 2006

Table 2.2: Complete list of identified risk factors within respective risk sources

Risk source	Identified risk factors	
1. Owner	1.1	Inadequate project planning by owner
	1.2	Selecting inappropriate contractors
	1.3	Delays in site delivery to contractor
	1.4	Delays in reviewing and approving design documents
	1.5	Change orders by owner
	1.6	Slow decision-making process by owner
	1.7	Delays in progress payments by owner
	1.8	Suspension of work by owner
	1.9	Poor coordination by owner between consultant and contractor
	1.10	Conflicts between joint-ownership of the project
2. Consultant	2.1	Delays in reviewing and approving design documents
	2.2	Delays in performing inspection and testing
	2.3	Delays in approving major changes in scope of work by consultant
	2.4	Inadequate consultant experience
	2.5	Poor consultant communication with contractor and owner
	2.6	Conflicts between consultant and design engineer
3. Contractor	3.1	Ineffective project planning by contractor
	3.2	Difficulties in financing project by contractor
	3.3	Incompetence or inexperience of contractor
	3.4	Inadequate site investigation
	3.5	Slow site mobilization
	3.6	Poor site management and supervision
	3.7	Delays due to unreliable subcontractors' work
	3.8	Frequent change of subcontractors
	3.9	Rework due to errors during construction
	3.10	Poor contractor communication with consultant and owner
	3.11	Conflicts between contractor and consultant and/or owner
4. Design	4.1	Inadequate design team experience
	4.2	Misunderstanding of owner's requirements by design engineer

	4.3	Delays in producing design documents
	4.4	Design errors/incomplete or unclear design drawings
5. Labor	5.1	Shortage of labor
	5.2	Unqualified or inadequate workforce
	5.3	Low productivity of labor
	5.4	Personal conflicts among labor
6. Materials	6.1	Shortage of construction materials in market
	6.2	Delays in delivery of materials
	6.3	Inadequate quality of materials
	6.4	Damage of sorted materials
	6.5	Changes in material types and specifications during construction
7. Equipment	7.1	Shortage of equipment
	7.2	Slow mobilization of equipment
	7.3	Low productivity and efficiency of equipment
	7.4	Frequent equipment breakdowns
	7.5	Improper equipment or lack of high-tech equipment
8. Project	8.1	Unsuitable type of project bidding and award (e.g. negotiation, lowest bidder, etc.)
	8.2	Mistakes or discrepancies in contract documents
	8.3	Original contract duration is too short
	8.4	Ineffective delay penalties
	8.5	Lack of communication between project parties
	8.6	Legal disputes between project participants
9. External	9.1	Delays in obtaining permits from municipality
	9.2	Changes in government regulations and laws
	9.3	Delays in providing services from utilities (e.g. water, electricity, telephones, etc.)
	9.4	Unexpected surface & subsurface conditions (e.g. soil, water table, etc.)
	9.5	Problems with neighbors
	9.6	Unfavorable weather conditions
	9.7	Accidents during construction
	9.8	Price fluctuations

and is linked to ten data variables. The first of these variables represents the dependent variable which is the extent of TO sustained by the project. The other nine variables, reflecting the nine different delay risk sources above, are considered independent variables. Table 2.3 shows a sample of the described dataset after all the operations pertaining to data collection and pre-processing (discussed next) were performed.

To compile the described dataset, a *second series* of meetings were arranged. In total, 112 meetings were held across the 28 contacted firms over the span of nine months to extract data from the 51 completed projects. The preliminary meeting with each firm involved explanations of the research aim and significance and the type of project data sought. Subsequently, one to two follow-up meetings were allocated to each project for collecting the necessary data. In these meetings, various project-related documents from the firm's historical records were investigated including contract documents, specifications, change orders, schedule baselines, monthly and quarterly updates, resource calendars, and risk registers. Based on these records and the knowledge of the risk factors constituting the risk sources, each risk source was assigned scores (index values) on two different index scales. The first of these scales relates to the consequence severity towards affecting the project time objective, and the second relates to the frequency of recurrence throughout the project, as shown in Tables 2.4 and 2.5, respectively. The overall risk source contribution values towards the TO are then

Table 2.3: Compiled dataset structure

Project ID	**Extent of TO**	**Risk source 1: Owner**	**Risk source 2: Consultant**	**…**	**Risk source 9: External**
Project 1	30-60% TO	Moderate	Very Low	…	High
Project 2	> 60% TO	Moderate	Very Low	…	Very High
Project 3	> 60% TO	Very High	Very Low	…	High
⋮	⋮	⋮	⋮	⋱	⋮
Project 50	30-60% TO	Very High	Very Low	…	Very Low
Project 51	> 60% TO	High	Very Low	…	High

Table 2.4: Consequence severity index scale

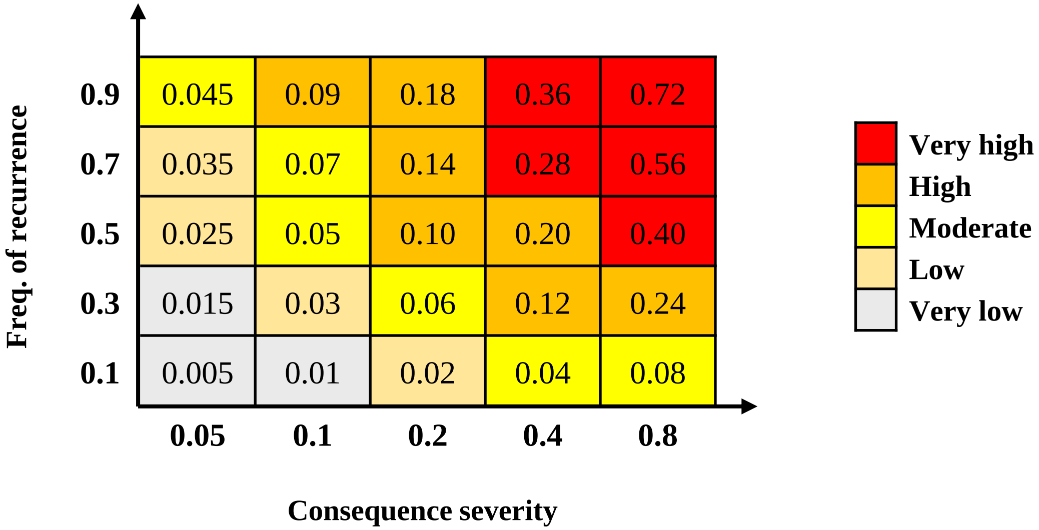
Index value	**Description**
0.05	Contributes to no or insignificant time overrun
0.1	Contributes to < 5% time overrun
0.2	Contributes to 5-10% time overrun
0.4	Contributes to 10-20% time overrun
0.8	Contributes to > 20% time overrun

Table 2.5: Frequency of recurrence (throughout project lifecycle) index scale

Index value	**Description**
0.1	Non-existing or very rare
0.3	Rare
0.5	Moderate
0.7	Frequent
0.9	Very frequent

numerically evaluated through multiplying the two corresponding scores (Assaf and Al-Hejji, 2006; Ismail et al., 2014; Xia et al., 2017). Afterwards, these numerical risk source contribution values were discretized into categorical risk source contribution levels based on the discretization matrix shown in Figure 2.2, where the values of the consequence severity are set along the horizontal axis, and those of the frequency of recurrence are set along the vertical axis. In this manner, each independent variable was classified into one of the following five categorical levels in terms of its contribution to TO: 1) Very low; 2) Low; 3) Moderate; 4) High; and 5) Very high. The two index scales and the discretization matrix were first adopted from literature (Assaf and Al-Hejji, 2006; Mahamid, 2011; Ismail et al., 2014; Kerzner, 2017; PMI, 2017; Xia et al., 2017), and then adapted based on input from the experts of the first series of meetings, and ultimately confirmed for adequacy to the projects investigated during the second series of meetings.

Furthermore, the dependent variable (i.e., TO) was also categorized into three class labels. One approach for such categorization is to divide the variable based on its frequency distribution into three *equal* class labels with each containing one-third of the variable count (i.e., exactly 17 project records). This approach would yield three class labels which are: 1) < 21% TO; 2) 21-27% TO; and 3) > 27% TO. Although this categorization approach reduces bias when implementing predictive data analytics, it is atypical for practical applications. It would thus be more beneficial to categorize TO in a way that provides greater



	0.05	0.1	0.2	0.4	0.8
0.9	0.045	0.09	0.18	0.36	0.72
0.7	0.035	0.07	0.14	0.28	0.56
0.5	0.025	0.05	0.10	0.20	0.40
0.3	0.015	0.03	0.06	0.12	0.24
0.1	0.005	0.01	0.02	0.04	0.08

Figure 2.2: Discretization matrix for assessing risk source contribution level towards time overrun

managerial insights and benefits. As such, the dependent variable was categorized into one of the following three class labels: 1) < 30% TO; 2) 30-60% TO; and 3) > 60% TO, which resulted in the class labels containing: 23, 16 and 12 project records, respectively. These three classes describe the extents of TO as: minor; moderate; and major, respectively, and were agreed on during the second series of meetings with construction firms.

2.4 DATASET CONSTRAINTS AND ANALYSIS CONSIDERATIONS

It should be noted that several constraints, considerations and limiting factors were enforced and/or encountered during data collection. First, the data gathered pertain to construction projects limited to those within Egypt and that are owned, financed, designed, built, managed and operated by national firms and entities. Second, the data collected were constrained to only building projects and do not concern other construction project categories (e.g., bridges) or disciplines. The third constraint pertains to project size, where data was only assembled from projects with contract values between 40 and 80 million Egyptian Pounds, and with contract durations within the range of 1 to 2 years. Fourth, data was only collected from projects that had incurred time overruns (i.e., TO > 0). Unifying the type of construction data within the compiled dataset was a key consideration for ensuring input consistency and homogeneity to the eventual predictive data analytics tool for attaining more reliable results, as mentioned earlier.

Furthermore, to ensure data integrity and quality, several carefully contemplated considerations were adopted. In this respect, only contractors with valid registrations in the Egyptian Federation for Construction and Building Contractors were considered. In addition, special attention was paid to consider only project cases having sufficient records and evidence to accurately infer credible and reliable data. Another important aspect was to avoid collecting data from projects with time intervals spanning across the period of the Egyptian revolution, which broke out on the 25th of January, 2011. Such projects would have experienced relatively extreme conditions and would disturb the consistency of risk levels among the wider apprehension of assembled projects.

These constraints and considerations resulted in including only 51 project data records for further analyses, which although constituted a valuable dataset of variables that is rarely reported or studied in literature, was nonetheless relatively small. As such, the selection of the predictive data analytics tool used in the present study was based on tools which performed well on small-sized datasets, as will be explained following the next section.

2.5 EXPLORATORY AND SENSITIVITY DATA ANALYSES

In order to better understand the system of delay-inducing risks within the construction sector, an exploratory data analysis, the results of which are displayed in Figure 2.3, was conducted to visually represent the properties of the collected data records and the correlations between different variables. The figure is a 10 × 10 matrix, in which the class labeled dependent variable (i.e., TO) and the nine independent variables (i.e., risk sources) are shown on the rows and columns. To facilitate its interpretation, the figure is divided into five blocks: the leftmost column – **Block 1**; the diagonal – **Block 2**; the uppermost row – **Block 3**; the lower-left triangle – **Block 4**; and the upper-right triangle – **Block 5**. The exploratory data analysis provides *three main insights*: 1) variable frequency counts and smoothed frequency curves; 2) a sensitivity analysis that explores the dependency of TO on the different risk sources; and 3) a sensitivity analysis that explores the dependencies of the different risk sources on one another.

Regarding the *first insight*, the boxes in **Block 1** and **Block 2** show the frequency counts and the smoothed frequency curves, respectively, of the data variables within each class. As mentioned earlier, the counts of projects incurring < 30% TO, 30-60% TO and > 60% TO were 23 (45%), 16 (31%) and 12 (24%), respectively. These percentages are consistent with the prevailing phenomenon of building projects experiencing less serious delays more frequently than more serious delays (Abd El-Razek et al., 2008; Singh, 2009). It can also be seen that

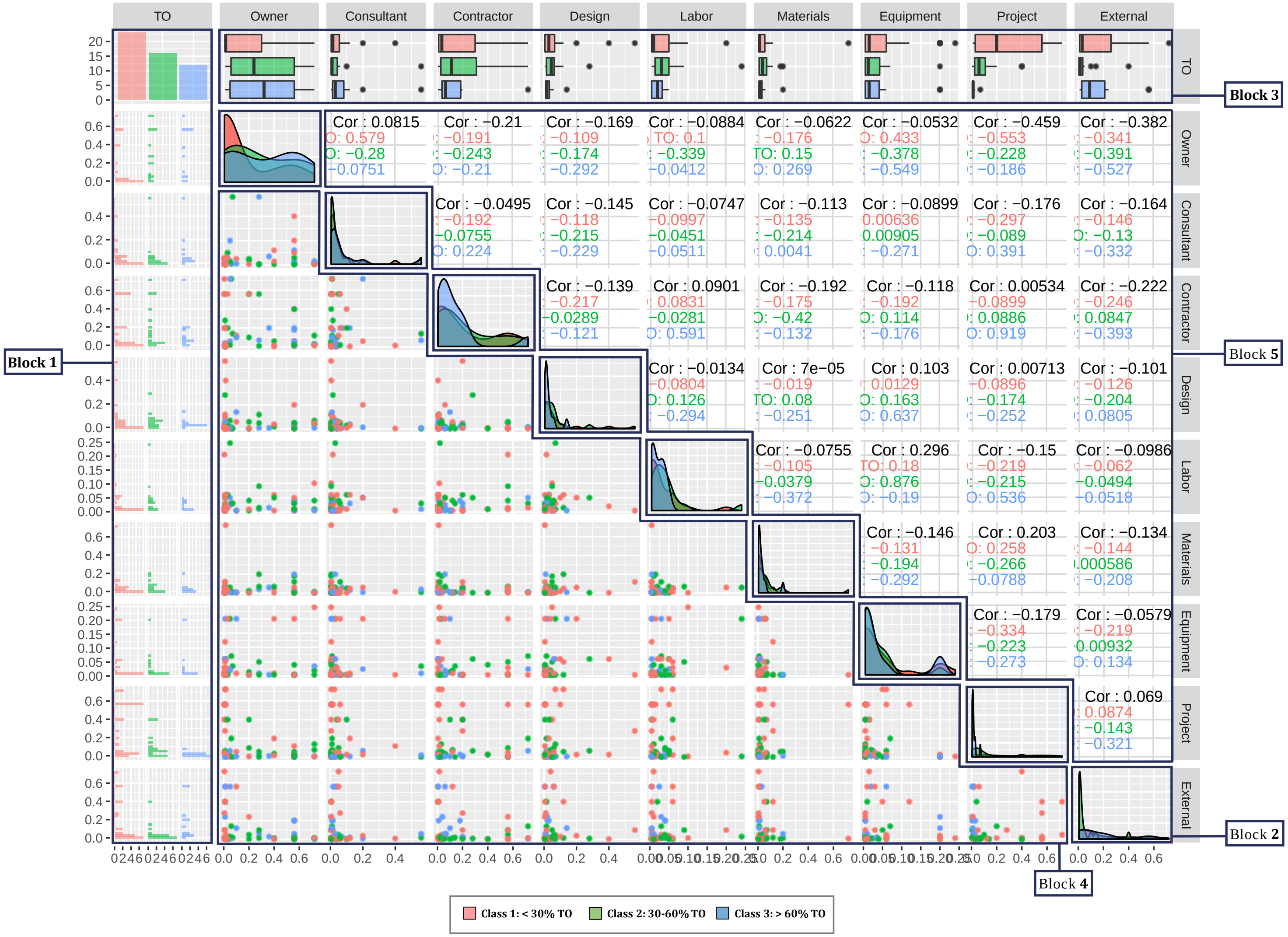


Figure 2.3: Exploratory and sensitivity data analysis of time overrun extent and the nine risk sources

the risk contribution values of the owner, contractor, project and external risk sources are higher than those related to other risk sources.

The *second insight* is inferred through the box plots located within **Block 3** that demonstrate how the changes of TO from one class to another are sensitive to/affected by any changes in the risk contribution values of different risk sources. These box plots illustrate the risk contribution value ranges and distributions pertaining to the nine risk sources for each class of TO, where the thick black bars and the middle boxes represent the median values and the inter-quartile ranges of these distributions, respectively. An important observation is that, for each of the nine risk sources, the risk value distributions for each class are highly overlapping. This overlapping indicates that no individual risk source is fit to provide a definite/clear distinction of the TO class, which subsequently entails that the risk sources are heavily interdependent with regards to influencing TO, and that the studied system exhibits a significant level of complexity. Nonetheless, some findings from the box plots will be described for completeness. Most notably, higher project risk values are associated with the < 30% TO class, while lower risk values are not distinctly related to a certain TO class. This implies that minor TO extents are sensitive to project-related risks, whereas moderate and major TO extents are insensitive. Moreover, lower owner risk values are associated with the < 30% TO class whereas with higher risk values, 30-60% and > 60% TO extents tend to occur. Accordingly, for the owner-related risks, minor, moderate and major TO extents are all sensitive, however, there is no clear separability/discrimination between the latter two since the risk value distributions for these two classes are nearly identical and almost fully overlapping. Through similar interpretations, it can be noted that for contractor risks, minor and moderate TO extents are sensitive albeit with a lack of clear separability between the two classes, and that major TO extents are insensitive. In addition, for External risks, only minor TO is sensitive. A final observation is that TO is insensitive to consultant, design, labor and materials risks, which is also attributed to the fact that these risks' ranges are all limited to lower risk values.

Regarding the *third insight*, **Blocks 4** and **5** represent a correlation matrix that illustrates the dependencies among the risk sources. The boxes in **Block 4** show scatter plots of the risk values of every two risk sources. Complementing these scatter plots, the boxes in **Block 5** show the corresponding correlation values (ranging from +1 to -1), which describe the strength and direction of the relationship between the risk sources, both for all project records collectively and for project records within each TO class separately. The magnitude of the correlation value is indicative of the strength of the relationship between any two risk sources (i.e., how much variance in one data variable is explained by the variance in the other variable), whereas the sign of the correlation value specifies the direction of this relationship (i.e., whether the data variable values vary together positively or negatively). As expected, no strong positive or negative correlation exists among any pair of risk sources, either for all project records collectively or for project records within a specific TO class separately. This implies that: 1) for the considered dataset, no two risk sources are related to one another, adding to the complexity of the system; and 2) the property of variable conditional independence for each class is manifested in the current dataset of project records. The latter finding played an important role in selecting one of the predictive analysis approaches, as will be discussed in the following section.

The *three key conclusions* from the exploratory and sensitivity data analyses conducted in this section can be summarized as: 1) delay risk sources are highly interdependent which is evident from the lack of any specific dependence trend of the TO on any individual risk source; 2) delay risk sources and TO classes are related in a complex manner which is evident from the class inseparability issue; and 3) delay risk sources are also, among themselves, related in a complex manner which is evident from the weak correlations between pairs of risk sources. These conclusions demonstrate the complexity of the studied system (i.e., the manner in which the independent and dependent variables are altogether related), and thereby asserting the complex nature of the construction sector and its inherent delay-inducing risks. It is this complexity that guided the selection of a capable predictive data analytics tool in order to realize an accurate predictive delay risk analysis model, as will be discussed in the next section.

2.6 Selection of Predictive Analytics Tool and Algorithms

2.6.1 Tool Selection

ML is one of the most promising tools in predictive data analytics. It combines methods from statistics, database analysis, data mining, pattern recognition and artificial intelligence to extract trends, interrelationships, patterns of interest and useful insights from complex datasets (Aburrous et al., 2010; Flath et al., 2012). In the present study, ML was selected over other predictive data analytics tools, as, for example, statistical learning (**SL**), for two main reasons.

First, as detailed by Breiman (2001), Tixier et al. (2017), and Dindarloo and Siami-Irdemoosa (2017), SL-based models require both formal model structures and data frequency distributions to be imposed a priori to the data fitting processes either based on some knowledge of the system or arbitrarily through assumptions. However, data generated from complex systems (such as those analyzed herein) would rarely have model structures and frequency distributions that are known or tractable. To reiterate, based on the conclusions of the exploratory and sensitivity data analyses conducted in the previous section, the complex nature of the studied system was apparent through: 1) the highly interdependent delay risk sources which lacked any specific trend for the subsequent TO extent – indicating that the underlying model structure is complex for such a system; and 2) the complex relationships between the risk sources, both with the TO classes and among themselves – indicating that the underlying variable frequency distributions for such a system are intractable.

The power of ML algorithms relies on their ability to avoid the limitations of any explicitly programmed instructions concerning the model structure and any hypothetical assumptions pertaining to the data frequency distributions. In fact, the underlying ML assumption is that the forms by which the independent variables and dependent variable are altogether related are complex and unknown. ML thus focuses on learning from the implicit data patterns through algorithms that continuously improve their performances through experience and induction. Based on the above, ML algorithms are not only effective in dealing with data variables having simple linear or nonlinear relationships, but also with variables having complex high order relationships, or even disjunctive variables. It can thus be argued that the adoption of conditioned analysis methods (e.g., SL) to analyze data collected from complex systems may result in imposed model structures and/or data frequency distributions that are poor representations of the actual system's phenomena. Such adoption would thus undermine the predictive accuracy of the resulting model compared to the models to be generated by the more versatile ML approach.

Second, ML uses optimization techniques to maximize the predictive performances (by minimizing the number of incorrect predictions) of the generated models, while SL focuses on the inferences induced from the relationships between the variables in the statistical model. Therefore, SL-based models typically face deficiencies in their predictive performances when dealing with datasets having a large number of variables, while ML-based models are known to be more suited to analyze such datasets and typically yield higher predictive accuracies (Kim et al., 2008; Aggarwal, 2016). As the data collected within this study consists of nine independent variables (risk sources) and one dependent variable (TO), and also based on the complexity of the system which was previously discussed above and in the "Exploratory and Sensitivity Data Analyses" section, ML was adopted in the present study.

2.6.2 ML Algorithms Selection

The **R** open source platform (R Core Team, 2013) was used by the present study and is a powerful computation tool that supports implementations of different ML techniques such as clustering, classification and regression. Classification is a supervised ML technique that is very effective in predictive data analytics. It is based on learning, via historical/training data, in order to facilitate mapping new input records (e.g., project cases) into specific dependent variable output classes (e.g., the extent of TO) based on relevant independent variable values (e.g., project-anticipated risk contribution levels). The present study focused on using the classification technique because of its capability of handling complex-related variables and is its effectiveness in dealing with categorical variables (Aggarwal, 2016).

Two classification algorithms were applied in this study which are the decision tree (**DT**) and naïve Bayesian (**NB**) algorithms. Through a review of the different ML classification algorithms reported in the literature, these two algorithms were selected, among other reasons, mainly because they are suited to small-sized datasets with a demonstrated history of satisfactory performance (Amor et al., 2004; Chi et al., 2012; Ashari et al., 2013; Aggarwal, 2016).

The DT classification algorithm produces an indictive classifier/model for segmenting new data records into class labels by modelling the classification process through a set of hierarchical decisions (rules) concerning the data variables. The induced decisions are arranged in a tree-like structure which is initiated by identifying the root node and then recursively splitting nodes until no further divisions are possible. The splitting criteria are derived from concepts of information theory which depend on the values of information gain, or entropy reduction, to assess the amount of information needed to generate decisions for segmenting a data record into a class label. Such information theoretic measures are key as they represent the criterion for assessing the hierarchical order of variables along the tree and the splitting of the nodes throughout (Caldas and Soibelman, 2003; Desai and Joshi, 2010; Aggarwal, 2016). Different available ML algorithms in the R platform can be used to develop DT classifiers. Some examples of such algorithms include: ID3 (Glur, 2018), rpart (Therneau and Atkinson, 2018), tree (Ripley, 2018), J4.8 (Hornik et al., 2009) and C5.0 (Kuhn and Quinlan, 2018). The *Recursive Partitioning and Regression Trees (rpart)* algorithm in R was selected within the present study. The rationale behind selecting this algorithm is its proven robustness against noisy data, its capability of learning disjunctive variable relationships, and its proven performance with small datasets (Chi et al., 2012; Aggarwal, 2016).

The NB classification algorithm, on the other hand, produces a classifier which identifies classes for new data records by calculating joint conditional probabilities of the previous data records' independent variable values given their dependent variable class labels. This algorithm is based on Bayes' theorem which quantifies the conditional probability of random variables (Gong et al., 2011). It also assumes that the naïve assumption, that variable values are conditionally independent for each class, holds (Ng and Jordan, 2002; Aggarwal, 2016). The outputs of the produced model are conditional probability scores and mutually exclusive class label designations based on the highest class label joint probability value for the data record (Gong et al., 2011; Bilal et al., 2016; Aggarwal, 2016). The *Naïve Bayes* algorithm in R (Meyer et al., 2017) was selected because it is ideal for small-sized datasets since it is known to converge quicker than other algorithms and, as such, requires less training data (Amor et al., 2004; Ashari et al., 2013). It should be noted, however, that NB algorithms are only suitable for analyzing datasets with conditionally independent variables, which was evident from the properties of the dataset considered in the present study, as explained previously in the “Exploratory and Sensitivity Data Analyses” section.

In terms of DT or NB previous applications within the construction research field, Caldas and Soibelman (2003) presented a model based on automatic hierarchical classification to enhance the access and organization of unstructured text documents within construction management information systems. Another study, carried out by Desai and Joshi (2010), utilized a DT classification mining algorithm to assess the most important factors influencing labor productivity in Indian construction projects. In addition, a study by Chi et al. (2012) applied four different DT classification algorithms to predict the cost performance of projects. Furthermore, an application of classification and regression trees was presented by Chou and Lin (2013) to proactively forecast disputes in the initiation phase of public-private partnership projects.

Jiang and Mahadevan (2008) proposed a Bayesian probabilistic methodology to assess the nonparametric damage detection of building structures. Bayesian learning methods were also employed by Gong et al. (2011) for identifying and classifying worker and heavy equipment actions in challenging construction environments from video datasets. Moreover, Bayesian networks were applied to analyze the specific causes of different types of accidents associated with the construction of embankments by Gerassis et al. (2016). Evidently, and to the best of the authors’ knowledge, it can be seen that although ML classification algorithms are widely used in various disciplines, their applicability has rarely been exploited in the delay risk analysis area.

2.7 DECISION TREE AND NAÏVE BAYESIAN CLASSIFIERS TRAINING

This section focuses on describing the analyses performed to achieve the second objective of the present study (see Figure 2.1). Both DT and NB classification algorithms were used to generate classifiers that predict the time performance of projects based on their risk source levels by partitioning each project into a class label describing the expected time delay. Each algorithm initially defines the dataset as an information system with a finite set of data records and variable values. Each row is considered a distinct project record, and each column is considered a distinct variable of that record. The model then identifies the independent variables (the nine risk sources) and the single dependent variable (TO). When implementing supervised classification learning, the ML algorithms are employed to learn the internal structure of the dataset to examine the effect of the variations of different variables on the degree of TO sustained. Each algorithm then forms a classifier to predict class labels for any new records.

2.7.1 Decision Tree Classifier

As previously discussed, the DT algorithm analyzes the training data for learning the influences of independent variables on partitioning data records into class labels. The algorithm then outputs a classifier in the form of a tree-like structure that describes the decision flow.

First, information theoretic measures of entropy and gain for all independent variables are calculated to be used as criteria for tree construction and node splitting. The classifier considers an independent variable to be more informative, and thus affects the dependent variable more significantly, when more information is induced by knowing the variable's value for predicting the dependent variable's class label. In other words, an independent variable is relevant if by eliminating knowledge about this variable, estimating the dependent variable can be substantially adversely affected. The algorithm then builds on these information theoretic insights and develops a knowledge-inductive decision tree for partitioning new records into predefined classes through conjunctive *if-then* rules.

Applying the DT algorithm to the described dataset used in this study generates the decision tree structure shown in Figure 2.4. The decision tree grows from the top node, referred to as the root node, and forms a hierarchical structure to map new data records (representing risk source levels of new construction projects) into class labels (describing the project's expected extent of TO).

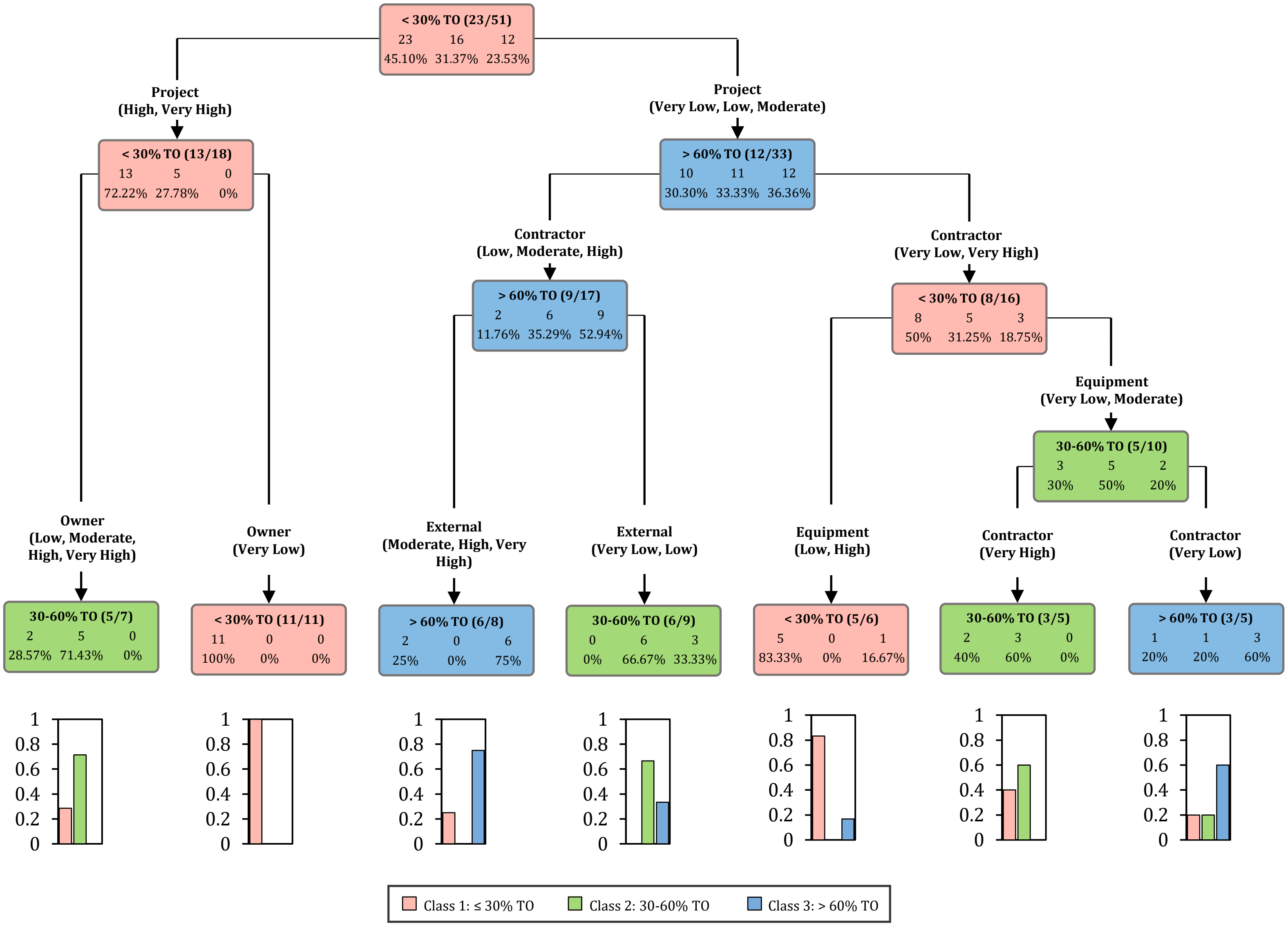


Figure 2.4: Decision tree for predicting project time overrun from risk source levels

Apart from the root node, the tree consists of internal nodes, leaf nodes and branches. In general, nodes represent class labels, and branches refer to the associated variables and variable values. Starting from the root node, data is recursively split to form new tree levels, where each level comprises internal nodes connected by branches. This splitting criterion is based on the previously explained information theoretic measures, where each possible split of the data is examined at each node and the variable with the highest information gain (i.e., highest influence on TO) is chosen for splitting the data. Accordingly, the branches represent possible variable values from the node which they originate.

In that sense, each node represents a subset of the data space defined by the combination of split criteria in the nodes above it, which is why the root node is the only node corresponding to the entire feature space. In addition, each node is labeled with the dominant class according to the distribution of classes within the node. Nodes also return information regarding this class distribution in the form of counts and percentages of the class labels at that node. This recursive partitioning process continues until there is no more benefit from further data segmentation, signaling that additional tree levels would cause data overfitting which would lead to higher levels of misclassification, thus diminishing the algorithm’s predictive performance. Nodes that are at the end of the last branches on the tree are called leaf nodes and play an important role when the tree is used as a predictive model. These leaf nodes represent the outcomes of all prior decisions and refer to the class label which all data records following the path to that leaf would be segmented into.

By referring once again to the generated decision tree from the 51 compiled projects dataset in Figure 2.4, the following observations can be made. Each node refers to one of the three classes, where red nodes correspond to a dominant Class 1 (< 30% TO), green nodes to a dominant Class 2 (30-60% TO), and blue nodes to a dominant Class 3 (> 60% TO). Each branch represents, out of the nine independent variables, the selected risk source to be split, and indicates its variable levels (Very low, Low, Moderate, High or Very high). The instances of project TO at the root node are all the instances in the dataset. As such, this root node contains 51 instances, of which 23 projects experienced TO of less than 30%, 16 projects between 30-60%, and 12 projects greater than 60%, as explained earlier. Consequently, the root node is labeled as a “Class 1” node. Furthermore, the hierarchical order of variables along the tree and the splitting of each node follow the criteria of information entropy and gain as previously discussed. The top three levels of the tree are occupied by the Project, Owner and Contractor sources, respectively, indicating their high significance towards influencing TO. Similar interpretations can be deduced from the remaining nodes and branches of the decision tree.

For further interpretation of the decision tree logic, the tree can be converted into a set of rules to be used for predicting the time performance of new project cases. Rules are generated by traversing each branch of the tree and collecting the variable values until a root node is encountered indicating the predicted class label. A confidence percentage is associated with each generated rule and describes the confidence in the class predicted by the rule (shown as a probability score in the leaf node). The model produced seven rules which follow a series of logical *if-then* statements. The rules produced from Figure 2.4, taken from top to bottom and from left to right, are shown in Figure 2.5. Interesting predictive patterns can be derived from these rules, and for further clarification, Rules 1 and 2 will be discoursed. Rule 1 states that for a project with High or Very high levels of Project-related risk factors, as well as Low, Moderate, High or Very high levels of Owner-related risk factors, there is a 71.4% chance that the project will be delayed beyond completion date by 30% to 60% of its original project duration. Rule 2 indicates that for a project with High or Very high levels of Project-related risk factors, and a Very Low level of Owner-related risk factors, it is certain that the project will experience a TO of less than 30%. In a similar manner, interpretations can be deduced from the remaining rules.

2.7.2 Naïve Bayesian Classifier

The NB classifier identifies mutually exclusive classes of TO for new project cases through calculations of conditional probabilities of variable values with relation to their class labels. It is based on the Bayes Theorem which quantifies the conditional probability of a random variable, and on the naïve

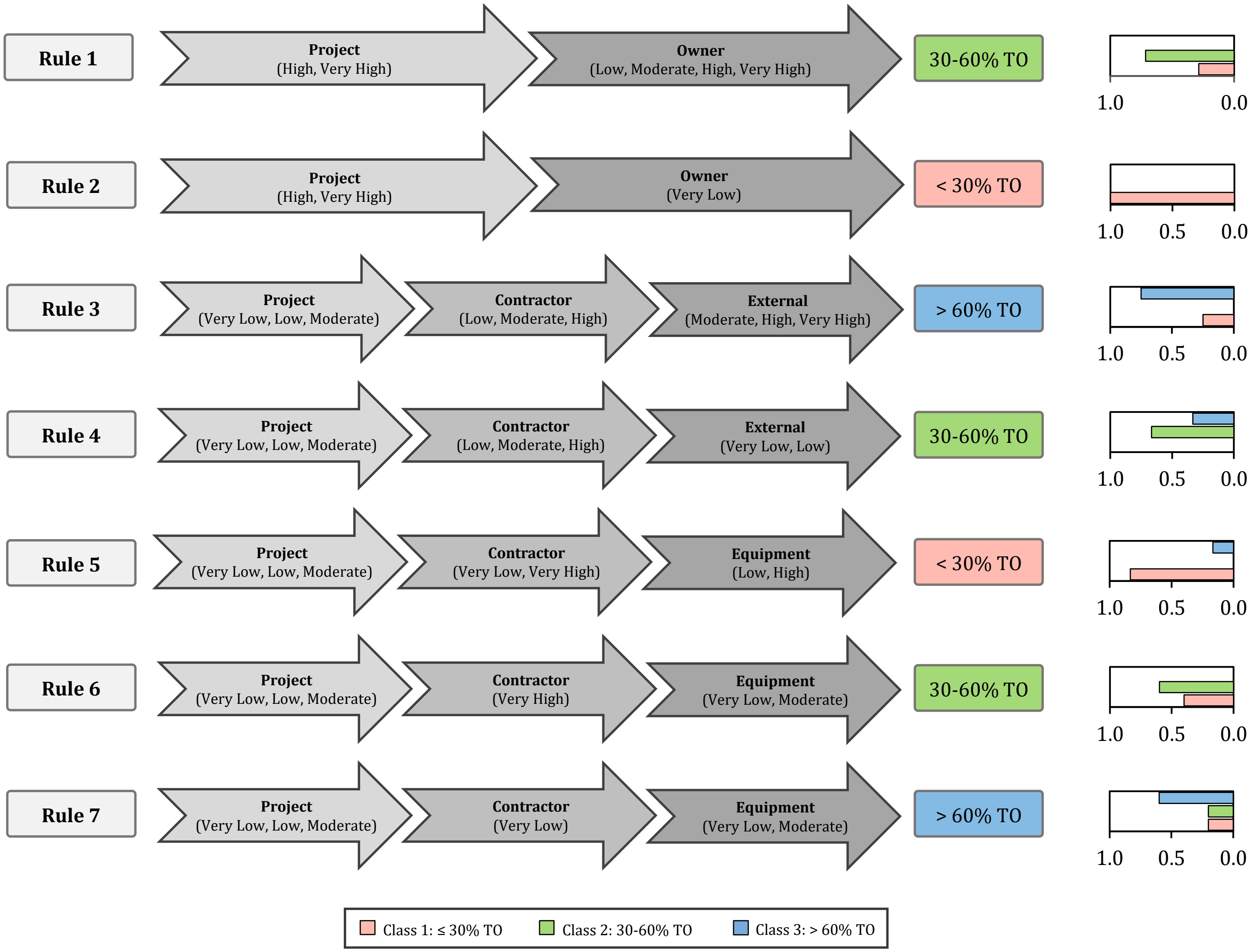

Rule 1
Project (High, Very High)
Owner (Low, Moderate, High, Very High)
30-60% TO
1.0
0.0
Rule 2
Project (High, Very High)
Owner (Very Low)
< 30% TO
1.0
0.5
0.0
Rule 3
Project (Very Low, Low, Moderate)
Contractor (Low, Moderate, High)
External (Moderate, High, Very High)
> 60% TO
1.0
0.5
0.0
Rule 4
Project (Very Low, Low, Moderate)
Contractor (Low, Moderate, High)
External (Very Low, Low)
30-60% TO
1.0
0.5
0.0
Rule 5
Project (Very Low, Low, Moderate)
Contractor (Very Low, Very High)
Equipment (Low, High)
< 30% TO
1.0
0.5
0.0
Rule 6
Project (Very Low, Low, Moderate)
Contractor (Very High)
Equipment (Very Low, Moderate)
30-60% TO
1.0
0.5
0.0
Rule 7
Project (Very Low, Low, Moderate)
Contractor (Very Low)
Equipment (Very Low, Moderate)
> 60% TO
1.0
0.5
0.0
🟥 Class 1: ≤ 30% TO
🟩 Class 2: 30-60% TO
🟦 Class 3: > 60% TO

Figure 2.5: Decision rules for predicting project time overrun from risk source levels

assumption of variable conditional independence. To describe the NB algorithm, the Bayes law must first be introduced for completeness. The Bayes law is shown in Equation (2.1), where *C* is the class label, *A* is the variable value of the new data record, and *P(C|A)* is the conditional probability of *C* given that *A* is observed. The Bayes theorem is useful for estimating *P(C|A)* when it is difficult to be attained from the training data, but other values as *P(A|C)*, *P(C)*, and *P(A)* can be obtained more easily.

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} \tag{2.1}$$

A more realistic approach would be to consider the case where a data record has several *(m)* independent variable values $A = (a_1, a_2, ..., a_m)$. The objective is to assign this record to a definite class C_i (which is one of *n* class labels) such that it corresponds to the maximum value of $P(C_i|A)$. In that sense, Equation (2.2) could be inferred. It is based on the product of conditional probabilities of independent variable values $a_1, a_2, ..., a_m$ given that class C_i is observed.

$$P(C_i|A) = \frac{P(a_1, a_2, \dots, a_m|C_i)P(C_i)}{P(a_1, a_2, \dots, a_m)}$$

$$= \frac{\left(\prod_{j=1}^{m} P\left(a_j|C_i\right)\right) \times P(C_i)}{P(a_1, a_2, \dots, a_m)} \quad \text{,}\textit{where } i = 1,2,\dots, n \tag{2.2}$$

Since the denominator is independent of the class, it thus suffices to only compute the numerator value in order to determine the class with maximum *$P(C_i|A)$*. Therefore, the NB model equation is simplified by removing the denominator as it will have no impact on the conditional probability outcome:

$$P(C_i|a_1, a_2, \dots, a_m) \propto \left(\prod_{j=1}^{m} P\left(a_j|C_i\right)\right) P(C_i) \qquad ,where\ i = 1,2,..., n \qquad (2.3)$$

It can be interpreted from Equation (2.3) that a data record with variable values $A = (a_1, a_2, \dots, a_m)$ is allocated to a class label C_i which returns the highest value of $P(C_i|a_1,a_2,\dots,a_m)$ which is proportional to the product of the various $P(a_j|C_i)$ multiplied by the probability of that class label existing in the data set, which is $P(C_i)$.

Upon application to the dataset, the model outputs are conditional probability scores of each independent variable level for each risk source given each of the three dependent variable class labels. The results of the first risk source (owner risk source) are shown in Table 2.6 as a sample. For any new project case, the model computes values of conditional probability products for its independent variable levels given each class label, multiplied by the probability of retrieving that class from the dataset. Subsequently, the model maps this project to its predicted class label based on the maximum of the three values corresponding to each of the classes. Evaluations of the NB classifier’s predictive accuracy and performance comparisons with the DT classifier will be discussed next.

Table 2.6: Class conditional probabilities for Owner risk source

Class label	Variable level				
	Very low	Low	Moderate	High	Very high
< 30% TO	0.609	0.043	0.000	0.087	0.261
30-60% TO	0.125	0.063	0.188	0.125	0.500
> 60% TO	0.167	0.000	0.167	0.083	0.583

2.8 Model Performance Evaluation and Validation

After introducing and applying both the DT and NB classification models, the purpose of this section is to validate their predictive performance and also evaluate/compare the effectiveness of both models in analyzing the project dataset. Primarily, references from the literature will be called upon to develop a well-rounded interpretation of the common performance of both models based on different domain aspects. Generally, NB classifiers do not require a large amount of data to acquire the internal structure of a dataset, and are therefore better than DT classifiers in learning from smaller training sets while reaching high levels of classification accuracy (Ashari et al., 2013). Moreover, from a computational perspective, NB classifiers are typically faster and more efficient in terms of both their learning and predictive capabilities (Amor et al., 2004). However, NB classification follows the laws of independent events' probability; and hence a central assumption in applying NB classifiers is that for each class, variable values are all conditionally independent of one another. Therefore, DT classifiers typically perform better in domains involving correlated variables. In other words, if two or more variables are highly correlated in NB classification, more weight is allocated to their influence on the predicted class label, which leads to a decline in predictive accuracy. DT models do not suffer from such an undesirable bias because it would not be possible to use two correlated variables for splitting the data of the training set, since this would lead to exactly the same split (Xhemali et al., 2009; Niuniu and Yuxun, 2010). It has been discussed that the considered project dataset is relatively small in size and the projects have variables that are conditionally independent of one another. For these two reasons, and based on the overall inferences presented above, the authors' preliminary hypothesis was that the NB model would generally outperform the DT model for a dataset of such properties.

2.8.1 Performance Evaluation Indices

The models' performance evaluations are facilitated by further developing the algorithms in R to return confusion matrices. Confusion matrices are specific table representations that describe the performance of classification models. The confusion matrices of the DT and NB classifiers are shown in Tables 2.7 and 2.8, respectively, where the classifiers were initially deployed to train on the entire dataset. Confusion matrices include integers reflecting the counts of certain classifications. Rows correspond to the number of actual classifications or total number of records within each class, while columns represent the number of predicted classifications. All correct predictions are located in the diagonal of the table, and this facilitates the visual inspection of the matrix for errors, which are any non-zero values outside the diagonal.

Table 2.7: Confusion matrix from DT classifications

Actual class	**Predicted class**			**Totals**
	< 30% TO	30-60% TO	> 60% TO	
< 30% TO	16	4	3	**23**
30-60% TO	0	14	2	**16**
> 60% TO	1	3	8	**12**
Totals	**17**	**21**	**13**	**--**

Table 2.8: Confusion matrix from NB classifications

Actual class	**Predicted class**			**Totals**
	< 30% TO	30-60% TO	> 60% TO	
< 30% TO	18	3	2	**23**
30-60% TO	3	12	1	**16**
> 60% TO	1	1	10	**12**
Totals	**22**	**16**	**13**	**--**

Before proceeding to evaluate the model performances from the matrices, a few key terms for each class need to be clarified first:

- True Positives (**TPs**): Number of predictions that were correctly assigned to a class (i.e., value in the matrix diagonal for the corresponding class).
- False Positives (**FPs**): Number of predictions that were incorrectly assigned to a class (i.e., sum of values in the corresponding class column excluding the TPs).
- False Negatives (**FNs**): Number of predictions that were incorrectly unrecognized as class assignments (i.e., sum of values in the corresponding class row excluding the TPs).
- True Negatives (**TNs**): Number of predictions that were correctly recognized as not belonging to a class (i.e., sum of values of all rows and columns excluding the row and column of that class).

Based on the aforementioned terminologies, confusion matrices enable analysts to extract numerical measures that act as indicators of the model performance. Such measures could be either overall performance indices or class performance indices, since this is a multi-class classification.

The two model overall performance indices used in this study are *accuracy* and *misclassification error*. Accuracy is a percentage of the total number of correct classifications to the total number of predicted classifications by a model, and correspondingly, the misclassification error (also referred to as the error rate) is a percentage of the direct misclassifications. In other words, the overall accuracy can be perceived as the ratio between the sum of diagonal values and the sum of the table. Thus, the confusion matrix of a highly performing model has large numbers in its diagonal and small numbers (ideally zero) outside the diagonal. Model class performance indices include *precision*, *sensitivity*, *specificity*, *false positive rate (FPR)*, and *false negative rate (FNR)*, and are calculated from the confusion matrices as shown in Equation (2.4), (2.5), (2.6), (2.7) and (2.8) respectively.

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.5}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.6}$$

$$FPR = \frac{FP}{FP + TN} \tag{2.7}$$

$$FNR = \frac{FN}{TP + FN} \tag{2.8}$$

2.8.2 Validation Approaches

The validation of both models was performed based on two approaches which reflect the two stages of a typical ML procedure: training and testing. In the first approach, and due to the small size of the compiled project dataset, *training performance* is evaluated by deploying the models to learn from the *entire dataset* and, subsequently, predict the class labels of the *same data* used for training. In the second approach, to evaluate the *testing performance*, the models were trained to learn from a subset of the entire dataset (a training set) and then predict class labels for the remaining part of the dataset (a testing set) in order to ensure an unbiased evaluation.

The *holdout method* is a common practice for investigating model testing performance, where the complete dataset is randomly split into 80-60% and 20-40% portions for training and testing sets, respectively. However, the major drawback of the holdout method include difficulties with arriving at a random testing set split that would be representative of the entire dataset in terms of: 1) the true variability of the independent variables; and 2) the distributions of the three class labels of the dependent variable in a way that avoids class imbalance (Kim et al., 2008; Chou and Lin, 2013).

In this respect, *k-fold cross validation* was adopted for evaluating the testing performance since it is known to be a reliable method that minimizes the bias and variance associated with the random splitting performed in the holdout method (Kohavi, 1995; Hastie et al., 2009; Arlot and Celisse, 2010; Seong et al., 2018). In *k*-fold cross validation, the complete dataset is divided into *k* distinct and almost equal subsets or folds, where *k* is a positive integer. The holdout method is then repeated *k* times where in each time one of the *k* folds is held out rotationally as the test set and the other *k*-1 folds are put together for training. For every repetition, a confusion matrix is obtained from which overall and class performance indices can be extracted. The final *k*-fold cross validation performance indices are then computed by averaging these *k* individual indices (Gong et al., 2011; Son et al., 2014; Tixier et al., 2017). As such, the advantage of this method is that the entire dataset is used in both training and testing whereas each fold, and hence each data record, is retained exactly once for testing. The present study employed 10-fold cross validation as many researchers have reported *k*=10 to be optimal in terms of computational time, estimation of error, and variance of indices (Kohavi, 1995; Hastie et al., 2009; Wei et al., 2013).

2.8.3 Analysis Results and Discussion

The training and testing performances of both DT and NB classifiers were evaluated based on the performance indices calculated from the confusion matrices. In general, unrepresentable model ability to predict the testing set is apparent from the results generated by the 10-fold cross validation. This is attributed to the small size of the dataset (51 data records) where a single test fold contains 5-6 records and training folds sum up to 45-46 records. Small-size testing and training sets often lead ML models to: 1) overfitting, where the models memorize the peculiarities of the training data rather than its general structure; and 2) generating performance indices with high variances among the test folds (Kim et al., 2008). Therefore, for small-size datasets, it is important to note that testing performance indices may not reflect individual model performance but may rather serve the comparative study between the DT and NB models.

Table 2.9 shows a comparison of the overall performance indices where the training (testing) values of accuracy and misclassification error for the DT classifier are 74.5% (47.2%) and 25.5% (52.8%), respectively, and for the NB classifier are 78.4% (51.2%) and 21.6% (48.8%), respectively. Given the complexity and interdependence of the system, it can be inferred that both models perform reasonably well in terms of training, with the NB classifier exhibiting relatively better performance in both training and testing abilities. Accordingly, the NB model is showing initial signs of exceeding the DT performance as suggested by the writers' hypothesis. Nevertheless, more performance measures need to be examined for a wider perspective.

Table 2.9: Comparison between classifiers based on overall performance indices

Performance Index	DT Classifier	NB Classifier
Accuracy	74.5% (47.2%)	**78.4% (51.2%)**
Misclassification Error	25.5% (52.8%)	**21.6% (48.8%)**

Note. Values without parentheses are training performance indices. Values in parentheses are testing performance indices. Values in bold indicate the superior performance.

Table 2.10: Comparison between classifiers based on class performance indices

Performance Index	DT Classifier			NB Classifier		
	< 30% TO	30-60% TO	> 60% TO	< 30% TO	30-60% TO	> 60% TO
Precision	**94.1%** (51.7%)	66.7% **(40.0%)**	61.5% (20.8%)	81.8% **(54.0%)**	**75.0%** (21.7%)	**76.9%** **(45.0%)**
Sensitivity	69.6% **(63.3%)**	**87.5%** (40.0%)	66.7% (35.0%)	**78.3%** (60.0%)	75.0% **(45.0%)**	**83.3%** **(45.0%)**
Specificity	**96.4%** (59.1%)	80.0% **(90.5%)**	87.2% (66.3%)	85.7% **(61.7%)**	**88.6%** (82.1%)	**92.3%** **(79.2%)**
False Positive Rate (FPR)	**3.6%** (40.8%)	20.0% **(9.5%)**	12.8% (33.7%)	14.3% **(38.3%)**	**11.4%** (20.3%)	**7.7%** **(20.0%)**
False Negative Rate (FNR)	30.4% **(36.7%)**	**12.5%** (60.0%)	33.3% (65.0%)	**21.7%** (40.0%)	25.0% **(55.0%)**	**16.7%** **(55.0%)**

Note. Values without parentheses are training performance indices. Values in parentheses are testing performance indices. Values in bold indicate the superior performance per class label.

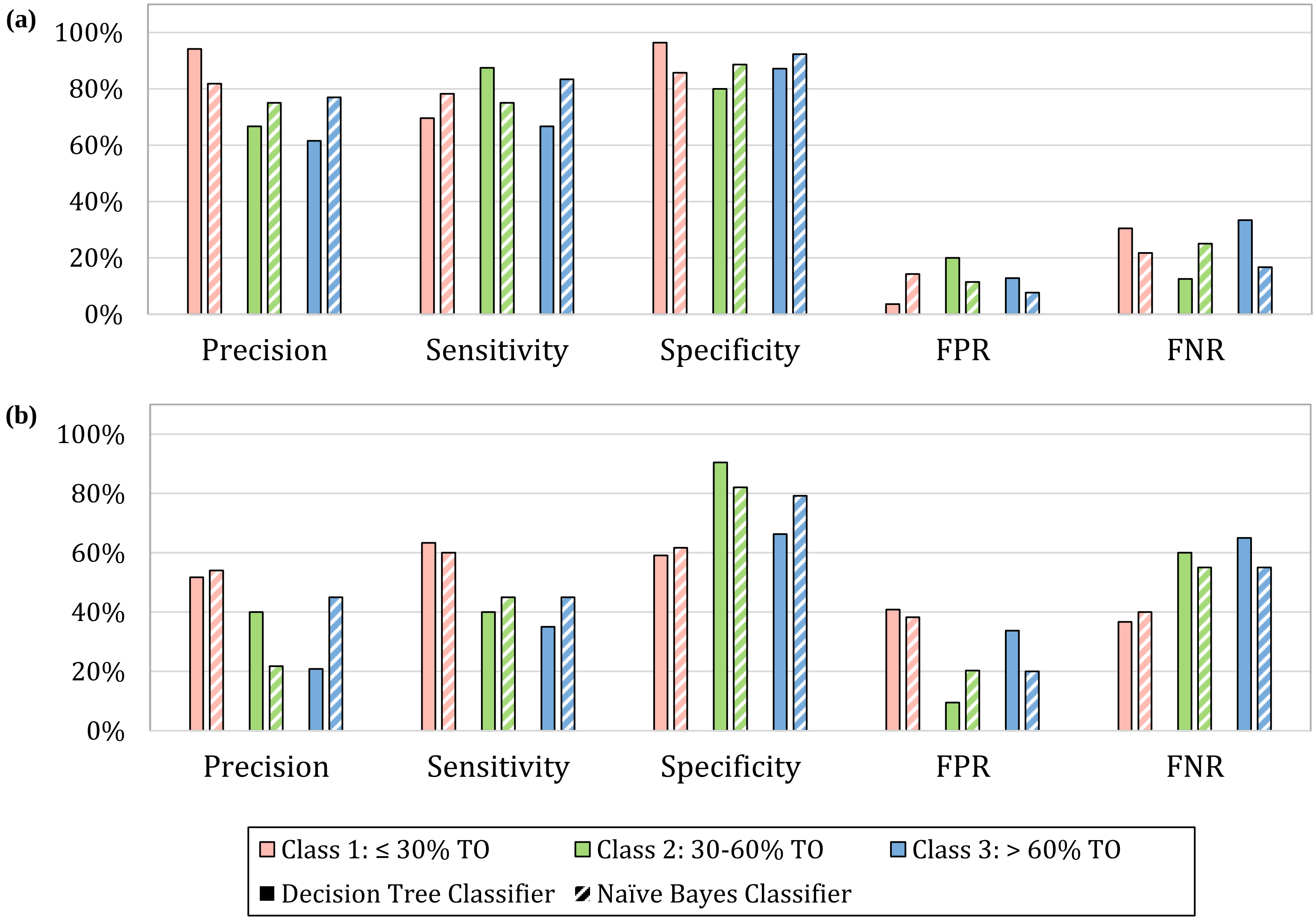


Figure 2.6: Comparison between classifiers in terms of (a) training and (b) testing performances based on class performance indices

Class performance indices are computed by both models for each of the three classes and Table 2.10 summarizes the comparison of training and testing abilities. Moreover, to enhance the visual interpretation of model performance by class, the class performance indices were plotted as shown in Figure 2.6. Overall, it can be stated that the predictive performance of the NB classifier is largely impressive. Apart from its high training accuracy of 78.4%, its training measures of precision, sensitivity and specificity do not fall below 75.0%, 75.0% and 85.7%, respectively, and its measures of FPR and FNR do not exceed 14.3% and 25.0%, respectively. Furthermore, the results display the NB classifier's consistent performance throughout the classes. In addition, the results show that the DT model also performs reasonably well, in terms of training, with a relatively low misclassification error and rather good class performances. The DT classifier's least values of precision, sensitivity and specificity are 61.5%, 66.7% and 80.0%, respectively, and highest values of FPR and FNR are 20.0% and 33.3%, respectively. Similar findings and trends can be interpreted for the testing performances of both DT and NB classifiers.

In terms of comparing both models, it can be inferred that the NB classifier's predictive performance exceeds that of the DT classifier in terms of both training and testing capabilities. The described results indicate the superiority of the NB classifier with regards to the minimum threshold attained for precision, sensitivity and specificity, and maximum threshold attained for FPR and FNR. Other important insights can be made by comparing the models' performance in each class. The NB classifier returns higher values in two out of three class comparisons concerning precision, sensitivity and specificity. As for FPR and FNR comparisons, the NB model was also found to have better performance since it returned lower values in two out of three class comparisons. It is also clear from the figures that the NB classifier displays a more consistent performance across the three classes compared to its DT counterpart. As such, the NB model outperforms the DT model in terms of overall performance, as well as in every class performance measure.

As a final statement, proactive project risk management entails the identification of new arising risk factors as well as the continuous monitoring of both the established and arising risk factors' dynamic behavior throughout the project lifecycle. In this respect, such monitoring involves the timely tracking and reassessment of the risk sources' expected risk severity and recurrence scores, and hence, their overall risk contribution levels. As such, the long-term goal of the present research is to create an analysis platform that, through modifying the independent variable input risk values, would facilitate continuous refinement of project duration prediction throughout the project lifecycle. As a first and key step towards meeting this long-term goal, the focus of the present study was to create trained ML models that are capable of conducting such dynamic analysis when such dynamic data becomes available.

2.9 CONCLUSION

The construction sector is a knowledge-based domain that deals with large volumes of objective, heterogeneous and interdependent data encapsulating abstract knowledge. In most cases, construction firms fail to capitalize on the opportunity presented by this data availability whereas, typically, conventional risk analysis methods, that are heavily dependent on subjective data sources and/or do not consider variable interdependencies within the data, are employed. Nonetheless, exploiting the power of ML data analytics tools can result in significant corporate benefit by enhancing the time performance of construction projects— regarded as one of the key indicators of a successful project.

The present study contributed to this endeavor by identifying and applying ML algorithms to develop two construction project delay risk predictive models based on decision tree and naïve Bayesian classification algorithms. This contribution was realized by reaching two key objectives. First, the main influential risk factors and sources affecting construction projects' delay were identified through a literature survey and consultations with construction sector experts. Subsequently, a dataset, comprising of previous building projects' extents of time overrun and the corresponding contributions of risk sources, was assembled through meetings with construction firms. Throughout this process, several key constraints were considered to ensure data consistency and quality. In addition, and through an exploratory and sensitivity data analysis, an understanding of the complex nature of the construction sector and the interdependence among the various delay risk sources was reached.

Second, a ML-based approach was considered the most suitable to handle such a complex system of interacting variables. Afterwards, two different ML algorithms were carefully selected based on the assembled project data's properties and were employed to create trained predictive models. Finally, the models were evaluated using 10-fold cross validation, among other methods, to generate overall and class performance indices through confusion matrices. The results confirmed the validity of both models and the effectiveness of their predictive performance. The analysis further revealed that, based on both training and testing results, the naïve Bayesian model outperforms the decision tree model in terms of overall performance, as well as in every class performance measure. This finding reflected a consensus with the preliminary hypothesis due to the conditional independence of the data variables.

It should be noted that, although the proposed ML analysis approach is thought to be applicable to tackle complex and interdependent systems of risk sources such as those generated within the construction sector, the specific constraints, properties and limitations associated with the dataset analyzed within this study renders the resulting numerical/categorical values not necessarily transferable to other cases/datasets, as is the case in any data-driven model. Nonetheless, the procedures described in the chapter can be applied to other project datasets, different from the one compiled herein which was studied mainly to facilitate understanding and demonstrate applicability of the proposed ML analysis approach. Subsequently, some recommendations pertaining to future adoption of the methodology described in the chapter are warranted.

First, only variables belonging to the nine identified delay risk sources were considered in the present study. As such, the dataset used in the study had multiple constraints on other external project variables to facilitate homogeneity and meaningful analyses. It is recommended, however, that the influences of other external project variables, that were constrained in the present study, are considered in future applications, as project location, type/end use, duration, contract value, contract type, technical complexity and surrounding area.

Second, the independent variables (nine risk sources) in the dataset were found to reflect properties of conditional independence with one another for each class of the dependent variable (TO). Such properties may not be present in other cases and variables may be correlated. It is thus recommended for future studies that careful sensitivity analysis be primarily carried out as a key step on which to base the selection of adequate ML algorithms for application.

Finally, dataset size can significantly impact ML model performance results. Small-sized datasets increase the chance of model overfitting thus adversely affecting model performance. In such cases, the authors highly recommend selecting models suited to small-sized datasets with a demonstrated history of satisfactory performance. On the other hand, larger datasets are more prone to noisy data which can also undermine model performance. Therefore, noise modelling and outlier analysis techniques are highly endorsed for larger datasets.

Ultimately, the developed methodology can be further incorporated into construction management information systems in support of a proactive project risk management approach that benefits project managers in a twofold manner. The first benefit is the ability to assess and anticipate the time performance of projects, described as the extent of time overrun, from the early planning stages based on the projects' inherent risk levels quantified from these stages. The second benefit pertains to the potential of facilitating continuously refined and more realistic estimates of project durations as the project progresses and while risk factors affecting construction delays are active and dynamic. Overall, such an intelligent platform would influence the state-of-the-practice by addressing the need for transforming multidimensional historical data of completed projects into useful corporate value. Such value would enable construction firms to make knowledge- and evidence-based changes and data-supported decisions to avoid future construction delays.

2.10 ACKNOWLEDGMENTS

The authors are grateful to the financial support of the Natural Science and Engineering Research Council of Canada (NSERC) CaNRisk-CREATE program, and the INTERFACE Institute and the INViSiONLab, both of McMaster University.

2.11 REFERENCES

Abd El-Razek, M. E., Bassioni, H. A., & Mobarak, A. M. (2008). Causes of delay in building construction projects in Egypt. Journal of construction engineering and management, 134(11), 831-841.

Aburrous, M., Hossain, M. A., Dahal, K., & Thabtah, F. (2010). Predicting phishing websites using classification mining techniques with experimental case studies. Proceedings of Information Technology: New Generations (ITNG), 2010 Seventh International Conference, 176-181. IEEE.

Aggarwal, C. C. (2016). Data mining: the textbook, 285-426. Springer.

Aibinu, A.A., & Jagboro, G.O. (2002). The effects of construction delays on project delivery in Nigerian construction industry. International Journal of Project Management, 20, 593–599.

Al-Momani, A. H. (2000). Construction delay: a quantitative analysis. International journal of project management, 18(1), 51-59.

Amor, N. B., Benferhat, S., & Elouedi, Z. (2004). Naive bayes vs decision trees in intrusion detection systems. Proceedings of the 2004 ACM symposium on applied computing, 420-424.

Arditi, D., Akan, G. T., & Gurdamar, S. (2006). Reasons for delays in public projects in Turkey. Construction management and economics, 3(2), 171-181.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.

Ashari, A., I. Paryudi, & A. M. Tjoa. (2013). Performance comparison between naïve Bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. Int. J. Adv. Comput. Sci. Appl. 4 (11): 33-39.

Assaf, S. A., & Al-Hejji, S. (2006). Causes of delay in large construction projects. International journal of project management, 24(4), 349-357.

Aziz, R. F. (2013). Ranking of delay factors in construction projects after Egyptian revolution. Alexandria Engineering Journal, 52(3), 387-406.

Bilal, M., Oyedele, L. O., Qadir, J., Munir, K., Ajayi, S. O., Akinade, O. O., Owolabi, H.A., Alaka, H.A., & Pasha, M. (2016). Big Data in the construction industry: A review of present status, opportunities, and future trends. Advanced Engineering Informatics, 30(3), 500-521.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3), 199-231.

Caldas, C. H., & Soibelman, L. (2003). Automating hierarchical document classification for construction management information systems. Automation in Construction, 12(4), 395-406.

Chan, D. W., & Kumaraswamy, M. M. (1997). A comparative study of causes of time overruns in Hong Kong construction projects. International Journal of project management, 15(1), 55-63.

Chao, L. C., & Chien, C. F. (2009). Estimating project S-curves using polynomial function and neural networks. Journal of Construction Engineering and Management, 135(3), 169-177.

Cheng, M. Y., & Wu, Y. W. (2009). Evolutionary support vector machine inference system for construction management. Automation in Construction, 18(5), 597-604.

Chi, S., Suk, S. J., Kang, Y., & Mulva, S. P. (2012). Development of a data mining-based analysis framework for multi-attribute construction project information. Advanced Engineering Informatics, 26(3), 574-581.

Chou, J. S., & Lin, C. (2013). Predicting disputes in public-private partnership projects: Classification and ensemble models. Journal of Computing in Civil Engineering, 27(1), 51-60.

Desai, V. S., & Joshi, S. (2010). Application of decision tree technique to analyze construction project data. Proceedings of International Conference on Information Systems, Technology and Management, 304-313. Springer Berlin Heidelberg.

Dindarloo, S. R., & Siami-Irdemoosa, E. (2016). Data mining in mining engineering: results of classification and clustering of shovels failures data. International Journal of Mining, Reclamation and Environment, 31(2), 105-118.

Elazouni, A. M. (2006). Classifying construction contractors using unsupervised-learning neural networks. Journal of construction engineering and management, 132(12), 1242-1253.

Ferson, S. (2008). What Monte Carlo methods cannot do. Human and Ecological Risk Assessment: An International Journal, 2(4), 990-1007.

Flath, C., Nicolay, D., Conte, T., van Dinther, C., & Filipova-Neumann, L. (2012). Cluster analysis of smart metering data. Business & Information Systems Engineering, 4(1), 31-39.

Fugar, F. D., & Agyakwah-Baah, A. B. (2010). Delays in building construction projects in Ghana. Construction Economics and Building, 10(1-2), 103-116.

Gerassis, S., Martín, J. E., García, J. T., Saavedra, A., & Taboada, J. (2017). Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. Journal of construction engineering and management, 143(2), 04016093.

Glur, C. (2018). data.tree: General purpose hierarchical data structure: R package version 0.7.6. Accessed August 10, 2018. https://CRAN.R-project.org/package=data.tree.

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. Bayesian analysis, 1(3), 403-420.

Gong, J., Caldas, C. H., & Gordon, C. (2011). Learning and classifying actions of construction workers and equipment using Bag-of-Video-Feature-Words and Bayesian network models. Advanced Engineering Informatics, 25(4), 771-782.

Guyonnet, D., Bourgine, B., Dubois, D., Fargier, H., Come, B., & Chilès, J. P. (2003). Hybrid approach for addressing uncertainty in risk assessments. Journal of environmental engineering, 129(1), 68-78.

Hastie, T., Friedman, J., & Tibshirani, R. (2009). The elements of statistical learning. New York, NY, USA:: Springer series in statistics.

Heravi, G., & Eslamdoost, E. (2015). Applying artificial neural networks for measuring and predicting construction-labor productivity. Journal of Construction Engineering and Management, 141(10), 04015032.

Hornik, K., Buchta, C., & Zeileis, A., (2009). "Open-Source Machine Learning: R Meets Weka." Computational Statistics, 24 (2), 225-232.

Huang, H. T., & Tserng, H. P. (2018). A Study of Integrating Support-Vector-Machine (SVM) Model and Market-based Model in Predicting Taiwan Construction Contractor Default. KSCE Journal of Civil Engineering, 1-10.

Ismail, I., Memon, A. H., & Rahman, I. A. (2014). Expert opinion on risk level for factors affecting time and cost overrun along the project lifecycle in Malaysian Construction Projects. International Journal of Construction Technology and Management, 1(2), 10-15.

Jiang, X., & Mahadevan, S. (2008). Bayesian probabilistic inference for nonparametric damage detection of structures. Journal of engineering mechanics, 134(10), 820-831.

Kerzner, H. (2017). Project management: a systems approach to planning, scheduling, and controlling. John Wiley & Sons.

Kim, H., Soibelman, L., & Grobler, F. (2008). Factor selection for delay analysis using knowledge discovery in databases. Automation in Construction, 17(5), 550-560.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, 1137-1145).

Kokkaew, N., & Wipulanusat, W. (2014). Completion delay risk management: A dynamic risk insurance approach. KSCE Journal of Civil Engineering, 18(6), 1599-1608.

Kuhn, M., & R. Quinlan. (2018). C50: C5.0 decision trees and rule-based models: R package version 0.1.2. Accessed August 10, 2018. https://CRAN.R-project.org/package=C50.

Lam, K. C., Lam, M. C. K., & Wang, D. (2010). Efficacy of using support vector machine in a contractor prequalification decision model. Journal of Computing in Civil Engineering, 24(3), 273-280.

Mahamid, I. (2011). Risk matrix for factors affecting time delay in road construction projects: owners' perspective. Engineering, Construction and Architectural Management, 18(6), 609-617.

Majid, I. (2006). Causes and Effect of Delays in Aceh Construction Industry. Master of Science thesis, University Technology Malaysia.

Mansfield, N. R., Ugwu, O. O., & Doran, T. (1994). Causes of delay and cost overruns in Nigerian construction projects. International journal of project Management, 12(4), 254-260.

Meyer, D., E. Dimitriadou, K. Hornik, A. Weingessel, & F. Leisch. (2017). Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. R package version 1.6-8. Accessed August 10, 2018. https://CRAN.R-project.org/package=e1071.

Mezher, T. M., & Tawil, W. (1998). Causes of delays in the construction industry in Lebanon. Engineering, Construction and Architectural Management, 5(3), 252-260.

Ng, A. Y., & Jordan, M. I. (2002). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In Advances in neural information processing systems, 841-848.

Niuniu, X., & Yuxun, L. (2010). Notice of retraction review of decision trees. Proceedings of 3rd IEEE International Conference for Computer Science and Information Technology (ICCSIT), 5, 105-109.

Odeh, A. M., & Battaineh, H. T. (2002). Causes of construction delay: traditional contracts. International journal of project management, 20(1), 67-73.

Project Management Institute, publisher, A guide to the project management body of knowledge, Sixth edition, Project Management Institute, Newtown Square, PA, 2017.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rezaie, K., Amalnik, M. S., Gereie, A., Ostadi, B., & Shakhseniaee, M. (2007). Using extended Monte Carlo simulation method for the improvement of risk management: Consideration of relationships between uncertainties. Applied Mathematics and Computation, 190(2), 1492-1501.

Ripley, B. (2018). tree: Classification and Regression Trees. R package version 1.0-39. Accessed August 10, 2018. https://CRAN.R-project.org/package=tree.

Sadeghi, N., Fayek, A. R., & Pedrycz, W. (2010). Fuzzy Monte Carlo simulation and risk assessment in construction. Computer‐Aided Civil and Infrastructure Engineering, 25(4), 238-252.

Sambasivan, M., & Soon, Y. W. (2007). Causes and effects of delays in Malaysian construction industry. International Journal of project management, 25(5), 517-526.

Seong, H., Son, H., & Kim, C. (2018). A Comparative Study of Machine Learning Classification for Color-based Safety Vest Detection on Construction-Site Images. KSCE Journal of Civil Engineering, 22(11), 4254-4262.

Singh, R. (2009). Delays and cost overruns in infrastructure projects: an enquiry into extents, causes and remedies. Centre for Development Economics, Department of Economics, Delhi School of Economics.

Son, H., Kim, C., Hwang, N., Kim, C., & Kang, Y. (2014). Classification of major construction materials in construction environments using ensemble classifiers. Advanced Engineering Informatics, 28(1), 1-10.

Therneau, T., & Atkinson, B. (2018). rpart: Recursive Partitioning and Regression Trees. R package version 4.1-13. Accessed August 10, 2018. https://CRAN.R-project.org/package=rpart.

Xhemali, D., Hinde, C. J., & Stone, R. G. (2009). Naive Bayes vs. decision trees vs. neural networks in the classification of training web pages. Int. J. Comput. Sci. 4 (1): 16-23.

Chapter 3:

MACHINE LEARNING-BASED DECISION SUPPORT FRAMEWORK FOR CONSTRUCTION INJURY SEVERITY PREDICTION AND RISK MITIGATION

ABSTRACT

Construction is a key pillar in the global economy, but it is also an industry that has one of the highest fatality rates. The goal of the current study is to employ data analytics (e.g., machine learning) to develop a framework capable of extracting hidden safety knowledge, upon which better-informed and interpretable injury-risk mitigation decisions can be made for construction sites. Central to the framework, generalizable decision tree and random forest models are developed and validated to quantitively predict injury severity levels from the interdependent effects of identified key injury factors. To demonstrate the framework utility, a dataset pertaining to construction site injury cases is utilized. By employing the developed decision support framework, safety managers can evaluate different construction sites' safety risk levels and the subsequent potential high-risk sites can be flagged for ultimately formulating targeted (site-specific) proactive risk mitigation strategies. Managers can also utilize the framework to explore interdependent factors and corresponding cause-and-effect relationships with injury severity which can enhance their understanding of the

underlying mechanisms that shape construction safety risk. Overall, the current study offers interpretable and generalizable decision-making insights for safety managers and workplace safety practitioners to better identify, understand, predict and control the factors influencing injuries on their construction sites and ultimately improve the safety level of their working environments and prevent or reduce injuries.

3.1 INTRODUCTION

3.1.1 Background

The construction industry continues to be one of the most dangerous industries worldwide (Jin et al. 2019; Marin et al. 2019; Alkaissy et al. 2020). For example, in the U.S., the construction industry accounted for approximately 22% of all work-related fatalities in 2019, despite employing only about 7% of the workforce, thus being the deadliest industry in the country with a fatality rate three times greater than the all-industry average (Hallowell et al. 2013; Bureau of Economic Analysis 2021; Bureau of Labor Statistics 2021a). More specifically in 2019, the U.S. construction industry suffered from over 1,000 fatal injuries in addition to more than 200,000 nonfatal injuries (Bureau of Labor Statistics 2021b). Such alarming examples of fatality and injury rates, along with their corresponding societal burdens and financial losses, have elevated the urgent research need to cultivate safer work environments across construction sites.

Primary safety research efforts were focused on identifying sets of construction injury factors (**CIFs**) or root causes (Zohar 1980; Dedobbeleer and Béland 1991; Mattila et al. 1994; Glendon and Litherland 2001). Such CIFs include but are not limited to: a) *project factors* as company size, project type, end use, duration, contract amount, and number of involved contractors; b) *work condition factors* as worksite environment, hazard exposure, and project hazard level; c) *human factors* as human error and mental state; d) *competence factors* as work familiarity, site experience, and safety training; e) *production pressure factors* as work pace, overload, and fatigue; f) *motivation factors* as wages, incentives and job satisfaction; g) *personal factors* as age, gender, and marital status; h) *weather factors* as temperature, humidity, wind speed, precipitation and snowfall; and i) *safety management factors* as safety programs, policies and compliance with procedures. Recent research studies attempted to quantify the impacts of different CIFs on injury incidence; however, they mainly relied on opinion-based data collection methods (e.g., structured interviews and questionnaire surveys). Specifically, such studies evaluated the relative influences of the different CIFs based on professional experience and intuition of relevant safety managers (McDonald et al. 2000; Mohamed 2002; Zohar 2002; Mearns et al. 2003; Fang et al. 2006; Pereira et al. 2018). More recent research studies centered around employing statistical analyses (e.g., multiple linear regression, bivariate correlation, and factor analyses) to such opinion-based collected data to study the relationships between CIFs and the incidence of injuries (Choudhry et al. 2009; Chen and Jin 2013; Marin et al. 2019; Pereira et al. 2020).

While the latter efforts are certainly valuable, they nonetheless suffer from two key drawbacks. *First*, subjective data can, in many cases, be biased/influenced by personal/subjective judgment and ‘gut-feeling’ (Akhavian and Behzadan 2013). *Second*, injuries can be viewed as the resulting outcome of combined and interdependent effects of multiple fundamental CIFs. For instance, the type of hazard exposure is typically influenced by changes to worksite environment and/or weather conditions (Feng et al. 2014). However, to simplify the resulting statistical models, the above research efforts considered CIFs in isolation (i.e., independently), thus reducing the reliability of such models and limiting their generalization and widespread adoption (Mohammadi et al. 2018). As such, these drawbacks have highlighted the necessity for an alternative safety decision support approach as presented in the current study.

3.1.2 Point of Departure

Empirical safety data and information are among the most valuable assets for organizations' safety decision-making. Currently, such data are becoming more available as a result of legislations requiring employers from various industries (i.e., including construction) to report work-related injury and fatality incidences along with reports that describe project and worksite circumstances surrounding these incidences (Huang et al. 2018; OSHA 2021). From such reports, key CIFs and construction injury severity levels (ISLs) can be extracted and organized into database formats to facilitate their use for further analysis.

An equally important asset is a suitable data analytics platform that addresses challenges of statistical analysis techniques, as discussed earlier, to accurately extract intrinsic injury-related patterns and thus hidden safety knowledge from such empirical datasets (Zhou et al. 2019). In this context, machine learning (ML) is known to model and predict complex phenomena (e.g., construction injuries) with interdependent variables (e.g., CIFs) and outcomes (e.g., ISL) due to its capability to discover the nonlinear complex relationships between such variables and outcomes without statistical assumptions (Rodriguez-Galiano et al. 2014; Siam et al. 2019; Gondia et al. 2020).

ML has been applied in construction safety research to predict the likelihood of incident types (Tixier 2016; Gerassis et al. 2017; Kang and Ryu 2019; Ayhan and Tokdemir 2020) and incident risk levels (Zhou et al. 2017; Poh et al. 2018; Sakhakarmi et al. 2019), and to assess construction safety climate scores across projects (Patel et al. 2015; Abubakar et al. 2018; Makki and Mosly 2021). However, most developed models within such works are essentially “black boxes” (e.g., random forests, artificial neural networks and support vector machines) that would, for instance, disable interpreting the underlying causation and interrelationships between CIFs and ISL, thus ultimately limiting qualitative safety judgment (Li et al. 2012; Kakhki et al. 2019). In that respect, there has been little discussion in construction safety prediction research about “glass-box” ML methods which can not only enable quantitative predictions, but also support qualitative judgement through interpretable insights that can also deepen managers understanding of the cause-and-effect relationships that exist between CIFs and ISL. For example, decision tree models are among the few powerful glass-box/transparent ML models that can explicitly link/map combinations of CIFs to different outcomes of ISL through rules. These rules can also be used to predict the most likely outcomes for new construction site circumstances (Chi et al. 2012). Random forests, on the other hand, adopt an ensemble of decision trees and aggregate their predictions, which improves the overall predictive performance of the models but again restricts their interpretability (Kuhn and Johnson 2013). What is also not yet clear in construction safety prediction research, is to what extent glass-box models are comparable to black-box counterparts in terms of predictive accuracy—a discussion which can facilitate rationale for interpretability/performance trade-off and model selection criteria. In this respect, a trade-off between decision trees' interpretability and random forests' possible performance enhancement within safety applications is still needed.

3.1.3 Goal and Objectives

Adapting specific ML models, the current study aims at developing an empirical data-driven construction safety decision support framework (see Figure 3.1) that enables quantitative construction injury prediction, while also supporting qualitative safety judgment and interpretation. As shown in Figure 3.1, key influential injury factors contributing to injury incidences are initially evaluated and identified, upon which top-priority safety decisions can be based. Subsequently, decision tree models are developed to facilitate predicting ISL from the combined and interactive effects of CIFs, upon which better-informed and interpretable safety decisions can be based. Adopted as candidates of black-box

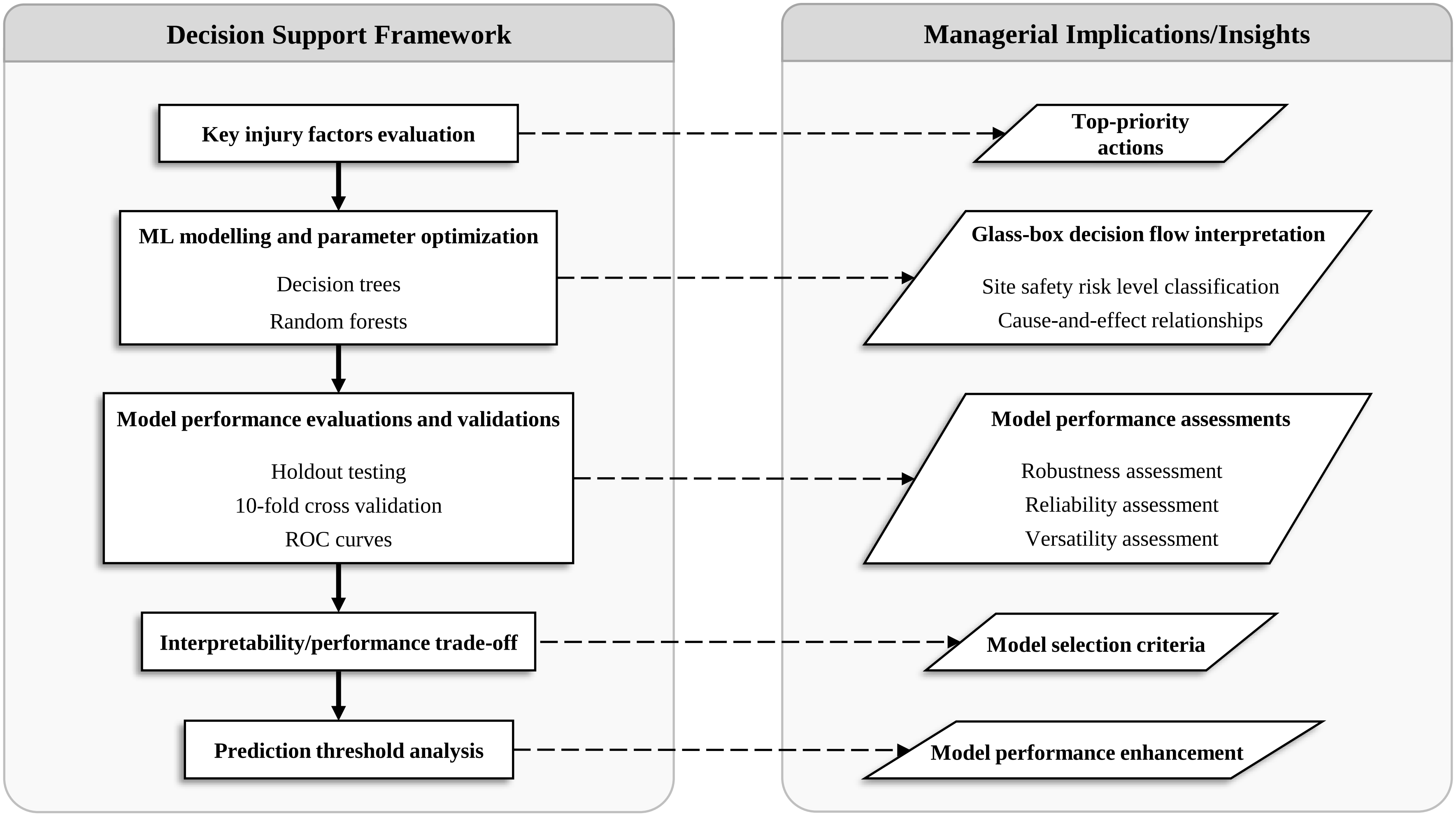

Decision Support Framework
Key injury factors evaluation
ML modelling and parameter optimization
Decision trees
Random forests
Model performance evaluations and validations
Holdout testing
10-fold cross validation
ROC curves
Interpretability/performance trade-off
Prediction threshold analysis
Managerial Implications/Insights
Top-priority actions
Glass-box decision flow interpretation
Site safety risk level classification
Cause-and-effect relationships
Model performance assessments
Robustness assessment
Reliability assessment
Versatility assessment
Model selection criteria
Model performance enhancement

Figure 3.1: Construction safety decision support framework architecture linked with key managerial implications

models, random forests are also developed herein. All model development considers the key step of parameter optimization to yield unbiased/generalizable models. The predictive performance of the developed models is further verified using cross-validation tests and multiple relevant performance evaluation measures. In that respect, guidance is provided on how to select between decision trees' glass-box interpretability and random forests' possible higher performance. Finally, receiver operating characteristics insights are provided to support decision making in adjusting model prediction thresholds to enhance overall model performance. To demonstrate the framework utility and examples of key learnings/managerial insights, an Occupational Safety and Health Administration construction site injury cases dataset is used. Using the developed models, safety managers can evaluate different construction sites' safety risk levels and classify them, with respect to injury severity levels, to potentially high-risk or low-risk zones for ultimately formulating and disseminating prevention strategies in a more targeted and proactive manner.

3.2 FRAMEWORK ARCHITECTURE

3.2.1 Key Injury Factors Evaluation

CIF evaluation and ranking can be performed on an empirical injury cases dataset to: 1) quantitatively evaluate the relative importance of CIF contributions to ISL; 2) identify key CIFs with the greatest influence on ISL to use for making top-priority safety decisions; and 3) proceed with such key CIFs to create a simplified dataset that would reduce computation time and increase the subsequent ML model accuracy (Gerassis et al. 2017; Zhang et al. 2020). The Information Gain Attribute Evaluator (**IGAE**) was used as the ranking algorithm due to its known accuracy with categorical variables and its ability to learn potentially disjunctive patterns (Chi et al. 2012; Aggarwal 2015). The IGAE is based on information-theoretic measures of entropy and gain that are calculated for each variable (i.e., CIF) in the dataset (Shannon 1948). Generally, information entropy (E) is a measure of uncertainty in a variable and denotes the lack of predictability from that variable, whereas information gain (G) is inversely proportional to entropy and infers the amount of information added to the prediction process by including such a variable. For example, variables with lower E values (i.e., higher G values) are more significant within the dataset. Given a variable with a distribution of $X = (x_1, x_2, .., x_m)$, the E of the variable is computed as given in Equation (3.1) (Wang et al. 2010; Aggarwal 2015). The greater the variable distribution randomness, the less E the variable contains, which indicates that the maximum E (i.e., least conveyed information) is observed when the variable is uniformly distributed.

$$E(X) = -\sum_{i=1}^{m} x_i \log_2(x_i) \tag{3.1}$$

When applying the IGAE to a dataset that contains multiple CIFs, each with varying numbers of categories/values, and an outcome ISL with several underlying classes, Equation (3.2), (3.3) and (3.4) are adopted to calculate class-based entropy (Aggarwal 2015; Gerassis et al. 2017; Zhang et al. 2020). First, the base entropy, $E(C)$, is calculated for the entire dataset using Equation (3.2) which is based on the counts of each class within the dataset. This is followed by computing the conditional entropy, $E(C_{kj})$, for each category within every CIF using Equation (3.3). This is the weighted summation of entropies pertaining to a category's distribution across the two classes. The weighted average entropy, $E(C_k)$, over all categories within each CIF is then calculated for each CIF, as shown in Equation (3.4). As such, CIFs with higher $E(C_k)$ imply a greater mixing of the two classes with relation to the distributions of the categories, while a CIF with an $E(C_k)$ value of zero implies perfect separation and therefore the greatest possible predictive power. The final step is to compute values of information gain, $G(C_k)$, for each CIF. The gain in information due to a specific CIF is the difference between the information conveyed for predicting the ISL from the dataset before and after the introduction of that CIF. Specifically, the $G(C_k)$ of the k^{th} CIF is the difference between the base entropy $E(C)$ of the dataset and the weighted average entropy $E(C_k)$ of that CIF, as demonstrated in Equation (3.5). As such, information gain indicates a reduction in entropy, and the most informative/influential CIF toward ISL is that with the highest $G(C_k)$.

$$E(C) = -\sum_{i=1}^{n} \frac{|C_i|}{|C|} \times \log_2 \frac{|C_i|}{|C|} \tag{3.2}$$

$$E(C_{kj}) = -\sum_{i=1}^{n} \frac{|C_{kji}|}{|C_{kj}|} \times \log_2 \frac{|C_{kji}|}{|C_{kj}|} \tag{3.3}$$

$$E(C_k) = \sum_{j=1}^{m} \frac{|C_{kj}|}{|C|} \times E(C_{kj}) \tag{3.4}$$

$$G(C_k) = E(C) - E(C_k) \tag{3.5}$$

where n is the total number of classes, i is the class, C is the total number of cases in the dataset, C_i is the number of cases of the i^{th} class, k is the variable or CIF, j is the category, C_{kji} is the number of cases of the j^{th} category within the k^{th} variable belonging to the i^{th} class, C_{kj} is the total number of cases of the j^{th} category within the k^{th} variable, and m is the total number of categories within the k^{th} variable.

3.2.2 ML Modelling and Parameter Optimization

The injury cases dataset is split into a training set (e.g., 80%) and a testing set (e.g., 20%). The training set is used to train/develop the ML models and the testing set is later introduced to evaluate the predictive performance of such models.

3.2.2.1 Decision Tree Models

3.2.2.1.1 Models Development

In the current framework, three different and commonly used decision tree models were applied: recursive partitioning and regression trees (**RPART**) (Therneau and Atkinson 2018), classification and regression trees (**CART**) (Ripley 2018), and C4.5 (Hornik et al. 2009) models. By comparing the performances of such three models against one another, their validity can be further tested and the most appropriate model for construction injury prediction can be identified. Generally, decision tree models are based on the recursive splitting of the data cases into subsets represented as nodes, where such nodes expand to form a top-down tree-shaped structure (see Figure 3.2a) that describes the decision (and thus prediction) flow (Breiman et al. 1984; Chi et al. 2012; Gondia et al. 2019). Starting from the root (i.e., top) node, the IGAE procedure described earlier is applied on the data subset comprising each node, and the most predictive CIF (i.e., most influential toward ISL) is then selected for splitting such node through two branches into two child nodes. Such node splitting is repeated until all cases within a node belong to only one of the ISL classes, marking a perfect classification and designating such node as a terminal node. Once all terminal nodes are reached, the splitting process concludes, and the resulting *final* tree can be used for predicting the classes of new cases. As a glass-box ML model, decision trees facilitate such prediction through an explicit/interpretable mapping of CIF combinations and corresponding categories to an ISL class by

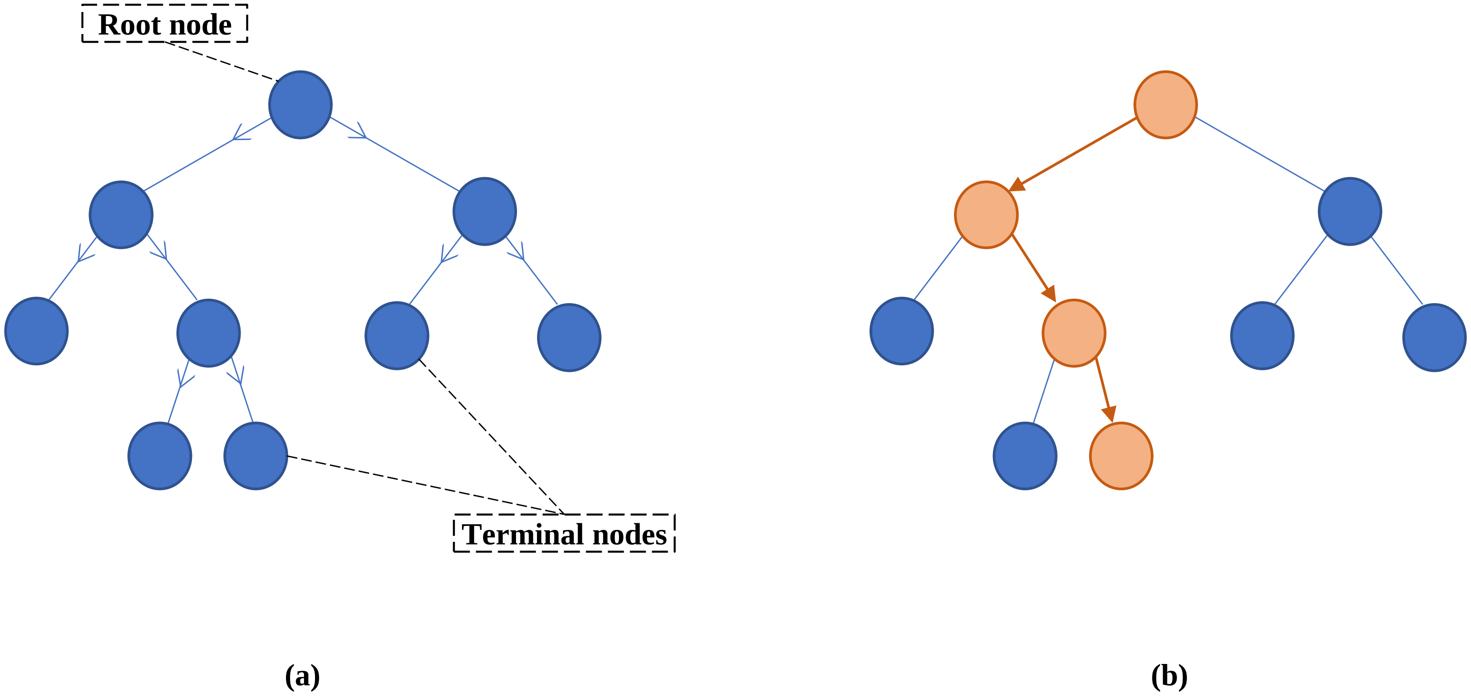


Figure 3.2: Decision tree: (a) structure and node splitting; and (b) mapping CIF combinations to ISL outcome

descending from the root node until reaching a terminal node, as shown in Figure 3.2b.

3.2.2.1.2 Parameter Optimization

If a decision tree model is allowed to grow indefinitely until all terminal nodes are reached, the resulting tree may become extremely complex due to its numerous nodes. Not only will a complex tree complicate its interpretability, but also such a tree will learn the unique peculiarities of the training set rather than its general structure, causing the tree to be incapable of generalizing to new cases—an issue known as overfitting (Kuhn and Johnson 2013). Therefore, tree pruning, through "snipping off" the least important splits, can avoid overfitting. A key step to such tree pruning is optimizing the tree parameters that control tree complexity. In this respect, the objective of tree parameter optimization is to find the optimal level of tree complexity that achieves the right trade-off between predictive accuracy on that training set and another set of new data (Bergstra and Bengio 2012). As such, the decision tree models in the current study are developed through a generalizable approach through tree parameter optimization where a 10-fold cross-validation procedure is applied repeatedly to the training set only, with nine training folds and a single alternating validation fold (Arlot and Celisse 2010). For each 10-fold cross-validation procedure, a specific tree parameter value is selected and the average predictive error (i.e., 1-accuracy) of the resulting trees over the ten validation folds is recorded. Several procedures are repeated to search through the parameter space, and parameters achieving the minimum average cross-validation error are selected to yield the optimum model.

For the RPART model, the parameter to be optimized is the *complexity parameter* (***CP***), which is the minimum improvement in model accuracy needed for a node to be allowed to split. Specifically, while growing the tree, any split is not pursued if it does not minimize the model's predictive error by a factor of *CP*. Regarding the CART model, the parameter available to be optimized is the *maxdepth* that refers to the maximum tree depth in which the tree is prevented from growing past that depth. The depth of a tree is the length of the shortest path from a root node to the deepest terminal node, where for example the tree in Figure 3.2a has a depth of three. As for the C4.5 model, the parameter that needs to be optimized is the *minsplit*, which refers to the smallest number of cases inside a node that could be further split. If a node is found to comprise a number of cases less than the set *minsplit* value, it is labeled a terminal node and does not continue to grow.

3.2.2.2 Random Forest Model

3.2.2.2.1 Model Development

Random forest (**RF**) models are also tree-based models that involve growing/training many single decision trees to form a forest that predicts new cases by combining the output of each tree (Breiman 2001; Zhou et al. 2019). This combination yields a RF model that can typically achieve a better predictive performance than any one individual tree (Rodriguez-Galiano et al. 2012). More precisely, for a dataset with a total of C cases (of which c are training cases) and k variables (i.e., CIFs), the RF procedure is implemented as illustrated in Figure 3.3. First, a number n of data samples (each of the same size c as the training set) is created using bootstrap random sampling with replacement from the original training set, which means that some cases may be repeated (within the same sample) or left out. Two-thirds of the c cases are randomly designated as in-bag cases and are used for tree training, while the remaining one-third is left out for parameter optimization (discussed next) and is known as the out-of-bag set (**OOB**). Second, each in-bag set is used to build a corresponding decision tree; however, at each node of the tree, a subset of k_{try} variables (which is not greater than the total number of k variables) is randomly selected and the best predictive variable from that subset is used for node splitting. Third, each tree contributes with a single vote of a predicted class to the forest and the final prediction result is obtained by considering the majority vote (i.e., the class with the higher frequency of votes). Overall, randomizing the sampling procedure and only trying a random subset of variables at each split result in dissimilar trees (as each is grown on a different data sample and variable subset) with reduced correlations, which ultimately improves the RF model generalization performance. A similar ML technique, bootstrap aggregating (bagging) has one (but key) difference with RF in that the former considers all variables at a node for splitting. This produces single trees that are highly correlated when subjected to a dataset having variables

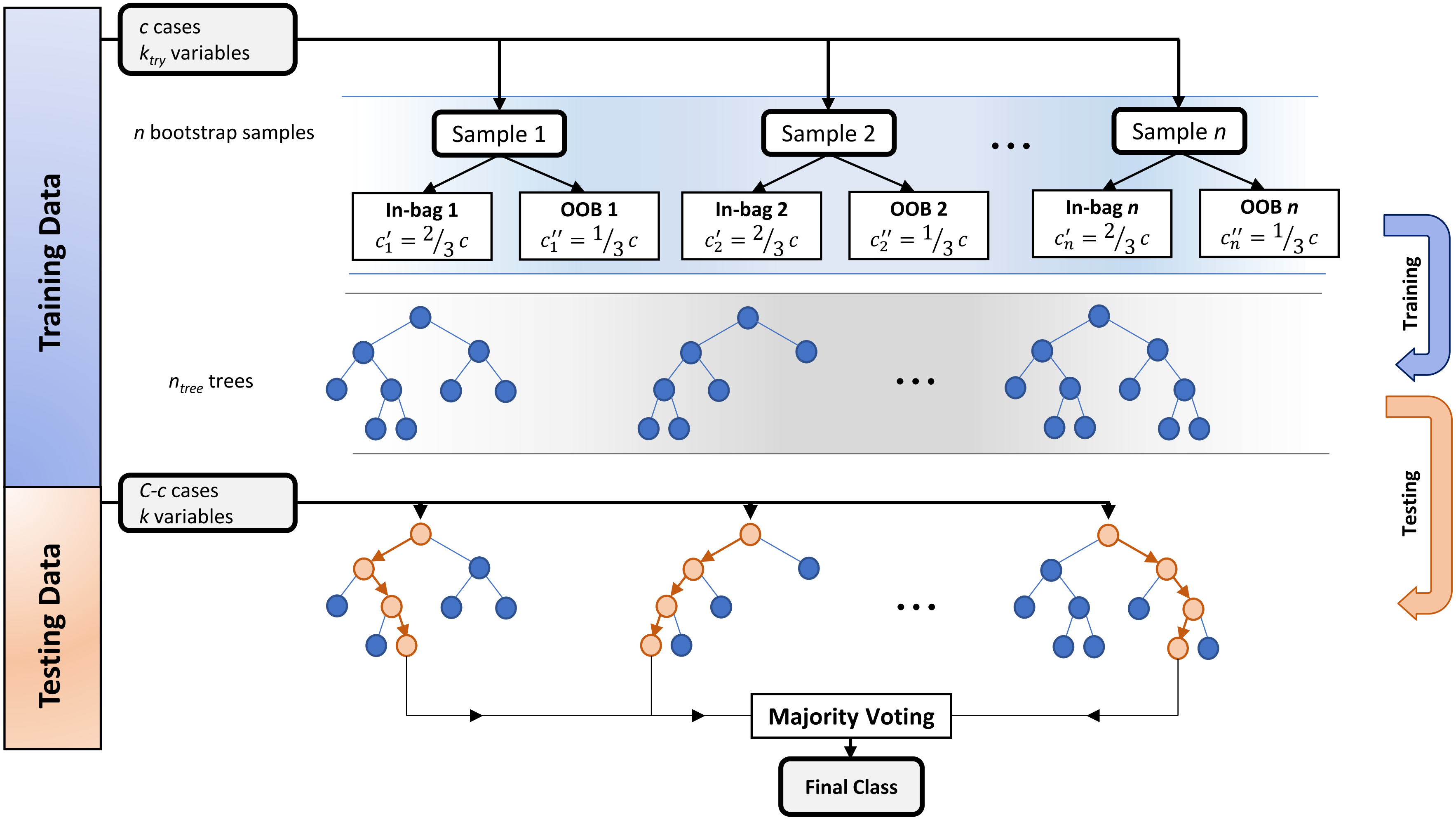


Figure 3.3: Random forest prediction procedure

that are strong predictors (Breiman 1996; Han et al. 2018). However, combining the results of single trees that are essentially similar/correlated does not typically lead to large improvements in generalizability (Zhou et al. 2019)—the effects of which will be evaluated in the application demonstration.

3.2.2.2.2 Parameter Optimization

Two major RF parameters need to be optimized, namely the number n_{tree} of trees in the forest (i.e., same as the number of n obtained bootstrap samples) and the number k_{try} of variables randomly considered at each node split. A too small n_{tree} value may result in insufficient training and a too small k_{try} value may lead to underfitting. In contrast, too large $ntree$ and k_{try} values may cause both overfitting and prolonged computation time. As such, different combinations of these two parameters are used in the current study to sequentially train RF models on the in-bag set (see Figure 3.3). This training is followed by introducing such resulting models to the OOB set (presenting new data that was not used in training) to calculate the average OOB error. The parameter combination resulting in the lowest OOB error is the optimum one.

3.2.3 Models Performance Evaluation and Validation

3.2.3.1 Evaluation Measures

To evaluate the performance of the developed ML predictive models within the framework, confusion matrices (Chou and Lin 2013; Seong et al. 2018) are primarily produced. As presented in Table 3.1, such a confusion matrix exhibits numbers of predicted and actual ISL classes, where the diagonal represents correct predictions and the off-diagonal represents incorrect predictions. For demonstration, the example in Table 3.1 assumes two ISL classes, fatal and nonfatal injuries, which can be used to mark construction sites as high- and low-risk zones, respectively. From such confusion matrices, several performance evaluation measures are then derived. ***Accuracy*** evaluates the overall performance of a model and describes the percentage of correct predictions relative to the total number of predictions (Equation 3.6). As it is particularly important to evaluate the model's ability to accurately predict fatalities, ***precision*** measures how many of the predicted fatal cases are correct. However, the implication of overlooking a fatality (i.e., a high-risk zone) by predicting it as a nonfatality (i.e., a low-risk zone) (**FN**) is more serious than incorrectly predicting a nonfatality as a fatality (**FP**). As such, true positive rate (**TPR**) is a suitable measure to also consider since it takes FN into account (Equation 3.8), unlike precision. Finally, true negative rate (**TNR**) describes the percentage of correctly predicted nonfatalities relative to the total number of actual nonfatalities (Equation 3.9).

Table 3.1: Confusion matrix example

Actual class	**Predicted class**	
	Fatal **(i.e., high-risk zone)**	**Nonfatal** **(i.e., low-risk zone)**
Fatal **(i.e., high-risk zone)**	True Positives (TP)	False Negatives (FN)
Nonfatal **(i.e., low-risk zone)**	False Positives (FP)	True Negatives (TN)

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{3.6}$$

$$Precision = \frac{TP}{TP + FP} \tag{3.7}$$

$$TPR = \frac{TP}{TP + FN} \tag{3.8}$$

$$TNR = \frac{TN}{TN + FP} \tag{3.9}$$

3.2.3.2 Validation Approaches

To validate the robustness, reliability, and versatility of the developed ML predictive models, three approaches were adopted in the current study as presented next.

3.2.3.2.1 Holdout Testing

For ***holdout testing***, the entire dataset is split into an 80% training set and a 20% testing set. The splitting is carried out in a stratified manner, where both training and testing sets have similar distributions of the ISL classes (Kohavi 1995; Fiore et al. 2016). The former set is used to train the models, while the latter set is used to present the trained model with essentially new data to test (assess) its predictive ***robustness*** through the confusion matrix and multiple evaluation measures, as discussed earlier.

3.2.3.2.2 10-fold Cross-validation

For ***10-fold cross-validation***, the entire dataset is divided into ten separate and almost equally-sized folds in a stratified manner with regards to the ISL classes. Nine folds are combined for training and the remaining fold is set aside for testing. Such training and testing are then repeated ten times such that each of the ten folds is used exactly once for testing. For each test fold, a confusion matrix is generated and a corresponding set of evaluation measures are derived and subsequently averaged over the ten folds. Compared to the holdout testing method, the main advantage of 10-fold cross-validation is that the entire dataset is utilized in training and the testing procedure is carried out ten times, thus allowing a better evaluation of the generalization capability of the developed model. This allows for an assessment of the model's ***reliability*** when applied repeatedly to several unseen future datasets.

3.2.3.2.3 ROC Curves

Another performance validation approach that is useful in applications where the seriousness of two prediction errors is significantly unequal (i.e., FP and FN) is the inspection of its ***receiver operating characteristic*** (**ROC**) ***curve*** and the area under that curve (**AUC**) (Son et al. 2014). Typically, the final prediction of a ML model is based on the highest predicted probability of the available classes, and thus for a prediction problem with two classes, the default prediction threshold is 0.5. In that regard, the receiver operating characteristic space is presented by a two-dimensional graph, where values of TPR are plotted on the Y-axis and values of false positive rate (**FPR**) (where FPR = 1-TNR) are plotted on the X-axis for multiple prediction thresholds ranging from 0 to 1. A model with a larger AUC indicates better ***versatility*** in predictive performance because this implies that a larger value of TPR can be achieved for each value of FPR across many different thresholds. As such, the AUC evaluates the ***versatility*** of the model's predictions irrespective of what threshold is used.

3.2.4 Interpretability/performance Trade-off

While decision tree models offer valuable glass-box interpretability merits, RF models may offer better predictive performance improvements, albeit, with restricted interpretability. As such, this subsection can provide managers with decision support pertaining to the criteria for which model to select under different characteristics of the injury cases data at hand. *First*, the RF models are more robust to unbalanced distributions of the outcome class compared to individual tree models (Zhou et al. 2016; Hong et al 2017). *Second*, the RF models can better handle data with large numbers of input variables (Abdel-Rahman et al. 2013; Liu et al. 2018). *Third*, the RF models may perform better than single trees on data with strong predictor variables alongside other potentially irrelevant variables (Sutton 2005; James et al. 2013), as discussed earlier. The application presented later will further demonstrate how managers can use this decision support tool to assess the trade-off between interpretability and performance according to their unique application requirements and data characteristics.

3.2.5 Prediction Threshold Analysis

As briefly described earlier, instead of considering only the default prediction threshold of 0.5, the ROC curve follows the process of 1) varying the threshold values in the range between 0 and 1; 2) storing the corresponding designations of actual and predicted ISL classes for each threshold value; 3) producing corresponding confusion matrices; and 4) plotting corresponding combinations of TPR and FPR. As such, the ROC curve provides a visual tool to better interpret the predictive performance of the model on a wide range of thresholds. Such curves demonstrate the trade-offs pertaining to increasing the model's probability of incorrectly designating a low-risk zone as a high-risk zone (i.e., FPR) for ultimately improving its probability of successfully detecting high-risk zones (i.e., TPR). Safety managers can use such a tool to adjust their model with a better threshold selection and yield a more suitable combination of FPR and TPR probabilities that meet the specific needs of a unique construction site, as will be further discussed in the demonstration application.

3.3 DEMONSTRATION APPLICATION

3.3.1 Dataset Description and Visualization

The dataset used in the current demonstration was obtained from the Occupational Safety and Health Administration (OSHA) injury cases dataset (OSHA 2019), where at the time of access, approximately 3,400 cases corresponded to U.S. construction projects. These reports contained CIF information describing the work circumstances at the time of the injury and the outcome ISL, whether fatal or nonfatal. Figure 3.4 shows the average number of injuries per day for each month. As can be seen in the figure, more injuries occurred during the non-winter months (e.g., May and June) compared to those during the winter months (e.g., December and January) since more construction activities are typically executed during the former months. Eight CIFs were provided, namely: 1) project type; 2) project end use; 3) contract amount; 4) worksite environment; 5) hazard exposure; 6) human error; 7) work familiarity; and 8) month. However, many of the nonfatal injury report cases contained missing CIF-related information. Therefore, such cases were not considered in the current study, thus reducing the dataset to a total of 1,981 injury cases with a complete set of eight CIFs and one corresponding outcome ISL (1,050 fatal and 931 nonfatal). This preprocessing step was necessary to convert the remaining cases into a readable dataset to enable the subsequent ML analysis.

Within the reduced dataset, injury numbers per *month* and distributions across ISL classes are shown in Figure 3.5. In addition, the distribution of categories of *project type*, *project end use* and *contract amount* across ISL classes are shown in Figures 3.6, 3.7 and 3.8, respectively. *Worksite environment* denotes the unsafe nature of the working conditions surrounding- or in close proximity to the worker, while *hazard exposure* is related to the dangers associated with the specific task performed by the worker at the time of the injury. The category distributions of these two CIFs are presented in Figures 3.9 and 3.10, respectively. In addition, the category distributions of *human error*, *work familiarity* and ISL are shown in Figures 3.11, 3.12 and 3.13, respectively.

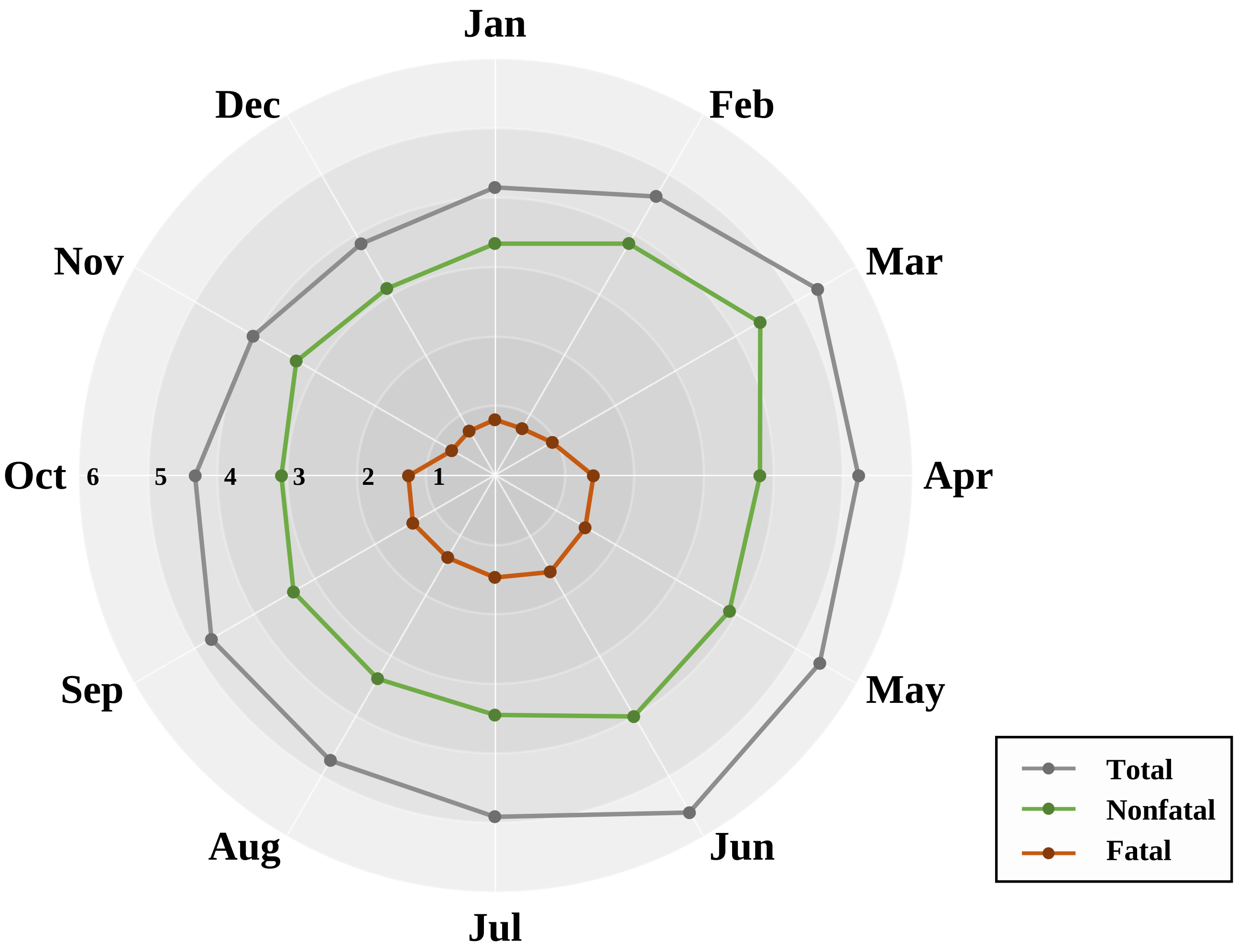


Figure 3.4: Average number of injuries per day (entire dataset)

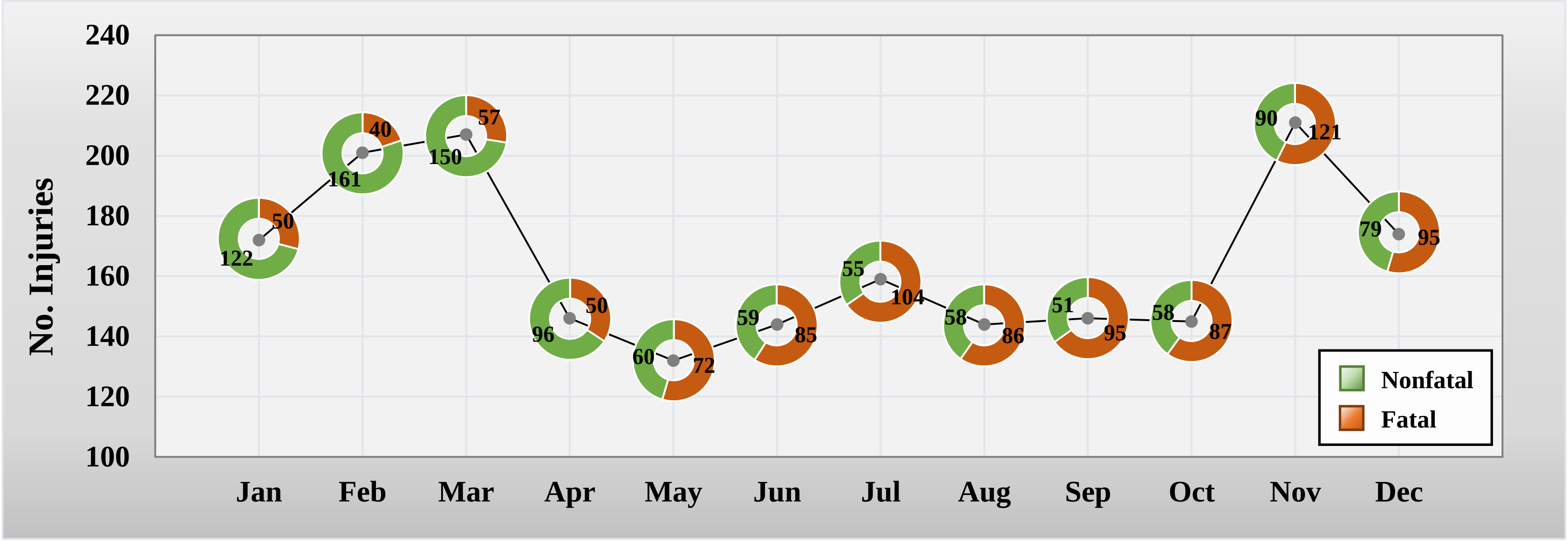


Figure 3.5: Number of injuries per month and across ISL

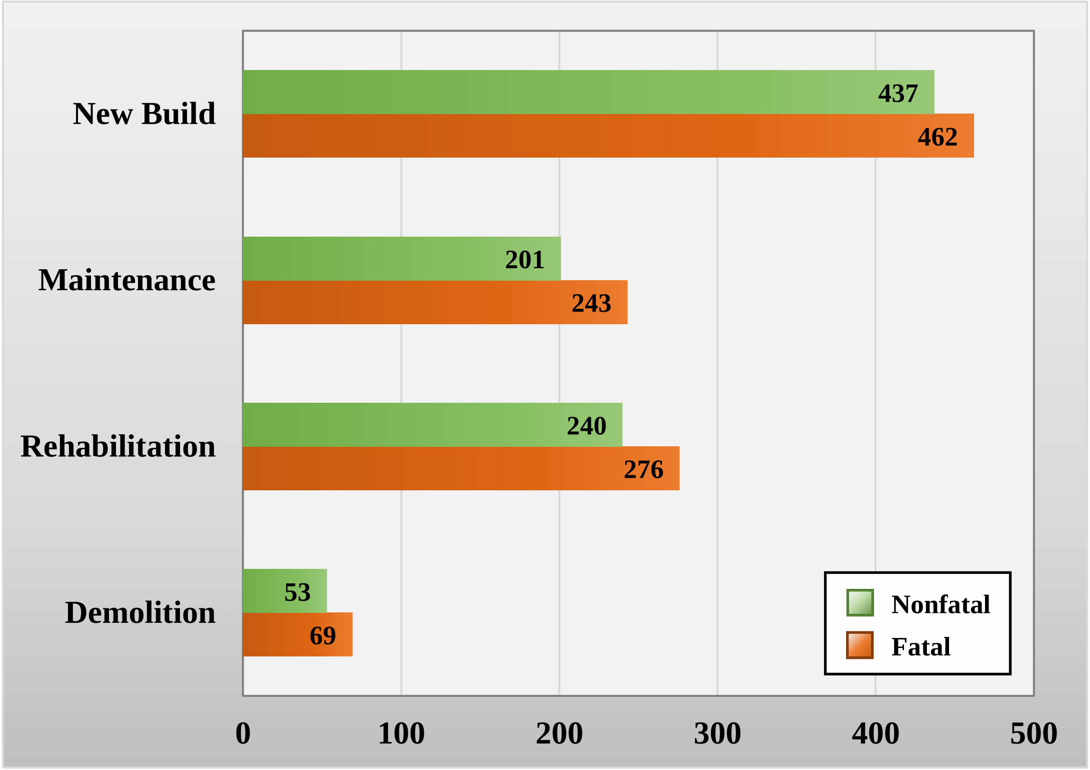


Figure 3.6: Project type category counts across ISL

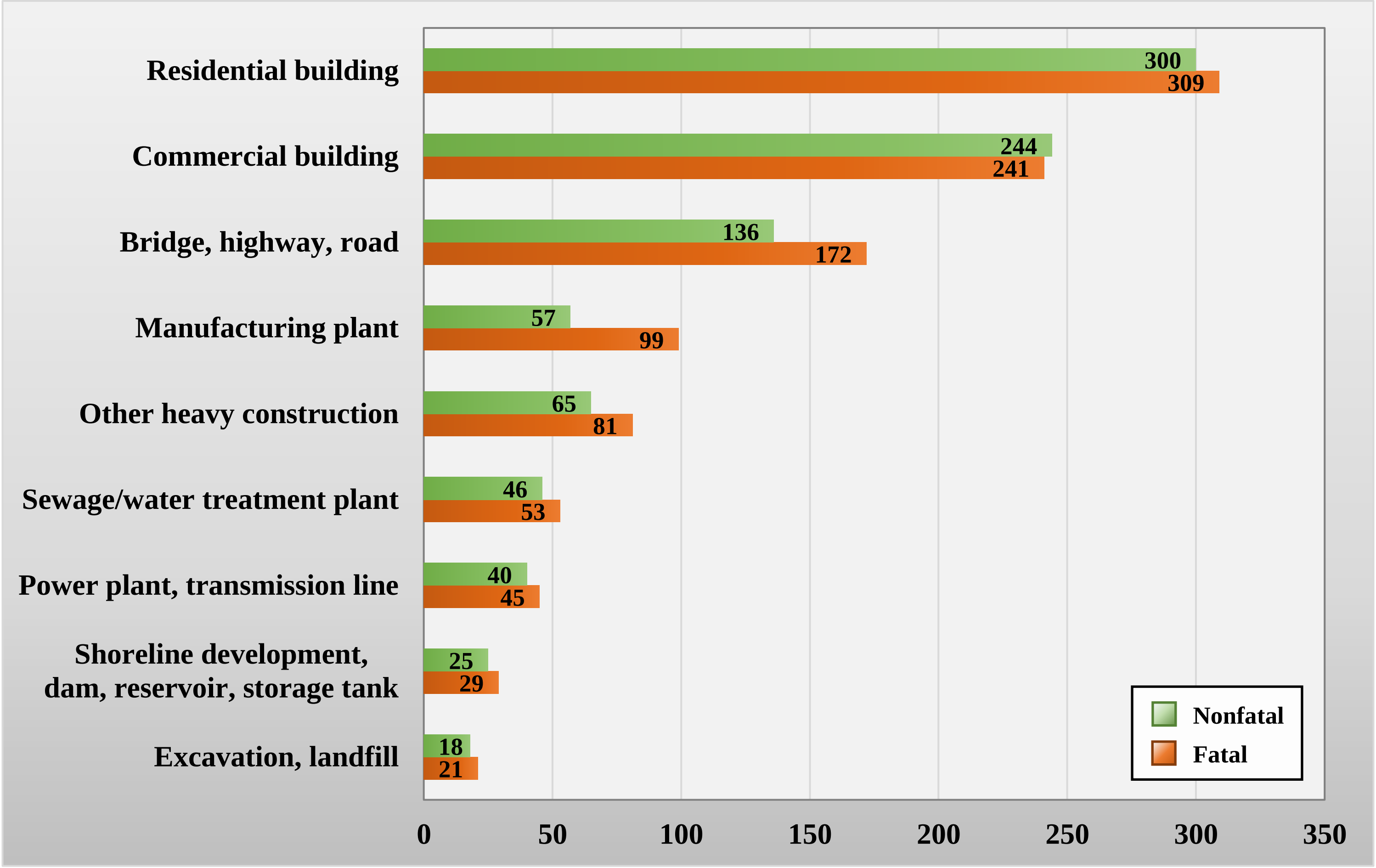


Figure 3.7: Project end use category counts across ISL

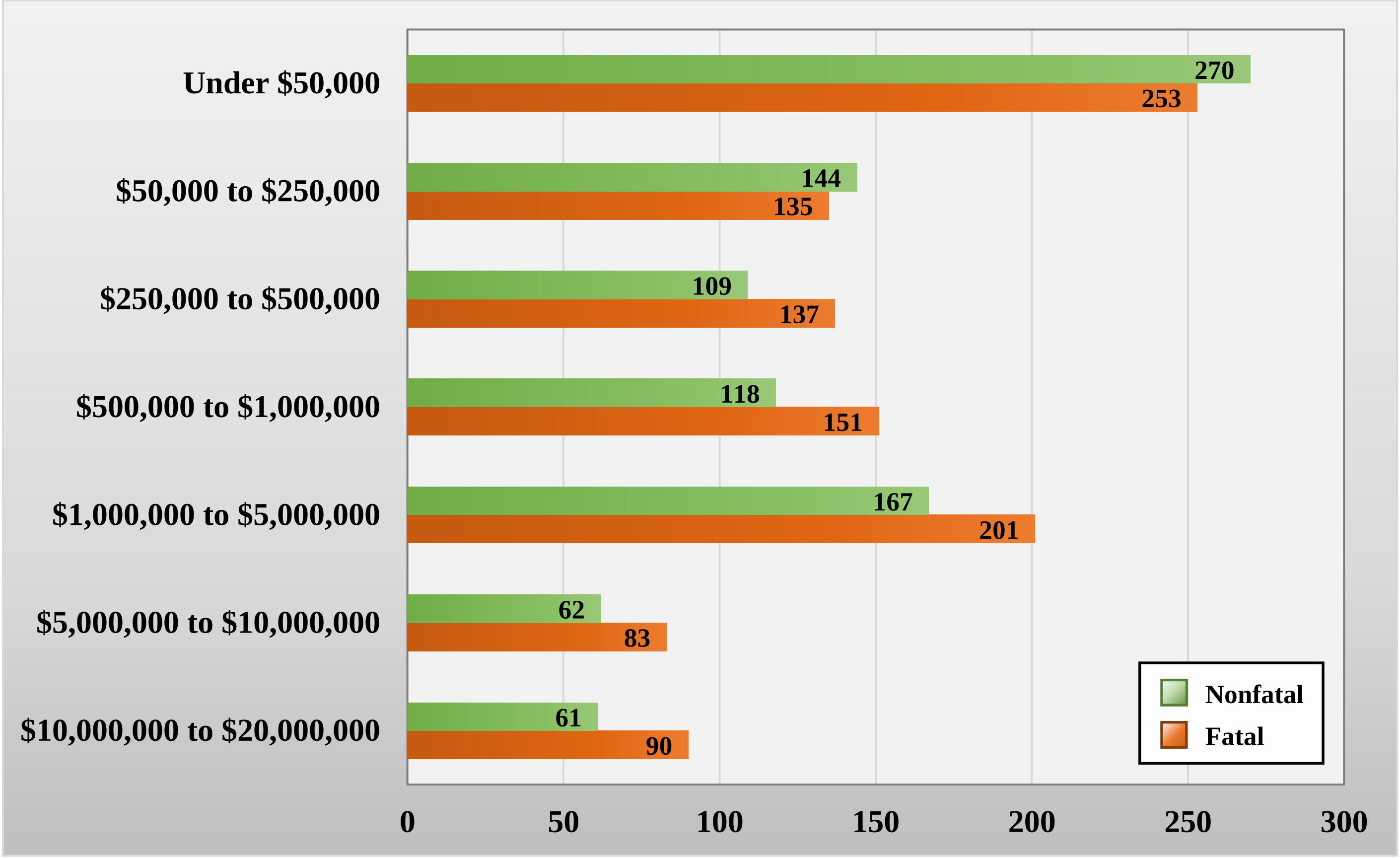


Figure 3.8: Contract amount category counts across ISL

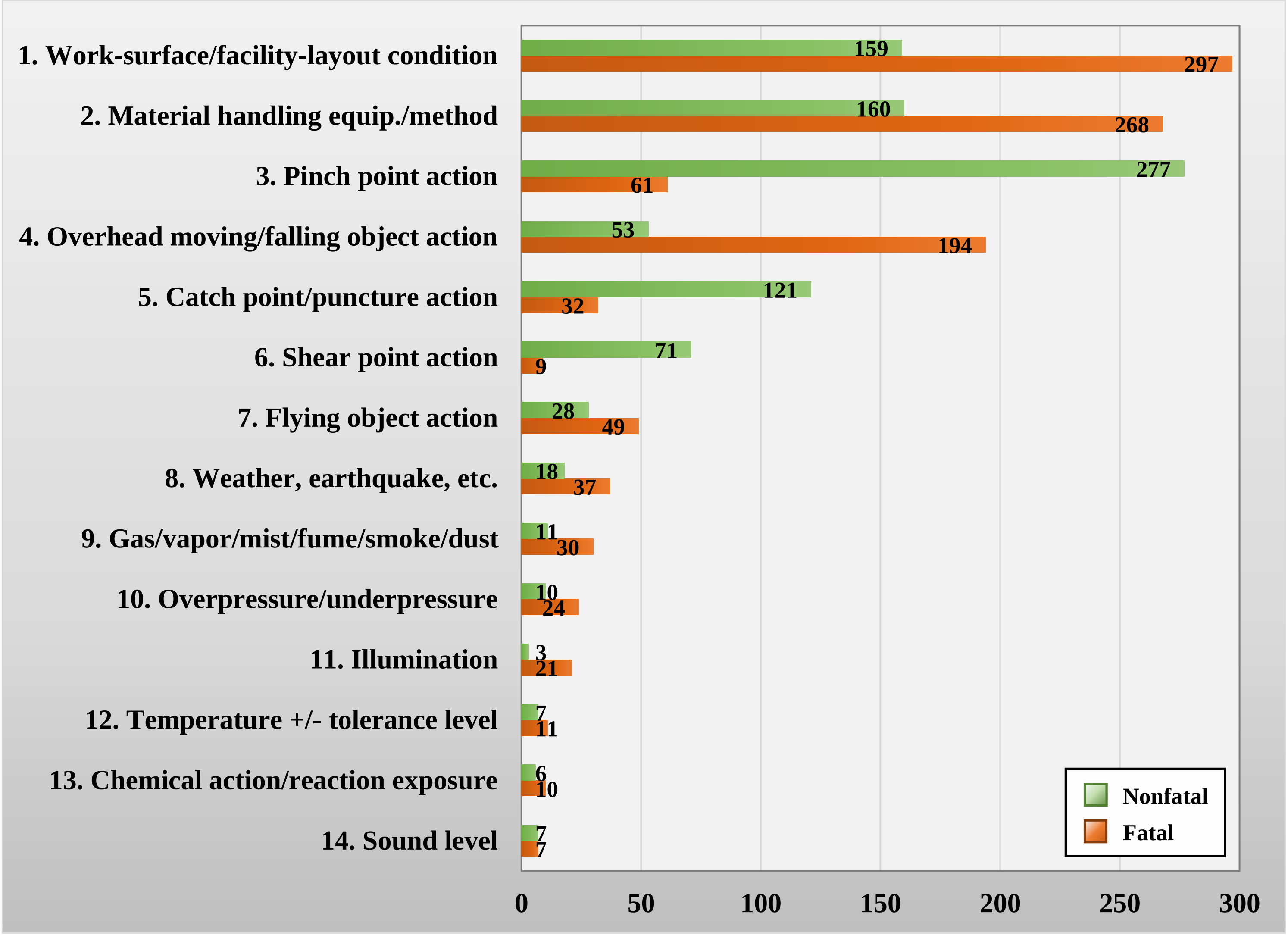


Figure 3.9: Worksite environment category counts across ISL

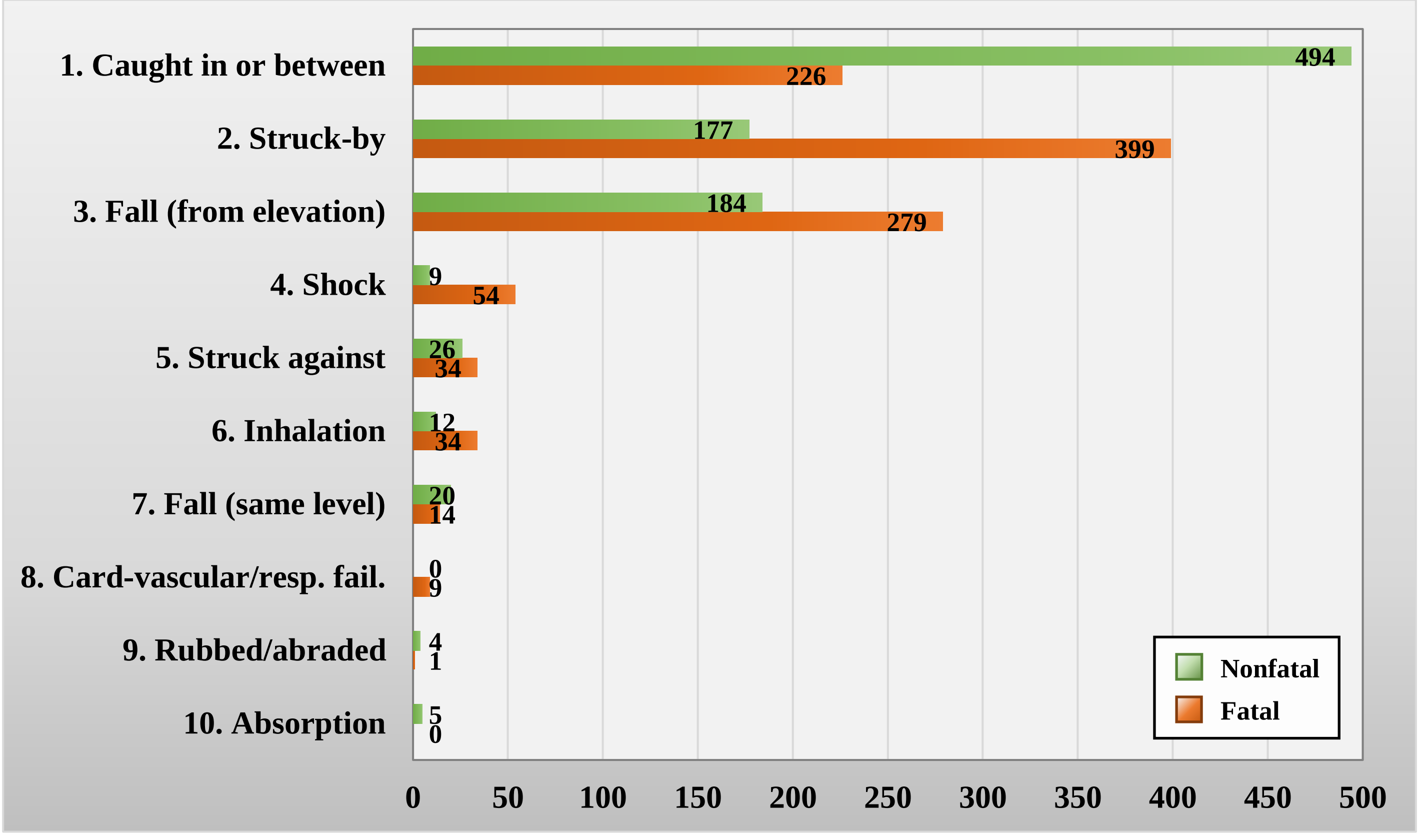


Figure 3.10: Hazard exposure category counts across ISL

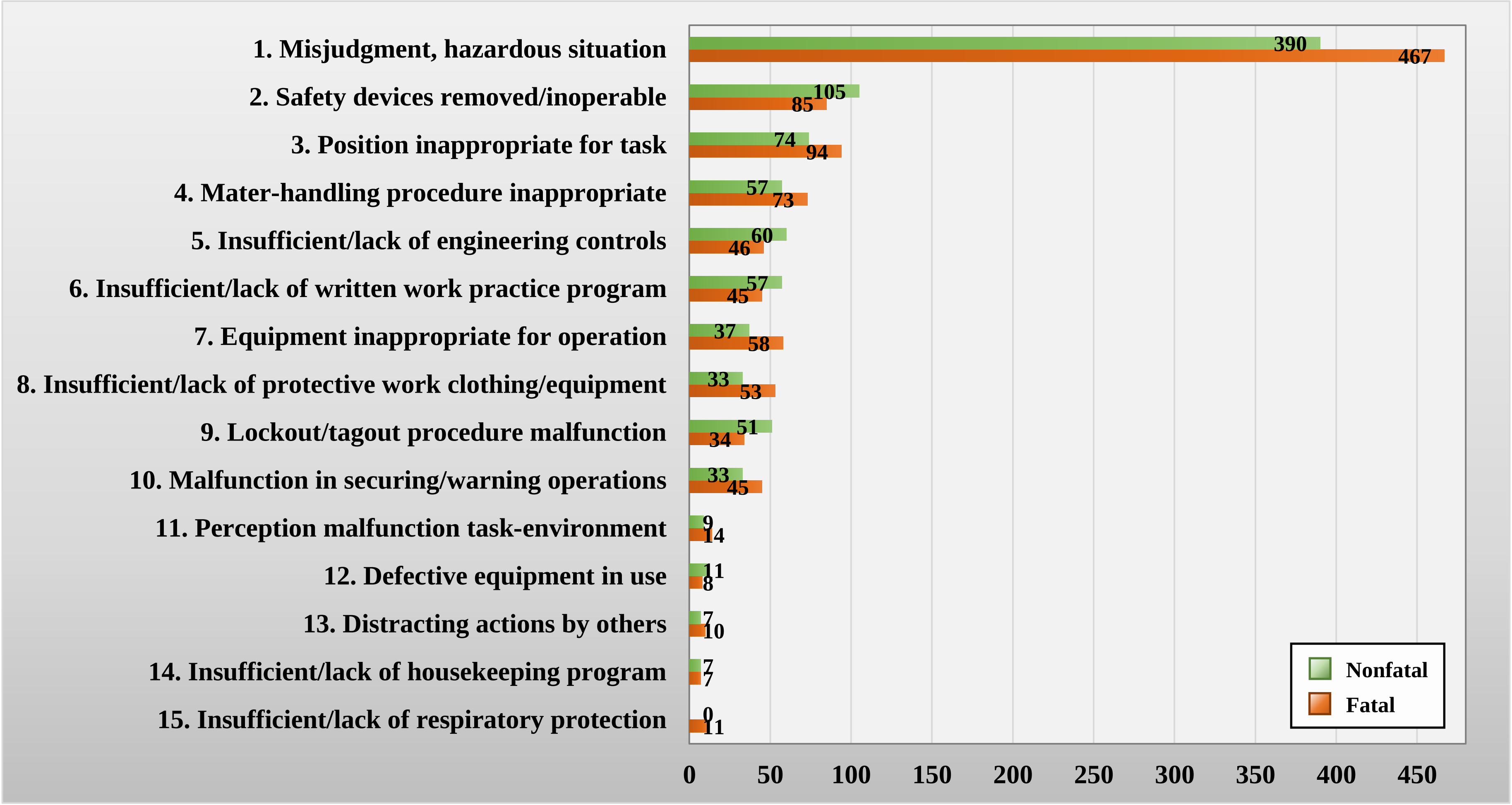


Figure 3.11: Human error category counts across ISL

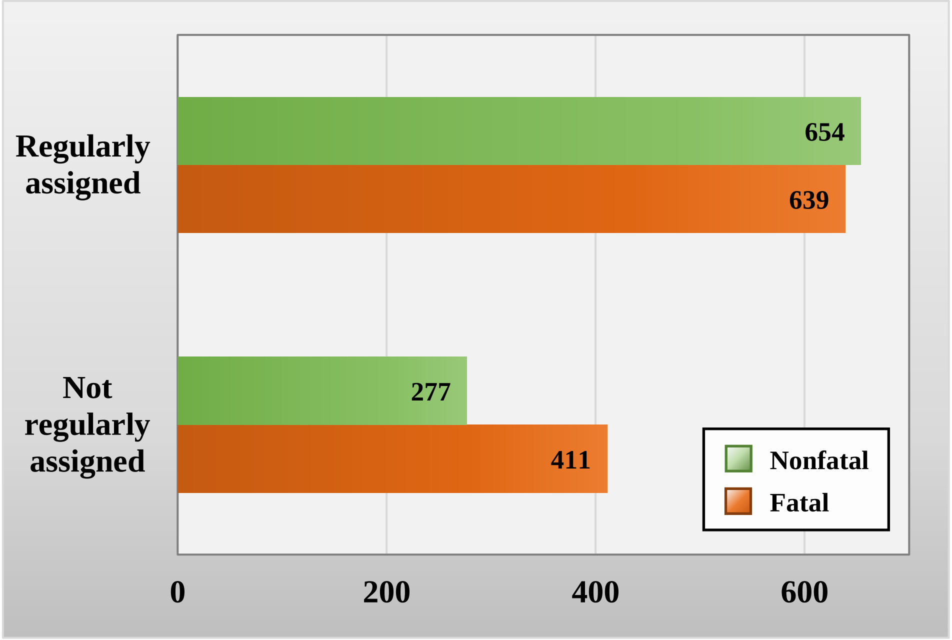


Figure 3.12: Work familiarity category counts across ISL

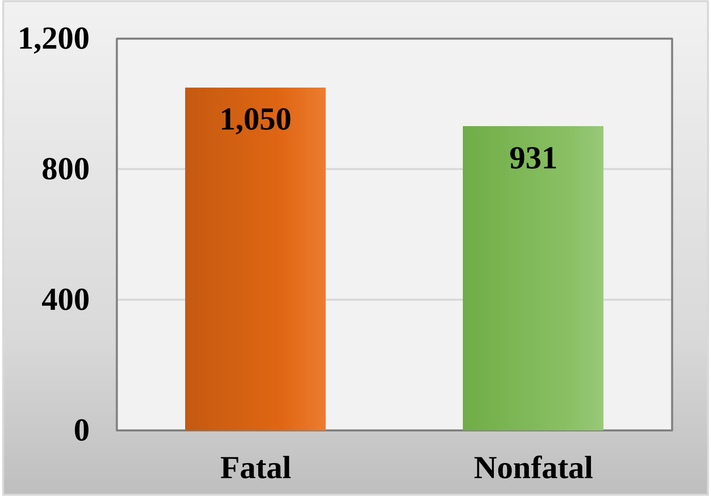


Figure 3.13: ISL class counts

As there were no spatial data included in the reports to match the date for deriving weather data, the *month* was considered as a CIF to represent weather factors. It can be seen from these figures that, although some CIFs have visible effects (i.e., clear distinctions) on ISL outcomes (e.g., worksite environment in Figure 3.9 and hazard exposure in Figure 3.10), other CIFs have less significant effects (i.e., balanced distributions) on ISL (e.g., project type in Figure 3.6). This finding suggests that different CIFs might have varying influences on ISL and thus varying significance toward the prediction procedure, which underlines the need for a CIF importance quantitative ranking.

3.3.2 Top-priority Actions

The IGAE presented in Equations (3.2) to (3.5) was applied to the dataset, where n = 2; i = 1 or 2 (i.e., 1 = nonfatal and 2 = fatal); C = 1,981; C_i = C_1 or C_2 (i.e., C_1= 1,050 and C_2= 931); k = 1 to 8 (e.g., 1 = project type, 2 = project end use, etc.), j is the category for each CIF (e.g., for project type, 1 = new build, 2 = maintenance, etc.), and m is the total number of categories within the k^{th} CIF (e.g., 14 for worksite environment). The information gain $G(C_k)$ results are shown in Figure 3.14. The top three ranked CIFs according to their corresponding $G(C_k)$ values are month (0.356), worksite environment (0.317) and hazard exposure (0.302). This finding is supported by the visualizations presented earlier, where the category distributions of such CIFs over the two ISL classes had visible separations in Figures 3.5, 3.9 and 3.10, respectively.

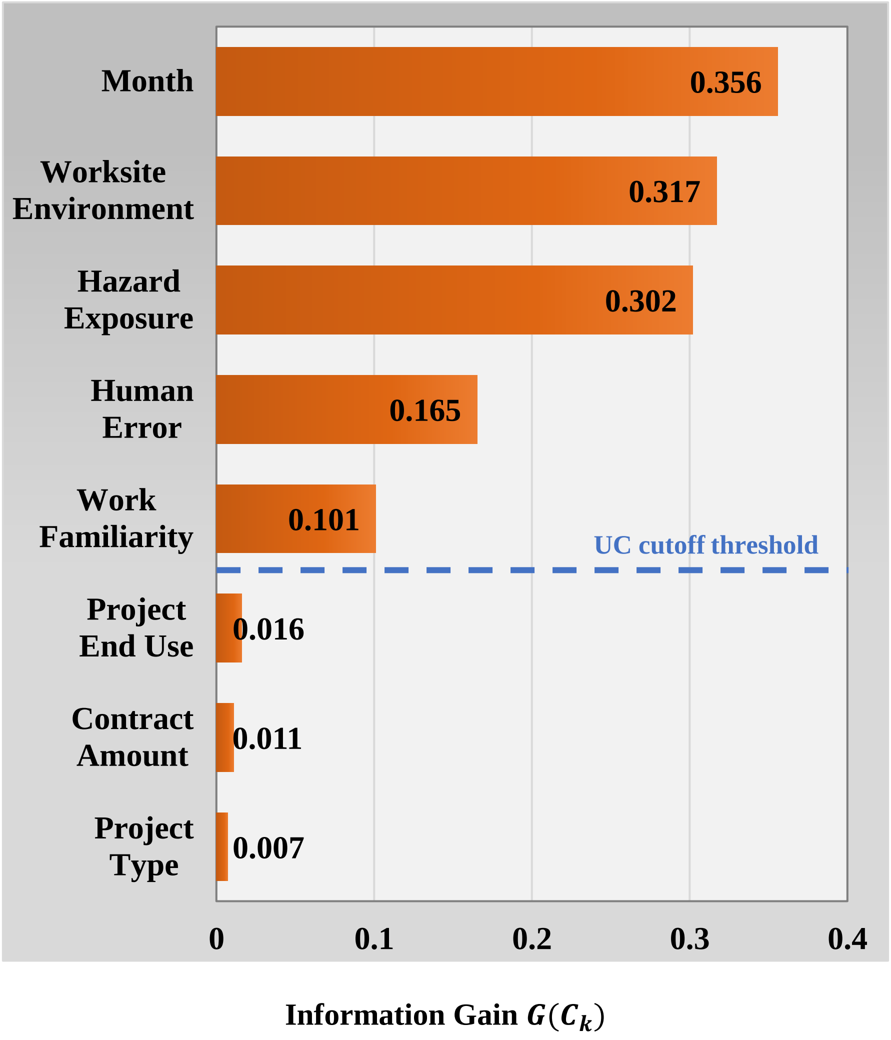


Figure 3.14: Ranking of CIF importance toward ISL

The fourth and fifth CIFs are human error (0.165) and work familiarity (0.101), which is also apparent from more balanced distributions of such CIF categories over the two ISL classes in Figures 3.11 and 3.12, respectively. The lowest ranked CIFs are project end use (0.016), contract amount (0.011) and project type (0.007), indicating these CIFs are the least among others in affecting ISL.

In selecting the key CIFs to proceed with, the average uncertainty coefficient (**UC**) is used as a cutoff threshold to exclude irrelevant CIFs (Desai and Joshi 2010). The UC for each CIF is calculated as the ratio of $G(C_k)$ to $E(C_k)$, and then the CIFs with a UC value greater than the average UC value are selected to proceed with during the subsequent ML analyses. Based on the UC values (see Figure 3.14), the top five CIFs are considered to be key CIFs, namely month, worksite environment, hazard exposure, human error and work familiarity, whereas the bottom three CIFs are excluded from the dataset, namely project end use, contract amount and project type. This selection is also in alignment with several previous studies (Cooper and Phillips 2004; Behm 2005; Choudhry et al. 2009) which highlighted that project-related factors and personal demographics (e.g., the bottom three CIFs) have less significant relationships with injury outcomes compared to safety-related factors and site situational conditions (e.g., the top five CIFs).

Injury prevention begins with having a clear understanding of CIFs that significantly influence safety in construction projects. As such, the described

IGAE procedure can help managers pinpointing those key CIFs to better invest in injury prevention and risk mitigation strategies that are of top priority. This procedure can be useful as a decision support tool in the planning stage of the construction project, especially when a large number of CIFs needs to be considered. In the current application, for example, the procedure pointed to weather factors, worksite environment and hazard exposure as CIFs on which efforts should be focused. Such actionable feedback may guide managers, from the onset, to devise prevention strategies that mitigate risks pertaining to: a) *weather conditions* through appropriate personal protective equipment regulations, emergency weather evacuation planning, and relevant weather safety training; b) *worksite environment* by preparing practical site-specific safety plans, hazardous conditions inspections and equipping workers with knowledge about physical protection in complicated sites and working from heights; and c) *hazard exposure* through job hazard analyses, pre-task safety planning meetings to ensure that hazards are recognized and communicated prior to worker exposure, safety programs for operating equipment, regular equipment maintenance, and emergency response drills.

3.3.3 Model Controls

As previously described, the dataset (1,981 cases) was split into an 80% training set (1,585 cases) and a 20% testing set (396 cases). The decision tree models were developed through a generalizable approach by means of tree parameter optimization through a 10-fold cross-validation procedure applied to the training set. The results for the RPART model are shown in Figure 3.15, where, as previously discussed, higher values of *CP* are more likely to restrict tree growth, while lower values of *CP* are more lenient to node splitting and result in larger tree sizes (i.e., number of terminal nodes). As can be seen in the figure, a *CP* value of 1.41% results in the smallest tree size of one split into two terminal nodes, while a *CP* value of 0.02% allows all splits and thus produces the largest tree possible with a tree size of 28. From the figure, the optimum *CP* value for the current RPART model is 0.27%, resulting in a minimum average cross-validation error of 18.49% and a tree size of 16. Also based on the 10-fold cross-validation results, the *maxdpeth* for the current CART model is limited to an optimum value of five which corresponds to a minimum average cross-validation error of 19.24%. Furthermore, the optimum *minsplit* value for the current C4.5 model is 15, resulting in a minimum average cross-validation error of 19.75%.

The RF model parameters are optimized based on the average OOB error. As shown in Figure 3.16, the range of n_{tree} is set from 5 to 500 with a step size of 1, and values of k_{try} between 2 and 4 (i.e., RF) as well as all 5 (i.e., bagging) variables are also considered—which means that 1,980 iterations of successive parameter combinations, and thus RF models, are evaluated. Based on the results, the optimum combination of parameters is n_{tree} = 315 and k_{try} = 4, resulting in a minimum average OOB error of 15.31%. The figure also shows that using a k_{try}

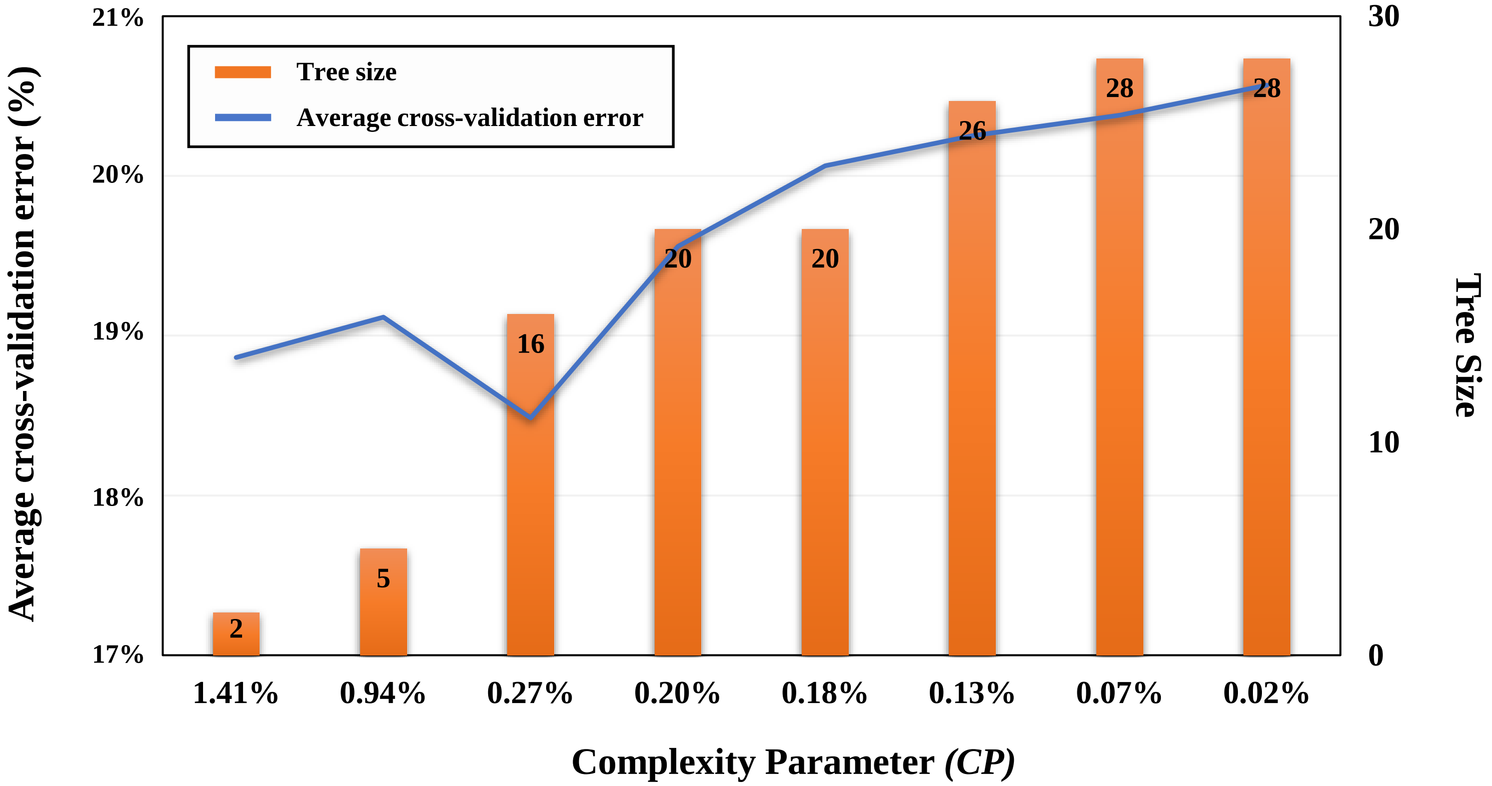


Figure 3.15: Parameter optimization for RPART model – Average cross-validation error under different values of *CP*

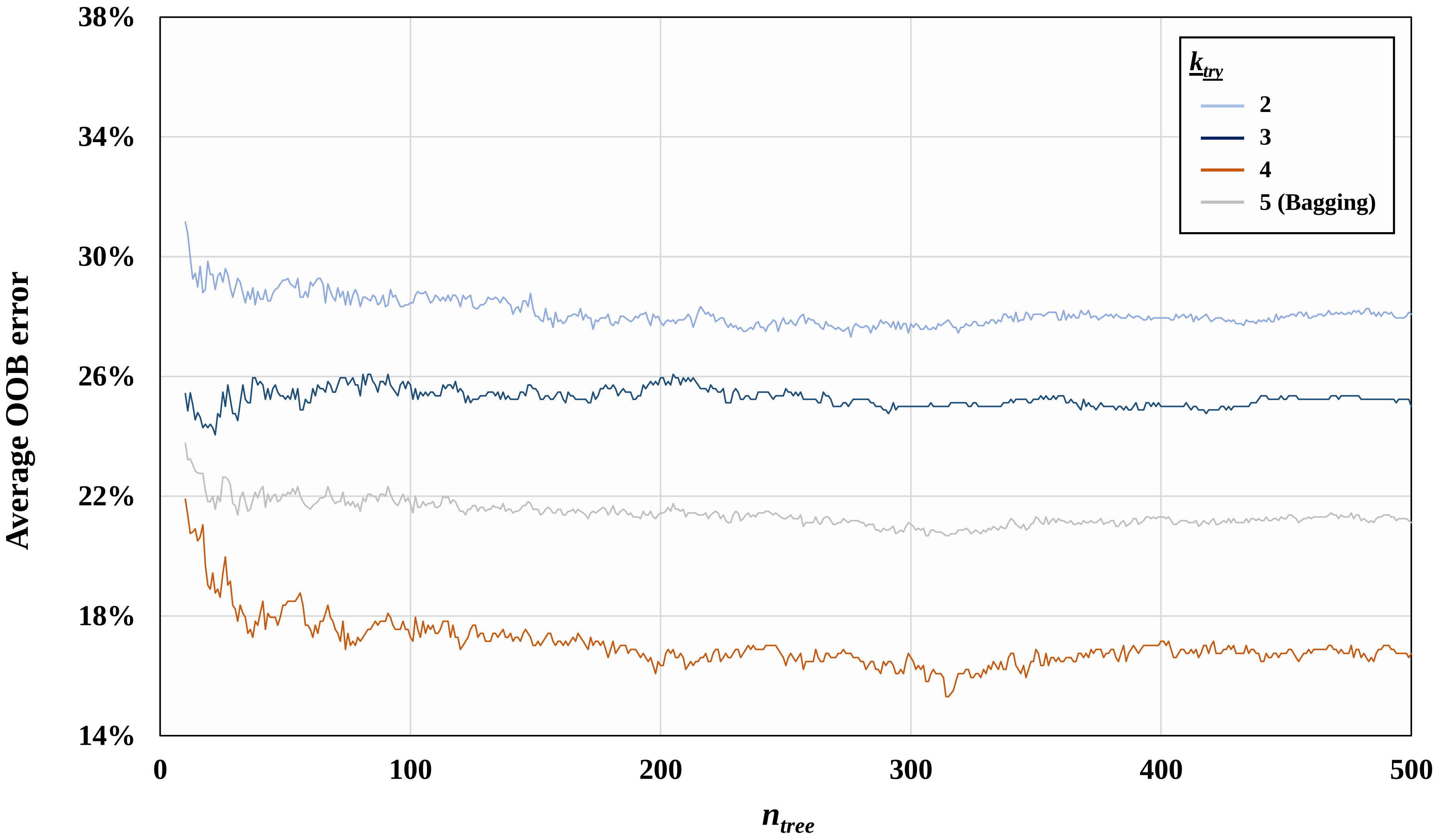


Figure 3.16: Parameter optimization for RF model – Average OOB error under different values of n_{tree} and k_{try}

value of all 5 variables does not provide the least average OOB error, which confirms the need for RF over bagging to inject more randomness in each bootstrap sample, thus producing less correlated trees and a more generalizable model, as previously discussed.

3.3.4 Model Performance Assessments

3.3.4.1 Robustness

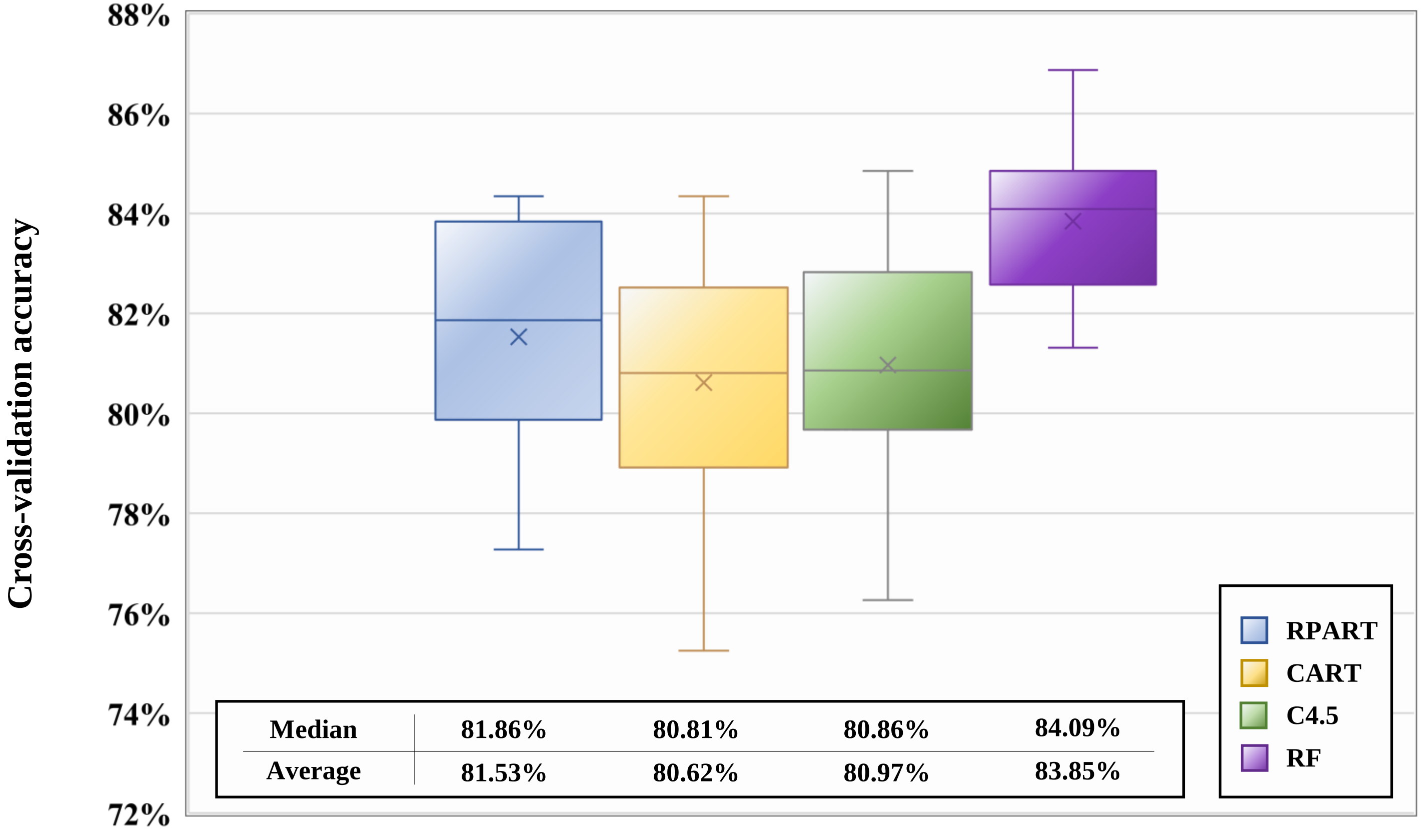
Using the aforementioned optimum parameters for training, the results of the holdout testing for the four models are reported in Table 3.2. The table includes performance evaluation measures of accuracy, precision, TPR, FPR and the average of such measures to establish an overall score for each model. Generally, the four models perform well as the average score of each model is always higher than 80%, which also confirms the reliability of the selected CIFs. Based on the measures in the table, the CART and C4.5 models demonstrate comparable performances with the former slightly outperforming the latter with respect to the average score. However, the RPART model is the best performing decision tree model in every measure including accuracy (81.82%), precision (81.08%), TPR (85.71%), TNR (77.42%) and thus average score (81.51%). As discussed, TPR is especially observed as the predictive performance of the fatal class is particularly important. The RPART model performs well leaving only a 14.29% chance that a fatality (i.e., high-risk zone) will be overlooked as a nonfatality (i.e., low-risk zone). Regarding the RF model, the results of the holdout testing in Table 3.2 indicate that this model outperforms its singular decision tree model counterparts, including the RPART model, in terms of all the performance evaluation measures as accuracy (83.84%), precision (82.59%), TPR (88.10%) and TNR (79.03%), attaining an average score of 83.39%.

Table 3.2: Model comparison results of holdout testing on testing set

Model	Accuracy (%)	Precision (%)	TPR (%)	TNR (%)	Average Score (%)
RPART	81.82%	81.08%	85.71%	77.42%	81.51%
CART	80.56%	80.09%	84.29%	76.34%	80.32%
C4.5	80.30%	80.28%	83.33%	76.88%	80.20%
RF	83.84%	82.59%	88.10%	79.03%	83.39%

3.3.4.2 Reliability

Figure 3.17 demonstrates the results of the 10-fold cross-validation as box plots representing the maximum, minimum, interquartile range, median and average of the 10 resulting accuracy values. Compared to the CART ad C4.5 models, the RPART model’s higher median (81.86%) and average (81.53%) accuracy indicate its better generalization capability among the tree models. In addition, the RF model’s better cross-validation performance is demonstrated through its: 1) highest median and average accuracies of 84.09% and 83.85%, respectively; 2) highest maximum and minimum recorded accuracies which are included between 86.87% and 81.31%, respectively; and 3) smallest max-min range over 10 testing folds of 5.56% which remains more stable than the ranges



Median	81.86%	80.81%	80.86%	84.09%
Average	81.53%	80.62%	80.97%	83.85%

Figure 3.17: Model comparison results of 10-fold cross validation accuracy on entire dataset

pertaining to singular trees (e.g., 7.07% of the RPART model). The latter finding confirms that the RF model not only generalizes better to unseen data but also can achieve consistent performances (i.e., is more reliable) over multiple new datasets.

3.3.4.3 Versatility

The receiver operating characteristic curves for the four models are presented in Figure 3.18. The figure visually shows that the curve for the RF model lies above the rest which, supported by the highest quantified AUC value of 0.91, speaks to the versatility of the RF model under different prediction thresholds and further underlines its ability in avoiding serious types of prediction errors (i.e., FNs).

3.3.5 Model Selection

As can be seen from the results discussed above, while the RF model achieves higher performance, it does not hugely outperform its decision tree counterparts. For instance, the RF model outperforms the RPART model by only 2.02% in accuracy, 2.39% in TPR and 0.03 in AUC. As such, within the current construction safety and injury prediction application, decision tree models can be recommended for use because such models 1) preserve good predictive performance; 2) provide valuable glass-box interpretation-related merits (discussed next); and 3) consume short computation time which is conducive to rapid decision-making.

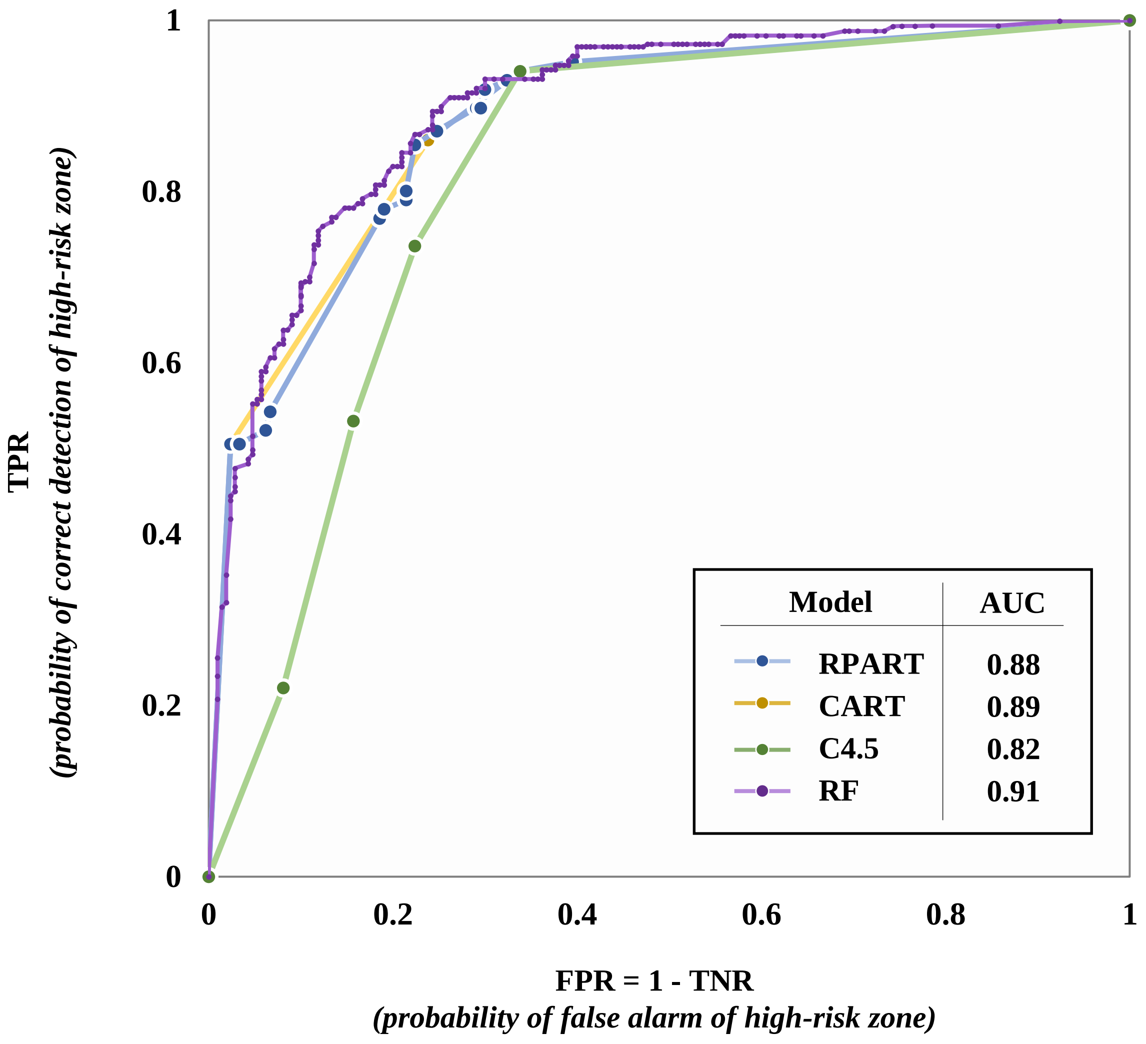


Figure 3.18: Model comparison results of ROC curves and AUC

The reasons behind this observed performance similarity can be discussed in the context of the three criteria/rationale discussed earlier. The first criterion is related to RF models being more robust to unbalanced distributions of the outcome class. Within the current application, the outcome ISL had a relatively balanced distribution (see Figure 3.13) which is why the RF model did not exhibit largely superior performance compared to its tree counterparts. Nonetheless, in dissimilar situations, managers may consider RF modelling. The second criterion is related to how RF models can better handle data with large numbers of input variables. For example, if the injury cases dataset used herein had contained a large number of CIFs describing more features of the work circumstances at the time of the injury, a RF model may be considered. The third criterion is in regards to how RF models may perform better than single trees on data with strong predictor variables alongside other potentially irrelevant variables. In applications where CIFs selection is not performed (unlike the current application), there may exist large differences in the influences of different CIFs on the ISL (e.g., see Figure 3.14), and managers may consider RF modelling in such cases. Despite the closeness in performance between the RPART and RF models within the current application, the gap between the model performances may widen under different safety applications and a RF model may be a more suitable selection in scenarios where considerable improvements in predictive performance can be achieved.

3.3.6 Glass-Box Decision Flow Interpretation

Figure 3.19 shows the developed RPART decision tree model which is selected for demonstration since it is the best performing single tree model in the current study. The tree has a depth of 8 and contains 16 terminal nodes which indicate the predicted ISL class (i.e., fatal or nonfatal) and the corresponding site safety risk level (i.e., high-risk zone or low-risk zone). Since the tree is earlier pruned to achieve better generalizability, the terminal nodes quantify the probabilities of each ISL, rather than providing perfect classifications, and the outcome ISL designation follows the more probable one (i.e., the default prediction threshold is 0.5, as discussed earlier). The tree branches comprise a series of logical decisions as each branch indicates one of the five key CIFs (i.e., month, worksite environment, hazard exposure, human error, or work familiarity) and corresponding categories. Some categories are referenced through numbering consistent with Figures 3.9, 3.10 and 3.11 in order to keep the tree aesthetic and size efficient. Since the IGAE is applied at each node to select CIFs for node splitting, the higher branches of the tree contain the more influential CIFs as recorded in Figure 3.14. For example, the first node in the tree, which contained all the training set, was split using the month CIF since it was the top influential CIF on the same set of data. Furthermore, the category groupings (i.e., months) used for that split were January-April on one branch and May-December on the other since these groupings provided the best child node homogeneity (i.e., the cleanest split in terms of the ISL classes) as can be noted from the class

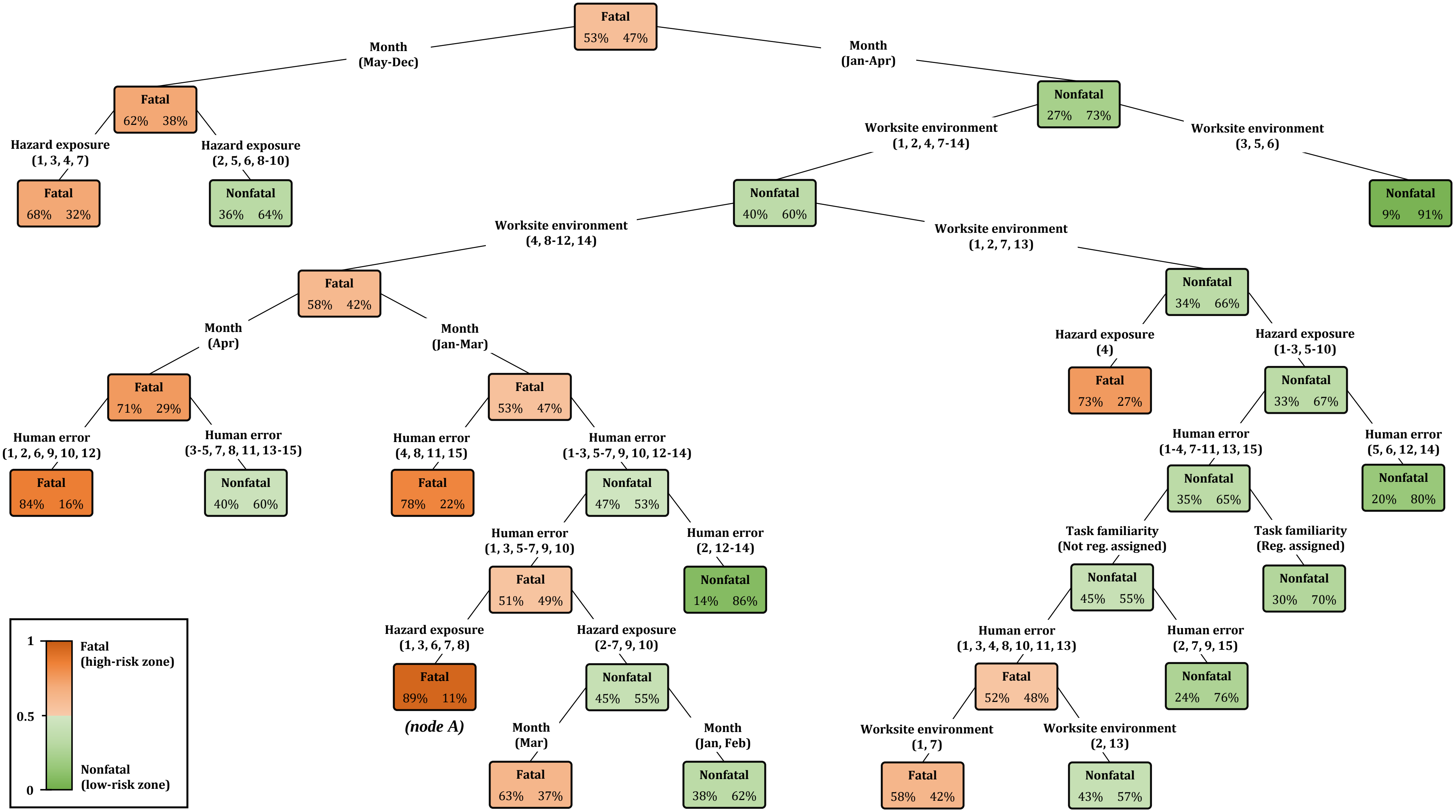


Figure 3.19: RPART decision tree for quantitatively predicting ISL and site safety risk level from qualitatively assessed CIFs

distributions over months in Fig. 3.5. As discussed earlier, such node splitting was repeated until the *CP* criteria was reached, thus yielding the final decision tree model in Figure 3.19. Such a glass-box model provides a transparent and interpretable decision flow structure that can empower safety managers with two key capabilities.

3.3.6.1 Site Risk Level Classification

The *first* is employing the model as a predictive tool that facilitates quantitatively classifying construction sites in accordance to their safety risk levels, thus flagging sites that are of particularly high-risk levels. At early planning stages, managers can qualitatively identify potential CIF categories (e.g., collected from Figures 3.9 to 3.12) for a work package that is scheduled to be executed within a certain site location. This can be achieved through knowledge of the: 1) time of work execution (month); 2) risks associated with the spatial working conditions relevant to the site (worksite environment); 3) nature of work and execution methods relevant to the work package (hazard exposure); and 4) assigned workers' experience and motivation (human error and work familiarity). Using such qualitative CIF categories, managers can utilize the tree to probabilistically quantify the outcome ISL and thus designate a site as a low- or high-risk zone. Armed with such leading safety insights, managers are able to make proactive and better-informed decisions so as any possible risks can be mitigated before reaching the construction site.

3.3.6.2 Cause-and-Effect Relationships

The *second* capability is utilizing the tree for qualitatively exploring intercausal reasoning within the construction injury phenomenon. The tree establishes explicit cause-and-effect relationships by mapping combinations of CIFs (causers) to an ISL (effect) which managers can use to evaluate the interrelatedness among certain CIF categories and their combined contributions toward ISL. For example, one can elaborate by isolating the most risk-prone terminal node in Figure 3.19, which is the sixth terminal node from the left marked as *node A*. The results of following combinations of CIF categories leading to this node infer that, within U.S. construction sites, work during the winter months (January to April) typically leads to worksite environments with difficult weather conditions [e.g., (12) temperature +/- tolerance level; (10) over/under-pressure; (9) gas/vapor/mist/fume/smoke/dust; and (11) illumination] and slippery/unstable work surfaces [e.g., (1) work-surface/facility-layout condition]. Such conditions are strongly connected to hazard exposures related to breathing difficulty [e.g., (6) inhalation; and (8) card-vascular/resp. failure], slips, trips and falls [e.g., (3) from elevation; or (7) same level] and falling objects. These events, coupled with any sort of human error, are almost certain (having an 89% probability) to result in a fatal injury. To minimize such a major source of fatalities, safety managers are encouraged to effectively put forward actions to better plan worksites. These actions should accommodate such working conditions and reduce their accompanying risks before workers are exposed and then forced to react to minimize these risks. Such timely measures can prove essential to protecting construction workers from occasional/seasonal or unexpected accidents. Similar interpretations can be made for CIF combinations funneling through other nodes, and such new knowledge can enhance the understanding of the underlying mechanisms that shape construction safety risk and create injuries.

3.3.7 Model Performance Enhancement

To complement the RPART tree model above, its ROC curve is shown in Figure 3.20. For example, within the current application, the default threshold of 0.5 corresponds to a combination of FPR and TPR values of 22.58% and 85.71%, respectively, as shown in Figure 3.20 and reported in Table 3.2. Managers may opt instead for a threshold of 0.2 and accept a reasonable increase in FPR (to 33.81%) but considerably raise TPR to 94.09%, whereas the TPR improvement due to a threshold of 0.1 does not justify the accompanying large FPR increase. A threshold of 0.2 means that terminal nodes in Figure 3.19 with a fatal ISL probability of 20% or more will be designated as fatal injuries and thus flagged as high-risk zones. With this adjustment, managers would further improve their model’s predictive performance by reducing the more serious types of predictive errors (i.e., FNs) and ultimately amplify the impacts of their managerial actions and prevention strategies.

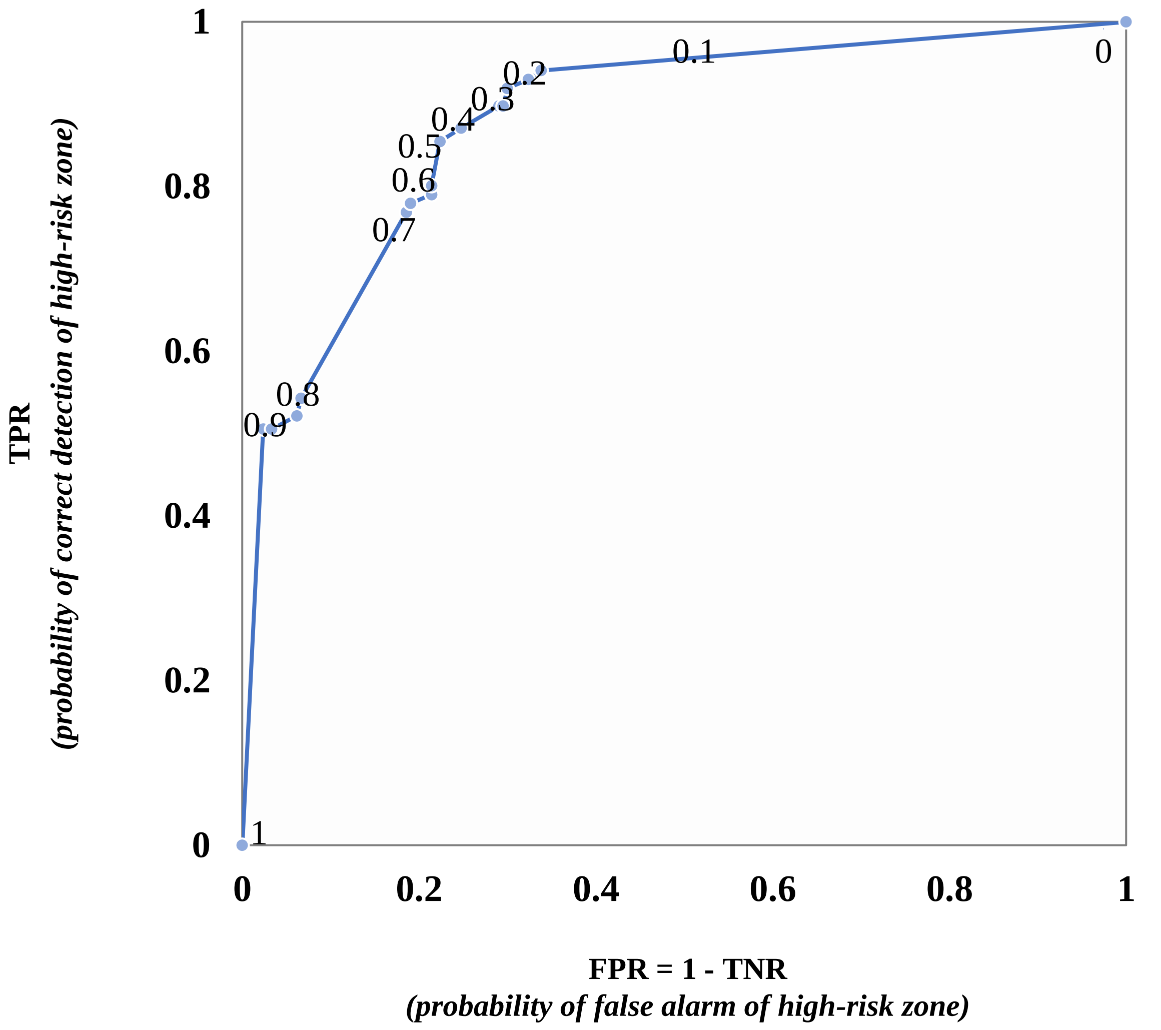


Figure 3.20: Receiver operating characteristic curve with prediction thresholds for RPART model

3.4 CONCLUSIONS

Construction remains one of the most hazardous industries worldwide. Learning from past incidents is key to future injury prediction and prevention. As such, machine-learning-based analyses of empirical data have the potential to transform the way organizations make their safety decisions through more accurate predictions and more effective prevention. In this respect, the current study develops an interpretable-machine learning-based framework for construction injury severity prediction and subsequent risk mitigation.

First, through evaluating the predictive power of different injury factors, a ranking algorithm procedure is utilized to pinpoint the key influential factors which safety managers can use for targeted efforts and priority actions. Subsequently, decision tree and random forest models are developed and optimized, and subsequently employed to quantitively predict injury severity level from the combined and interdependent effects of the key identified injury factors. These quantitative predictions are also supported by interpretable/explainable qualitative insights through leveraging the glass-box merits of such models. The predictive robustness, reliability and versatility performances of the developed models are verified using cross-validation tests and multiple relevant performance evaluation measures. Decision support is also provided to equip managers with knowledge on when to use decision trees or random forest models through a trade-off between interpretability and performance. Finally, receiver operating characteristics insights are presented to aid in further adjusting model prediction thresholds in order to improve overall model performance according to unique safety applications and requirements.

A demonstration application using the OSHA injury cases dataset is subsequently presented to showcase how the decision support framework can be utilized to provide safety managers with the following key insights: i) Awareness of the key influential construction injury factors at specific sites such that targeted efforts and top-priority preventative strategies can be deployed; ii) Guidance on identifying high-risk sites so that hazards can be eliminated proactively, thus enhancing workplace safety; iii) Qualitative exploration of the underlying interdependence between injury factors as well as their interactive cause-and-effect relationships with injury severity (i.e., root causes of injuries); iv) Decision support on model selection based on a trade-off between interpretability and performance; and v) Deepened understanding of machine learning model mechanisms and outputs for ultimately selecting prediction thresholds that can improve model performance. Ultimately, through the ability to better understand, predict, and prevent the occurrence of construction injuries, the framework developed and described herein should empower safety managers and workplace safety practitioners with better-informed and safer decision-making that would foster safer sites and save lives.

3.5 ACKNOWLEDGMENTS

The authors are grateful to the financial support of the Natural Science and Engineering Research Council of Canada (NSERC) CaNRisk-CREATE program, and the INTERFACE Institute and the INViSiONLab, both of McMaster University.

3.6 REFERENCES

Abdel-Rahman, E. M., Ahmed, F. B., & Ismail, R. (2013). Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. International Journal of Remote Sensing, 34(2), 712-728.

Aggarwal, C. C. (2015). Data mining: the textbook. Springer.

Akhavian, R., & Behzadan, A. H. (2013). Knowledge-based simulation modeling of construction fleet operations using multimodal-process data mining. Journal of Construction Engineering and Management, 139(11), 04013021.

Alkaissy, M., Arashpour, M., Ashuri, B., Bai, Y., & Hosseini, R. (2020). Safety management in construction: 20 years of risk modeling. Safety science, 129, 104805.Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. Statistics surveys, 4, 40-79.

Behm, M. (2005). Linking construction fatalities to the design for construction safety concept. Safety science, 43(8), 589-611.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. Journal of machine learning research, 13(Feb), 281-305.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth. International Group, 432, 151-166.

Bureau of Economic Analysis (BEA) (2021). Employment by NAICS Industry. Retrieved July 27, 2021 from https://www.bea.gov/data/employment/employment-by-industry.

Bureau of Labor Statistics (2021a). Census of Fatal Occupational Injuries (CFOI). Retrieved July 27, 2021 from https://www.bls.gov/iif/oshcfoi1.htm.

Bureau of Labor Statistics (2021b). Survey of Occupational Injuries and Illnesses Data. Retrieved July 27, 2021 from https://www.bls.gov/iif/soii-data.htm#dafw.

Chen, Q., & Jin, R. (2013). Multilevel safety culture and climate survey for assessing new safety program. Journal of Construction Engineering and Management, 139(7), 805-817.

Chi, S., Suk, S. J., Kang, Y., & Mulva, S. P. (2012). Development of a data mining-based analysis framework for multi-attribute construction project information. Advanced Engineering Informatics, 26(3), 574-581.

Chi, S., Suk, S. J., Kang, Y., & Mulva, S. P. (2012). Development of a data mining-based analysis framework for multi-attribute construction project information. Advanced Engineering Informatics, 26(3), 574-581.

Chou, J. S., & Lin, C. (2013). Predicting disputes in public-private partnership projects: Classification and ensemble models. Journal of Computing in Civil Engineering, 27(1), 51-60.

Choudhry, R. M., Fang, D., & Lingard, H. (2009). Measuring safety climate of a construction company. Journal of construction Engineering and Management, 135(9), 890-899.

Cooper, M. D., & Phillips, R. A. (2004). Exploratory analysis of the safety climate and safety behavior relationship. Journal of safety research, 35(5), 497-512.

CPWR - The Center for Construction Research, and Training. (2018). The construction chart book: The US construction industry and its workers.

Dedobbeleer, N., & Béland, F. (1991). A safety climate measure for construction sites. Journal of safety research, 22(2), 97-103.

Desai, V. S., & Joshi, S. (2010). Application of decision tree technique to analyze construction project data. In International Conference on Information Systems, Technology and Management (304-313). Springer, Berlin, Heidelberg.

Fang, D., Chen, Y., & Wong, L. (2006). Safety climate in construction industry: A case study in Hong Kong. Journal of construction engineering and management, 132(6), 573-584.

Feng, Y., Teo, E. A. L., Ling, F. Y. Y., & Low, S. P. (2014). Exploring the interactive effects of safety investments, safety culture and project hazard on safety performance: An empirical analysis. International Journal of Project Management, 32(6), 932-943.

Fiore, A., Quaranta, G., Marano, G. C., & Monti, G. (2016). Evolutionary polynomial regression–based statistical determination of the shear capacity equation for reinforced concrete beams without stirrups. Journal of Computing in Civil Engineering, 30(1), 04014111.

Gerassis, S., Martín, J. E., García, J. T., Saavedra, A., & Taboada, J. (2017). Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. Journal of construction engineering and management, 143(2), 04016093.

Glendon, A. I., & Litherland, D. K. (2001). Safety climate factors, group differences and safety behaviour in road construction. Safety science, 39(3), 157-188.

Gondia, A., Ezzeldin, M., & El-Dakhakhni, W. (2020). Forthcoming. Mechanics-guided Genetic Programming Expression for Shear Strength Prediction of Squat Reinforced Concrete Walls with Boundary Elements. Journal of Structural Engineering.

Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2019). Machine Learning Algorithms for Construction Projects Delay Risk Prediction. Journal of Construction Engineering and Management, 146(1), 04019085.

Hallowell, M. R., Hinze, J. W., Baud, K. C., & Wehle, A. (2013). Proactive construction safety control: Measuring, monitoring, and responding to safety leading indicators. Journal of construction engineering and management, 139(10), 04013010.

Han, T., Jiang, D., Zhao, Q., Wang, L., & Yin, K. (2018). Comparison of random forest, artificial neural networks and support vector machine for intelligent diagnosis of rotating machinery. Transactions of the Institute of Measurement and Control, 40(8), 2681-2693.

Hong, H., Tsangaratos, P., Ilia, I., Chen, W., & Xu, C. (2017). Comparing the performance of a logistic regression and a random forest model in landslide susceptibility assessments. The Case of Wuyaun Area, China. In Workshop on world landslide forum (1043-1050). Springer, Cham.

Hornik, K., Buchta, C., & Zeileis, A. (2009). Open-source machine learning: R meets Weka. Computational Statistics, 24(2), 225-232.

Huang, L., Wu, C., Wang, B., & Ouyang, Q. (2018). Big-data-driven safety decision-making: a conceptual framework and its influencing factors. Safety science, 109, 46-56.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, 3-7). New York: springer.

Jin, R., Zou, P. X., Piroozfar, P., Wood, H., Yang, Y., Yan, L., & Han, Y. (2019). A science mapping approach based review of construction safety research. Safety science, 113, 285-297.

Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. Safety science, 117, 257-262.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, 1137-1145).

Kuhn, M., & Johnson, K. (2013). Measuring performance in classification models. In Applied predictive modeling (247-273). Springer, New York, NY.

Li, Z., Liu, P., Wang, W., & Xu, C. (2012). Using support vector machine models for crash injury severity analysis. Accident Analysis & Prevention, 45, 478-486.

Liu, X., Song, Y., Yi, W., Wang, X., & Zhu, J. (2018). Comparing the random forest with the generalized additive model to evaluate the impacts of outdoor ambient environmental factors on scaffolding construction productivity. Journal of Construction Engineering and Management, 144(6), 04018037.

Marin, L. S., Lipscomb, H., Cifuentes, M., & Punnett, L. (2019). Perceptions of safety climate across construction personnel: Associations with injury rates. Safety science, 118, 487-496.

Mattila, M., Hyttinen, M., & Rantanen, E. (1994). Effective supervisory behaviour and safety at the building site. International Journal of Industrial Ergonomics, 13(2), 85-93.

McDonald, N., Corrigan, S., Daly, C., & Cromie, S. (2000). Safety management systems and safety culture in aircraft maintenance organisations. Safety Science, 34(1-3), 151-176.

Mearns, K., Whitaker, S. M., & Flin, R. (2003). Safety climate, safety management practice and safety performance in offshore environments. Safety science, 41(8), 641-680.

Mohamed, S. (2002). Safety climate in construction site environments. Journal of construction engineering and management, 128(5), 375-384.

Mohammadi, A., Tavakolan, M., & Khosravi, Y. (2018). Factors influencing safety performance on construction projects: A review. Safety science, 109, 382-397.

Occupational Safety and Health Administration (OSHA) (2019). Injuries Illnesses, and Fatalities. Retrieved February 17, 2019 from https://www.bls.gov/iif/oshoiics.htm.

Occupational Safety and Health Administration (OSHA) (2021). Recommended Practices for Safety and Health Programs. Retrieved April 6, 2021 from https://www.osha.gov/safety-management.

Patel, D. A., & Jha, K. N. (2015). Neural network approach for safety climate prediction. Journal of management in engineering, 31(6), 05014027.

Pereira, E., Ahn, S., Han, S., & Abourizk, S. (2018). Identification and association of high-priority safety management system factors and accident precursors for proactive safety assessment and control. Journal of Management in Engineering, 34(1), 04017041.

Pereira, E., Ahn, S., Han, S., & Abourizk, S. (2020). Finding causal paths between safety management system factors and accident precursors. Journal of Management in Engineering, 36(2), 04019049.

Ripley, B. (2018). “Classification and regression trees: R package version 1.0-39.” Accessed August 10, 2018. https://CRAN.R-project.org /package=tree.

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93-104.

Rodriguez-Galiano, V., Mendes, M. P., Garcia-Soldado, M. J., Chica-Olmo, M., & Ribeiro, L. (2014). Predictive modeling of groundwater nitrate pollution using Random Forest and multisource variables related to intrinsic and specific vulnerability: A case study in an agricultural setting (Southern Spain). Science of the Total Environment, 476, 189-206.

Sakhakarmi, S., Park, J., & Cho, C. (2019). Enhanced machine learning classification accuracy for scaffolding safety using increased features. Journal of construction engineering and management, 145(2), 04018133.

Seong, H., Son, H., & Kim, C. (2018). A comparative study of machine learning classification for color-based safety vest detection on construction-site images. KSCE Journal of Civil Engineering, 22(11), 4254-4262.

Shannon, C. E. (1948). A mathematical theory of communication, Part I, Part II. Bell Syst. Tech. J., 27, 623-656.

Siam, A., Ezzeldin, M., & El-Dakhakhni, W. (2019). Machine learning algorithms for structural performance classifications and predictions: Application to reinforced masonry shear walls. Structures,22, 252-265.

Son, H., Kim, C., Hwang, N., Kim, C., & Kang, Y. (2014). Classification of major construction materials in construction environments using ensemble classifiers. Advanced Engineering Informatics, 28(1), 1-10.

Su, Y., Mao, C., Jiang, R., Liu, G., & Wang, J. (2021). Data-Driven Fire Safety Management at Building Construction Sites: Leveraging CNN. Journal of Management in Engineering, 37(2), 04020108.

Sutton, C. D. (2005). Classification and regression trees, bagging, and boosting. Handbook of statistics, 24, 303-329.

Therneau, T., and B. Atkinson. (2018). "Recursive partitioning and regression trees: R package version 4.1-13." Accessed August 10, 2018. https://CRAN.R-project.org/package=rpart.

Trading Economics (2020). United States GDP From Construction. Retrieved May 8, 2020 from https://tradingeconomics.com/united-states/gdp-from-construction.

Wang, W., Jiang, X., Xia, S., & Cao, Q. (2010). Incident tree model and incident tree analysis method for quantified risk assessment: an in-depth accident study in traffic operation. Safety Science, 48(10), 1248-1262.

Zhang, Y., Javanmardi, A., Liu, Y., Yang, S., Yu, X., Hsiang, S. M., ... & Liu, M. (2020). How Does Experience with Delay Shape Managers' Making-Do Decision: Random Forest Approach. Journal of Management in Engineering, 36(4), 04020030.

Zhou, J., Li, X., & Mitri, H. S. (2016). Classification of rockburst in underground projects: comparison of ten supervised learning methods. Journal of Computing in Civil Engineering, 30(5), 04016003.

Zhou, Y., Li, S., Zhou, C., & Luo, H. (2019). Intelligent approach based on random forest for safety risk prediction of deep foundation pit in subway stations. Journal of Computing in Civil Engineering, 33(1), 05018004.

Zohar, D. (1980). Safety climate in industrial organizations: theoretical and applied implications. Journal of applied psychology, 65(1), 96.

Zohar, D. (2002). The effects of leadership dimensions, safety climate, and assigned priorities on minor injuries in work groups. Journal of organizational behavior, 23(1), 75-92.

Chapter 4:

MACHINE LEARNING-BASED CONSTRUCTION SITE RISK MODELS

ABSTRACT

In the last decade, injury statistics have not exhibited significant improvement within the construction industry. Current management strategies are typically deployed reactively in response to safety lagging indicators (e.g., injury rates, lost workdays, and post-incident inspections). This situation suggests that proactive safety management approaches that rely on leading indicators for key decision-making need to be developed. In this respect, the current study is aimed at developing site risk models that generate predictions of safety risk leading indicators across different zones and over project lifecycles. Such leading indicators can be used to proactively anticipate worksite risks such that preventive measures are implemented in advance and can also be adjusted in real-time as projects progress to dynamically monitor and enhance safety performance. The developed models are driven by ensemble machine learning algorithms trained using previous injury precursors and outcomes. Specifically, the ensemble algorithms consider five base algorithms which are subsequently tuned and validated: naïve Bayes, decision trees, random forests, support vector machines, and artificial neural networks. A demonstration application is also presented herein, where the ensemble algorithms are employed to develop a risk model that

forecasts leading indicators of site-specific risk levels, including financial implications of potential injuries and most likely affected body parts. Such a model, which can be extended to other safety-related settings, can have a tangible impact on construction worksite safety through transforming datasets of historical incidents and injuries into actionable insights that support proactive and real-time safety management and decision-making within the construction industry.

4.1 INTRODUCTION

4.1.1 Background

The construction industry is constantly ranked as one of the most unsafe industries globally (Behm 2005; Choudhry et al. 2008; Choudhry et al. 2009; Hallowell et al. 2013; Feng et al. 2014; Jin et al. 2019; Marin et al. 2019). For example, Figure 4.1 demonstrates that the U.S. construction industry in 2019 accounted for more than 200,000 worksite injuries, of which approximately 80,000 were injuries that resulted in days away from work (Bureau of Labor Statistics 2021). The figure also shows that construction garnered the third highest injury rate (i.e., 1.1 injuries with days away from work per 100 workers) among all major industries in the U.S. that year (Bureau of Labor Statistics 2021). However, perhaps what is most concerning is that, in the last decade, the overall trends of injuries have remained remarkably constant, where a flat trend/plateau of injury rates can be inferred in Figure 4.2 through the past 10 years (i.e., hovering between 1.5 and 1.1 injuries per 100 workers per year) (CPWR 2018; Bureau of Labor Statistics 2021). Such a safety plateau has been observed/reported not only in the U.S. but also in several other countries/regions such as Europe (Lander et al. 2016; Misiurek and Misiurek 2017), Asia (Mohammadi et al. 2018; Kang and Ryu 2019) and North America (Tixier et al. 2016a, b; Chen et al. 2018). This situation suggests that traditional safety strategies by governments, industry and academia may have reached the end of

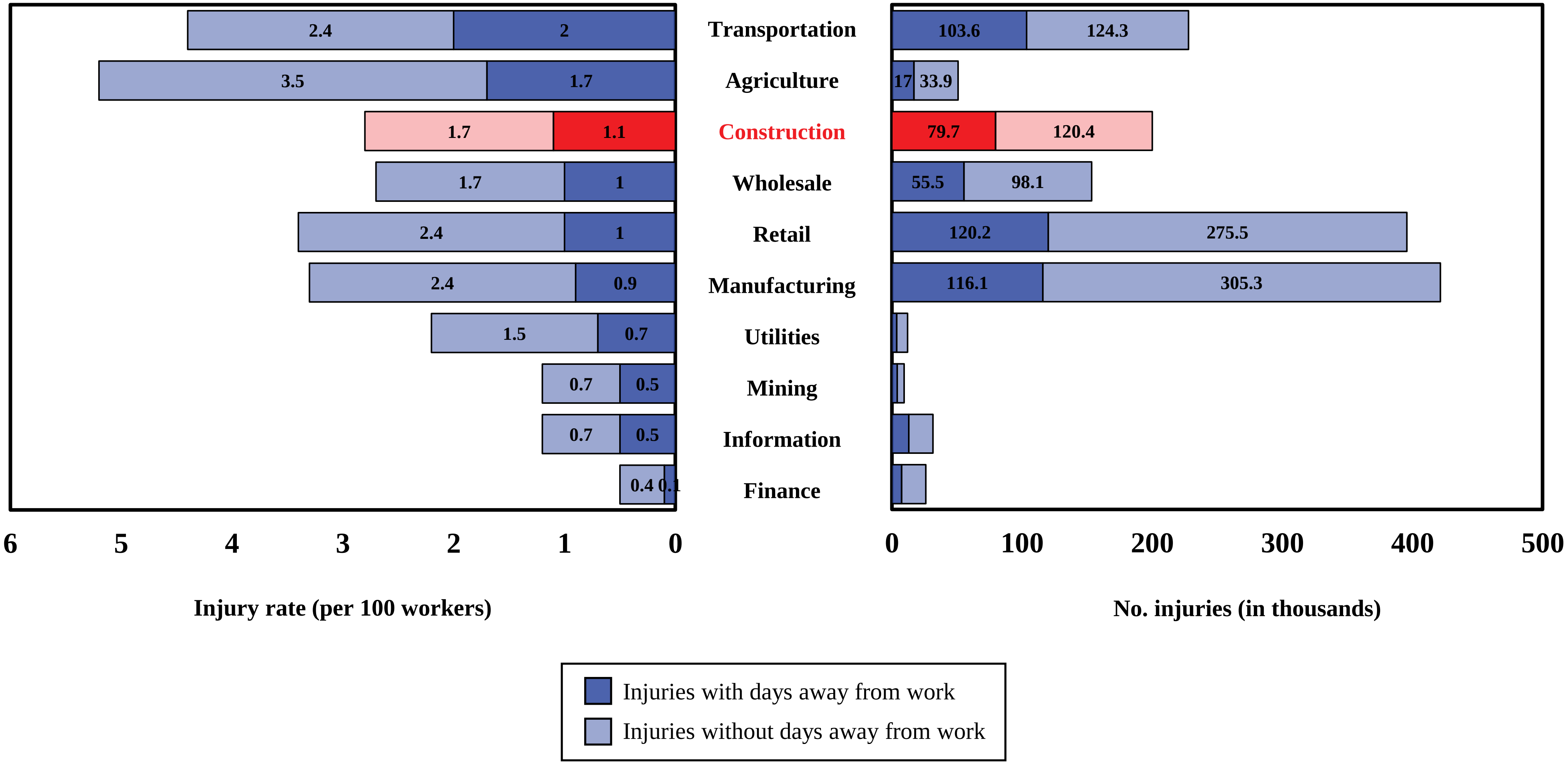


Figure 4.1: U.S. occupational injury rates and numbers by major industry, 2019 (based on data from Bureau of Labor Statistics, 2021)

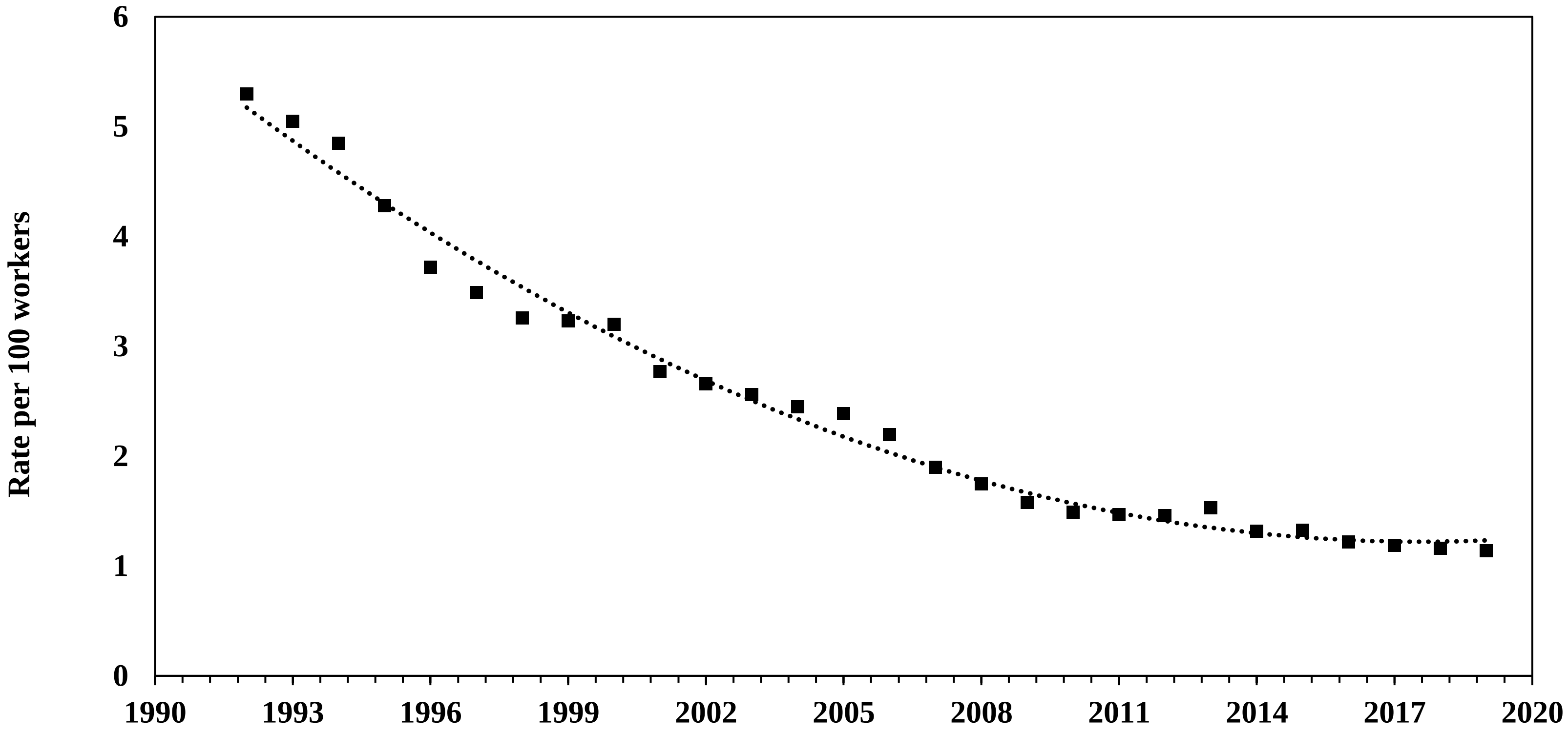


Figure 4.2: Rates of U.S. construction occupational injuries resulting in days away from work, 1992-2019 (based on data from Bureau of Labor Statistics, 2021)

their abilities (Esmaeili and Hallowell 2012; Tixier et al. 2016b), which highlights the need to continue to develop innovative approaches to break through this plateau.

Where reported, the safety plateau has been attributed to two key reasons. First, safety has traditionally been measured and managed *reactively*. For example, reactive/lagging indicators, such as injury rates (e.g., Figure 4.1), lost workdays and experience modification ratings, which are essentially outcome-based measures, are continuously monitored and have long been relied on to measure safety performance, evaluate past safety management strategies and indicate progress (Chen and Jin 2013). Other lagging indicators are post-incident inspections and incident learning analysis, which support management actions that are responsive to a worker being injured, a new standard or regulation being published, or an inspection finding a problem that must be fixed (Hallowell et al. 2013; OSHA 2021). Alternatively, identifying hazards before they cause injuries is a far more effective approach, which is why proactive safety risk leading indicators and management strategies are still needed and should occur from the frontend planning phases of projects (Zou 2011). Second, traditional safety strategies are *neither real-time nor dynamic*. Different sets of conditions, and thus risks/injury precursors, exist with each spatial or temporal change in a project (Sacks et al. 2009; Villanova 2014). It is therefore important to employ strategies that can dynamically monitor and improve safety performance such as models that can allow for adjustments as the project progresses in order to reflect the current/real-time state of a project (Tixier et al. 2017; Chen et al. 2018). However, traditional strategies do not translate well to different work scenarios, thus preventing the efficient capture of the dynamic nature of construction work (Tixier et al. 2016b). As such, with proactive, real-time and dynamic capabilities lacking from traditional strategies, such strategies may be limited when used to compete in everchanging/dynamic construction worksite environments and to meet the proactive and real-time demands for effective safety decision making.

Proactive real-time strategies can be based on predictive algorithms to forecast injury outcomes and safety risks (Poh et al. 2018). Predictive algorithms have been gaining increasing attention in safety research for the past decade, as will be explained later in more detail. Such algorithms can learn from measures of injury precursors/root causes as worksite-, work method- or worker-related hazards or conditions that are relevant to worksite safety and injury outcomes (Guo and You 2016). Products of such predictive algorithms are proactive/leading indicators that can facilitate identifying and controlling potential hazards before they result in injuries in the worksite (Hallowell et al. 2013).

4.1.2 Related Work

Recent research suggests that we have entered into the so-called *third wave* of construction safety management research which harnesses intelligent systems and emerging technologies (Jin et al. 2019; Niu et al. 2019). This follows the *first wave* characterized by a reliance on hard protection as personal protective equipment as a physical buffer between workers and hazards (Zohar et al. 1980; Dedobbeleer et al. 1991; Glendon and Litherland 2001); and the *second wave* of studies on safety climate and culture and their impacts on workers' safety perceptions and behaviors and on promoting safety and health in the construction team (Mohamed et al. 2002; Mearns et al. 2003; Cooper et al. 2004; Fang et al. 2006; Choudhry et al. 2007; Choudhry et al. 2009; Chen et al. 2013; Feng et al. 2014).

The mainstream directions of the third wave of construction safety management research can be represented by four clusters: 1) ***Information technology applications*** (Moon et al. 2014a, b; Shen and Marks 2016; Martínez-Aires et al. 2018; Akram et al. 2019); 2) ***Computer vision applications*** (Park and Kim 2013; Sacks et al. 2013; Sacks et al. 2015; Le et al. 2015; Wang et al. 2018); 3) ***Wearable sensing technologies applications*** (Hwang et al. 2016; Yang et al. 2016; Guo et al. 2017; Nath et al. 2017; Ahn et al. 2019); and 4) ***Data-driven (analytics) applications*** (Salas and Hallowell 2016; Amiri et al. 2017; Goh and Ubeynarayana 2017; Zou et al. 2017; Kim and Chi 2019) which is more related to the current study focus. This fourth cluster has been inspired by the large amounts of structured, semi-structured and unstructured heterogeneous safety-related data continuously collected from construction sites, as worksite, work method and/or worker descriptors upon accident/injury occurrence—data characteristics too difficult for traditional computing methods to effectively support related processing, analysis and computation (Huang et al. 2018).

Data-driven research within construction safety management include applications of a) *natural language processing and text analytics* (Tixier 2016b; Zhang et al. 2019; Barker et al. 2020); b) *data mining* (Rivas et al. 2011; Cheng et al. 2012; Hsueh et al. 2013; Tixier et al. 2017; Huang et al. 2018); and c) *predictive algorithms* which employs statistical learning and machine learning (**ML**) techniques to produce validated safety/injury risk leading indicators or outcomes upon which informed safety decision-making can be based such as how organizations effectively prioritize their safety management resources. ML applications include, for example, decision tree and random forests algorithms which were employed to predict construction injury severity levels and flag high-risk sites (Gondia et al. 2021). Random forest algorithms were also used to predict energy types produced from construction incidents (Tixier 2016a) and accident types to occur in Korean construction projects (Kang and Ryu 2019). Support vector machine algorithms were used to: i) classify construction accidents in Singapore based on the number of man-days lost (Poh et al. 2018); ii) determine safety risks associated with the construction of deep pit foundations in subway infrastructure projects in China (Zhou et al. 2017); and iii) forecast failure modes of large and complex scaffolds on construction sites (Sakhakarmi et al. 2019). Artificial neural network algorithms were utilized to predict construction project safety climate scores in India (Patel et al. 2015) and to classify accidents in the Turkish construction sector (Ayhan and Tokdemir 2020). Bayesian algorithms were used to determine accident types in mining and embankment construction projects in Spain (Gerassis et al. 2017). Studies in the area of ML-based predictive algorithms, such as those cited above, have made great strides by opening the gate for the first time to leveraging big and objective safety-related data in order to develop algorithms that proactively forecast injury outcomes and safety risks. These studies are also not limited by statistical assumptions and have demonstrated validation of their resulting leading indicators.

4.1.3 Ensemble Approaches

Before selecting the predictive algorithm to predict injury outcomes as safety risk leading indicators, it is important to note how construction safety environments are increasingly being reported as inherently complex systems in nature, and how injuries are the emergent outcomes of interdependent interactions within such systems (Tixier et al. 2016a; Alkaissy et al. 2020). Such complexity can be attributed to the unique characteristics of construction projects including the various involved trades, transient workforce, dynamic work environments, and often unstraightforward required construction techniques (Guo et al. 2015). Such characteristics thus render construction injuries as the resulting outcome of the joint presence of a worker and the interplay among several injury precursors within that worker’s surrounding complex environment. These precursors are related to numerous attributes including those related to construction means and methods, environmental conditions and human behavior, and can be either observed before an injury occurs or collated after the fact. These unique characteristics set construction apart from many other industries and pose significant challenges to its safety management.

Integrated ML approaches as multiple algorithm (ensemble) approaches are useful ways to approach such complex systems and predict their phenomena since they have been reported to be considerably successful in complex domains compared to individual algorithm methodologies (Chou and Lin 2013; Son et al. 2014). Specifically, a ML ensemble algorithm is comprised of a set of underlying ML base algorithms whose individual learning strengths are integrated to better adapt to complex systems by modelling their multiple complex facets including the relationships between input (e.g., injury precursors) and response (e.g., injury outcomes) attributes (Zhou et al. 2019). When utilizing such an ensemble algorithm for future predictions, the base algorithm prediction results are combined such that better predictive performances and improved generalization capabilities can be achieved by reducing the variance in the individual algorithm predictions (Gholizadeh et al. 2018).

In this respect, the goal of the current study is to develop construction site risk models that generate predictions of injury outcomes and safety risk leading indicators across different site zones and over project lifecycles, thus empowering a proactive and real-time approach to construction safety management. In this respect, ensemble algorithms, which drive the generation of predictions within these models, are trained and validated to learn from previous injury precursors and outcomes. The study employs naïve Bayes, decision tree, random forest, support vector machine and artificial neural network as the base algorithms of the ensemble algorithm. A demonstration application is subsequently considered, where the ensemble algorithms are employed to develop a site risk model that supports proactive and real-time decision making by generating leading indicator forecasts of site-specific risk levels, injuries' financial implications and body parts most likely affected.

4.2 ML Algorithm Approach

Supervised ML algorithms have gained significant recognition for their regression (i.e., when the response attribute is quantitative) and classification (i.e., when the response attribute is qualitative) capabilities. When used for classification, ML algorithms/classifiers learn from historical/training data to map a set of input attribute categories (e.g., injury precursors) into one of several response attribute classes (e.g., injury outcomes). In the current study, five different ML algorithms were considered: naïve Bayes, decision tree, random forest, support vector machines and artificial neural networks—subsequently serving as the base algorithms for the final ensemble algorithms. The training, hyperparameter tuning and evaluation procedure followed for each developed algorithm are illustrated through the flowchart shown in Figure 4.3.

4.2.1 Training and Evaluation

The dataset is first split into an 80% training set and a 20% testing set, where the dataset featured in Figure 4.3 pertains to the one adopted in the demonstration application discussed later. The splitting is carried out in a stratified manner, where both training and testing sets have similar distributions of the response attribute (Kohavi 1995; Fiore et al. 2016). The training set is employed to train the base ML algorithms and to tune/optimize each algorithm's hyperparameters, whereas the testing set was later introduced to present the algorithms with new data to validate and evaluate their performances. To ensure consistency in the evaluation, the same training set and testing set are used for the different ML algorithms. The results of the performance evaluation play a key role in the ensemble algorithms development as discussed later. The measures used to evaluate the ML algorithms herein include: **accuracy**, **precision**, **sensitivity** and **specificity** (Chou and Lin 2013; Son et al. 2014; Seong et al. 2018). In applications with class imbalance (as will be discussed later), it is also endorsed to assess the **F-measure** (McDonald et al. 2012; Kang and Ryu 2019) which aggregates the precision and sensitivity measures as:

$$(2 \times precision \times sensitivity) \,/\, (precision + sensitivity) \tag{4.1}$$

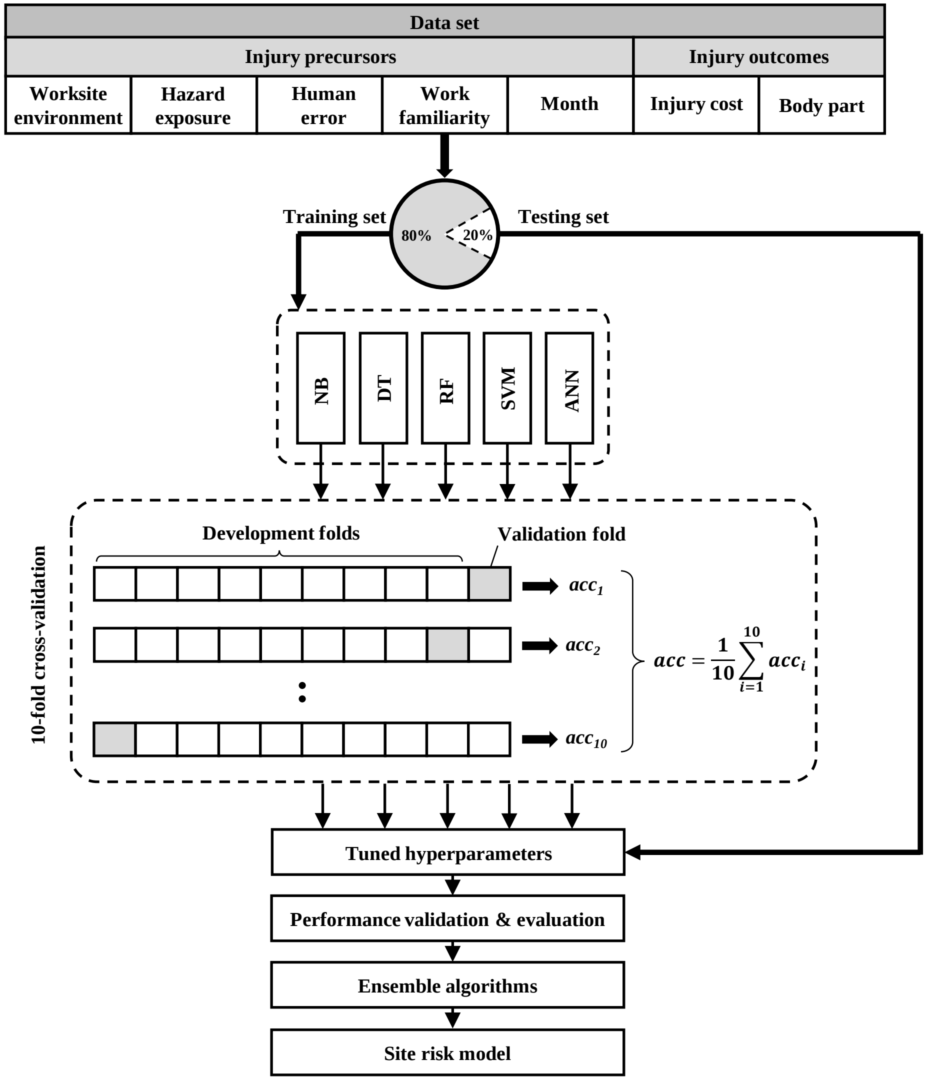


Figure 4.3: Algorithm training, hyperparameter tuning and validation approach

To compound the effect of the five described measures, an overall average performance score ($\boldsymbol{S}$) for each algorithm was also calculated as:

$$S = 1/m \sum_{i=1}^{m} d_i \tag{4.2}$$

where m is the number of distinct evaluation measures (e.g., 5 herein) and d_i is the i^{th} measure.

4.2.2 Hyperparameter Tuning

Each ML algorithm comprises one or more hyperparameters that make up that algorithm's architecture. These hyperparameter values need to be tuned to yield an algorithm that neither overfits (i.e., learns the unique pattern of the data) nor underfits (i.e., unable to capture the relationships between inputs and outputs well enough) with respect to the training data, but rather strikes the right balance such that the algorithm preserves good generalization capabilities to new data. To carry out algorithm hyperparameter tuning, grid search techniques are used in conjunction with a 10-fold cross-validation approach, as shown in Figure 4.3. For each combination of hyperparameter values, a 10-fold cross-validation procedure is carried out, where the training set is divided into ten separate and almost equally-sized folds; nine for algorithm development and one for its validation. This process is then repeated ten times where each fold is used once for validation. Ten algorithms are thus developed using the same hyperparameter combination resulting in ten accuracy values on the validation folds which are averaged to yield one accuracy value for each hyperparameter combination. Grid search techniques are adopted to search through the hyperparameter space and record average accuracies against hyperparameter combinations, and the combination with the highest average accuracy provides the optimal algorithm.

4.2.3 Base Algorithms Overview

4.2.3.1 Naïve Bayes

Naïve Bayes (**NB**) is a probabilistic classifier based on the Bayes theorem of conditional probability. NB simplifies the calculation of probabilities by assuming conditional independence in the input attributes belonging to a given response attribute's class (John and Langley 2013; Goh and Ubeynarayana 2017). During training, class conditional probabilities are calculated for each input attribute category (Gondia et al. 2019; Seong et al. 2018). For a new case, NB calculates the maximum posterior probability and then assigns the most likely class based on the case's input attributes (Rish 2001; Gerassis et al. 2017). NB is widely used as a classification technique because, despite its simplicity, it has shown good predictive performances in various classification tasks (Liu et al. 2013; Bhowmik 2015; Moreira et al. 2016; Kakhki et al. 2019). Specifically, NB algorithms usually perform well when applied to high-dimensional datasets or in binary classification problems (Marucci-Wellman et al., 2017; Seong et al. 2018), and were thus adopted as part of the current ensemble algorithm.

4.2.3.2 Decision Tree

Decision trees (**DT**) (Breiman et al. 1984) are based on the recursive binary splitting of the data cases across multiple nodes through a divide-and-conquer approach. Starting from the root node which consists of all the data cases, the most influential input attribute is determined via calculations of entropy and its resultant information gain (Poh et al. 2018; Gondia et al. 2019). All attribute categories are considered and those providing the best homogeneity (i.e., the cleanest split in terms of the response attribute classes) are selected to split the data into two subset nodes. The process is then repeated, where the cases within each node become more homogeneous than those in the previous nodes, and stopped once full homogeneity is reached or some other stopping criterion is satisfied (Chou and Lin 2013). The resulting tree can be used to generate a set of classification rules from the data which sends a new case to the lowest node in the tree based on its input attributes (Son et al. 2014). Usually, full homogeneity is not reached since this can cause overfitting, but instead, tree pruning is employed using cross-validation (Zhou et al. 2019). Tree pruning is controlled by the complexity parameter (***cp***) which describes the minimum improvement in the algorithm accuracy needed to be achieved to allow a node to split. DT were adopted herein since they have shown good classification performances, when pruned, across various applications (Navada et al. 2011; Somvanshi et al. 2016).

4.2.3.3 Random Forest

Random forests (**RF**) (Breiman 2001) are ensemble algorithms of DTs which combine two powerful techniques: bootstrap aggregating (bagging) (Breiman 1996) and random attribute selection (Ho 1998). In bagging, $\boldsymbol{n_{tree}}$ bootstrap data samples are drawn from the training set by sampling with replacement, where each sample is used to independently build a DT (Zhang et al. 2020). Trees in RF are constrained to be simple, where each tree is grown to a maximal depth (set as 5 in the current application) and no tree pruning is applied (Nitsche et al. 2014). During tree construction, instead of considering all input attributes as in DT, RF selects only a subset of $\boldsymbol{k_{try}}$ attributes (less than the total number of attributes) which are randomized at each node split (Zhou et al. 2019). For a new case, each tree in the forest classifies its final response attribute class, and the majority class among the collection of trees is the RF final classification (Goh and Ubeynarayana 2017). The RF tuning hyperparameters are $\boldsymbol{n_{tree}}$ and $\boldsymbol{k_{try}}$. Compared to DT, RF is known to be more robust in terms of generalizability to new data and better in cases with input attributes with varying predictive powers (Nitsche et al. 2014; Poh et al. 2018). This is because through randomizing the procedures of data sampling and attribute selection at each node split, RF can reduce bias to training data and to attributes with strong predictive powers/influence on the response attribute, and can thus form a valuable part of the ensemble algorithm.

4.2.3.4 Support Vector Machine

Support vector machines (**SVM**) (Vapnik 1999) are known as strong classifiers due to their ability to map training data to a higher dimensional feature space using nonlinear kernel functions when such data cannot be separated linearly (Meyer 2001; Olson et al. 2012). In the transformed feature space, nonlinear class boundaries are more easily separable through linear hyperplanes which appear nonlinear in the original feature space (Son et al. 2014). To achieve good generalizability, SVM detects the separating hyperplanes that maximize the margins between the underlying classes (Seong et al. 2018). Popular kernel functions include the polynomial, sigmoid and radial basis functions (RBF), where the latter is used in the current application due to its renowned high performance (Zhou et al. 2017; Kakhki et al. 2019). The RBF kernel, which is computed for pairs of training vectors (e.g., x_i and x_j), takes the form (James et al. 2013; Nitsche et al. 2014; Zhang et al. 2020):

$$K(x_i, x_j) = \exp\left(-\gamma\left\|x_i - x_j\right\|^2\right) \tag{4.3}$$

The SVM tuning hyperparameters are the cost/penalty of training cases that violate the separating plane ($\boldsymbol{C}$), and the RBF kernel parameter ($\boldsymbol{\gamma}$). Low values of $\boldsymbol{C}$ indicate more tolerance of violations (acceptable errors) and thus a wider margin which yields a classifier with potentially higher training error but better generalizability. On the other hand, high values of $\boldsymbol{C}$ indicate narrow margins that are rarely violated which results in a classifier with improved fitting performance but deteriorated generalization (i.e., overfitting) (James et al. 2013). The $\boldsymbol{\gamma}$ parameter plays an important role in optimizing the separating plane as it denotes how far the influence of a single training case reaches in terms of defining the plane's boundary. Low γ values denote far reaches and high values denote close reaches, and so $\boldsymbol{\gamma}$ can be seen as the inverse of the radius of influence of the selected support vectors on the boundary. When $\boldsymbol{\gamma}$ is too small, the region of influence can include the whole training set, thus preventing the classifier from capturing the complexity or shape of the data. If $\boldsymbol{\gamma}$ is too large, the region of influence can only include the support vectors, thus yielding a classifier prone to overfitting (Zhou et al. 2017). Combinations of $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ provide a wide range of feature space transformations and separating hyperplanes, which enables RBF SVMs to accommodate for complex data mappings. As such, it is known to perform well in complex classification problems, which justified its consideration herein.

4.2.3.5 Artificial Neural Networks

Artificial neural networks (**ANN**) are inspired by the organization and functioning of biological neural systems, as that of human brains, and are effective in simulating relationships between input and response attributes that are part of complex nonlinear systems (Mangalathu and Jeon 2018; Kulkarni et al. 2017). According to most researchers (Arditi et al. 1998; Arditi and Tokdemir

1999; Kulkarni et al. 2017; Waziri et al. 2017), a feed-forward back-propagation neural network was used in the current study, as shown in Figure 4.4. Typically, an ANN consists of an input layer, an output layer and one or more hidden layers. Increasing the number of hidden layers can enhance the modelling capabilities in highly complicated systems but also can increase the risk of overfitting. In this study, one hidden layer was used which was i) proven sufficient to approximate any continuous nonlinear function (Cybenko 1989; Hossein et al. 2010; Gandomi and Roke 2015); and ii) capable of solving complex classification problems (Lippmann 1987; Ripley 2002; Son et al. 2014). As can be seen in Figure 4.4, the input layer contains a set of *m* neurons representing the input attributes, the output neurons serve as discriminators between the response attribute's *n* classes, and the hidden layer contains several $\boldsymbol{h}$ computation neurons connected to the input and output neurons by numeric weights in order to pass information between the input and output layers. Every neuron in the hidden layer undergoes two computations, where the first determines its net input through a biased weighted summation of all its connected input neurons as:

$$A_j = \sum_{i=1}^{m} w_{ij} x_i + b_j \qquad , \quad j = 1, .., h \tag{4.4}$$

where A_j is the net input of hidden neuron *j*, w_{ij} is the weight connecting input neuron *i* to hidden neuron *j*, and b_j is the bias or threshold term for hidden neuron *j*. For the second computation, the net input result is passed into an

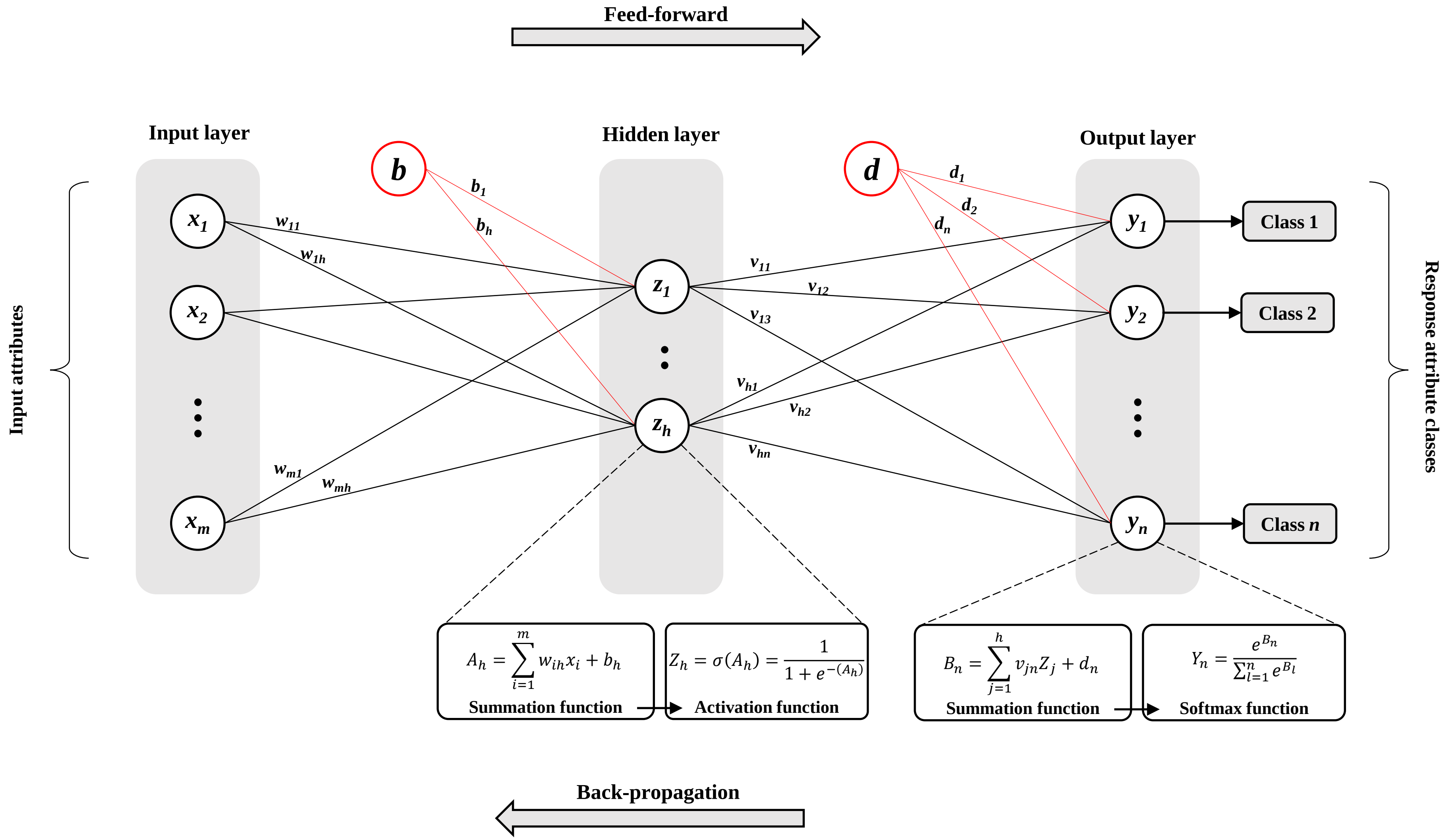


Figure 4.4: Architecture diagram for feed-forward back-propagation neural network with one hidden layer

activation or transfer function that controls the contribution of a hidden neuron's output. Popular activation functions include the unit function, rectified linear function, hyperbolic tangent function or sigmoid function, where the latter is typically preferred and was used in this study in order to introduce nonlinearity into the algorithm and due to its ability to capture the complexity in systems (Moayed and Shell 2011; Ayhan and Tokdemir 2020). As such, the output (Z_j) of the j^{th} neuron in the hidden layer can be described as:

$$Z_j = \sigma(A_j) = \frac{1}{1 + e^{-(A_j)}} \quad , \quad j = 1, .., h \tag{4.5}$$

The results of the hidden layer are ultimately mapped into the output layer through another two computations, where the input into the k^{th} output neuron is:

$$B_k = \sum_{j=1}^{h} v_{jk} Z_j + d_k \quad , \quad k = 1, .., n \tag{4.6}$$

where v_{jk} is the weight connecting hidden neuron j to output neuron k, and d_k is the bias in output neuron k. In classification problems, the softmax function in Equation (4.7) is commonly used and allows for a final transformation of the inputs (e.g., B_k) into positive probabilities that sum to one (Hastie et al. 2009).

$$Y_k = \frac{e^{B_k}}{\sum_{l=1}^{n} e^{B_l}} \quad , \quad k = 1, .., n \tag{4.7}$$

where Y_k is the final probability that an instance becomes classified as class *k*. The network architecture is completed by determining the hidden layer size (i.e., number of hidden neurons ***h***) which is key since too small or large ***h*** values can lead to underfitting or overfitting, respectively. The current study determined an upper bound for ***h*** using Kolmogorov's theorem (Hecht-Nielsen 1987; Gandomi and Roke 2015) as: $h \leq 2m + 1$, where *m* is the number of inputs as stated earlier, and the exact number of ***h*** was subsequently tuned using cross-validation, as will be described later. Training the ANN is a two-stage process: i) a forward pass of training cases through the network to produce predicted classes which are compared with the actual classes for computing error rates; and ii) a backward pass to continuously adjust the weights and biases to reduce errors and iteratively capture the input-output relationships enshrined within the data. For a new case, the input attributes flow through the hidden layer of neurons to the output layer and the class *k* with the highest probability Y_k is the predicted class.

4.2.4 Ensemble Algorithm

For each of the described base algorithms, the prediction output is initially in the form of a probability distribution over the outcome classes, rather than a definitively selected class. For example, as presented in Table 4.1, if the prediction probabilities of classes A, B and C are 55%, 32% and 13%, respectively, then class A is the prediction. The ensemble methodology subsequently aggregates the prediction probabilities of the base algorithms in order to determine the final prediction—thus producing an ensemble algorithm that performs at least as well as the best-performing base algorithm and oftentimes better (Hou and Ramani 2007; Zhou et al. 2011).

Table 4.1: Demonstration example for different ensemble algorithm techniques

Confidence score	Algorithm	Class A	Class B	Class C
80.87%	Base algorithm 1	55.10% *	31.59%	13.31%
37.54%	Base algorithm 2	17.83%	39.55%	42.62% *
22.81%	Base algorithm 3	8.52%	44.70%	46.78% *
—	Ensemble algorithm 1 (majority voting)	—	—	100% *
—	Ensemble algorithm 2 (average voting)	27.15%	38.61% *	34.24%
—	Ensemble algorithm 3 (confidence-weighted average voting)	37.67% *	35.82%	26.51%

Note. * Final prediction per algorithm.

Several ensemble techniques are available including majority voting, average voting and confidence-weighted average voting (see Table 4.1). In majority voting, each base algorithm casts a vote (i.e., a predicted class) and the maximum vote is the final prediction. For example, if two out of three base algorithms predict class C, then the final prediction is class C. In the case of average voting, the base algorithms' prediction probabilities are simply averaged, and the final prediction is the class with the highest average. In the current study, confidence-weighted average voting is adopted, where the prediction probabilities are weighted according to a confidence score for each base algorithm. The confidence score used herein for each base algorithm is its overall average performance score (S) which was introduced earlier. More specifically, the ensemble algorithm's prediction probability for a certain class *k* is calculated as:

$$P_k = \frac{\sum_{i=1}^{n} S_i p_{ik}}{\sum_{i=1}^{n} S_i} \qquad , \quad \forall k \qquad (4.8)$$

where n is the number of base algorithms and p_{ik} is the prediction probability of the i^{th} base algorithm for that class *k*. The final prediction is the class with the highest prediction probability as per the confidence-weighted average calculations (Hastie et al. 2009; Chou and Lin 2013; Son et al. 2014).

4.3 DEMONSTRATION APPLICATION

4.3.1 Data Description

The data adopted in the current study to demonstrate an ensemble algorithm approach for developing a site risk model were obtained from the Occupational Safety and Health Administration (OSHA) occupational injury and illness cases database (OSHA 2019). At the time of access, 1,981 cases corresponded to construction workplace injuries which were used herein. The dataset consists of seven attributes; **five injury precursors** (input attributes) and **two injury outcomes** (response attributes). The five precursors include: 1) *worksite environment*—describing the unsafe nature of the working conditions surrounding- or in close proximity to the worker at the time of the injury; 2) *hazard exposure*—describing the hazards associated with the specific work method performed by the worker to complete a task at the time of the injury; 3) *human error*—describing worker negligence of controllable circumstances within their responsibilities; 4) *work familiarity*—describing whether the task assigned was considered regular work performed by the worker; and 5) *month*—describing weather conditions at the time of injury, since the dataset did not include detailed injury location information to match to the date for deriving weather data. The precursor categories and corresponding injury counts within the adopted dataset are presented in Table 4.2.

Table 4.2: Injury precursor categories and injury counts

Precursor	Category	Injury count	Total count
Worksite environment	1. Work-surface/facility-layout condition	456	1,981
	2. Materials handling equip./method	428	
	3. Pinch point action	338	
	4. Overhead moving/falling object action	247	
	5. Catch point/puncture action	153	
	6. Shear point action	80	
	7. Flying object action	77	
	8. Weather, earthquake, etc.	55	
	9. Gas/vapor/mist/fume/smoke/dust	41	
	10. Overpressure/underpressure	34	
	11. Illumination	24	
	12. Temperature +/- tolerance level	18	
	13. Chemical action/reaction exposure	16	
	14. Sound level	14	
Hazard exposure	1. Caught in or between	720	1,981
	2. Struck-by	576	
	3. Fall (from elevation)	463	
	4. Shock	63	
	5. Struck against	60	
	6. Inhalation	46	
	7. Fall (same level)	34	
	8. Card-vascular/resp. failure	9	
	9. Rubbed/abraded	5	
	10. Absorption	5	
Human error	1. Misjudgment, hazardous situation	857	1,981
	2. Safety devices removed/inoperable	190	
	3. Position inappropriate for task	168	
	4. Mater-handling procedure inappropriate	130	
	5. Insufficient/lack of engineering controls	106	
	6. Insufficient/lack of written work practice program	102	
	7. Equipment inappropriate for operation	95	

	8. Insufficient/lack of protective work clothing/equipment	86	
	9. Lockout/tagout procedure malfunction	85	
	10. Malfunction in securing/warning operations	78	
	11. Perception malfunction task-environment	23	
	12. Defective Equipment in use	19	
	13. Distracting actions by others	17	
	14. Insufficient/lack of housekeeping program	14	
	15. Insufficient/lack of respiratory protection	11	
Work familiarity	1. Regularly assigned	1,293	1,981
	2. Not regularly assigned	688	
Month	January	172	1,981
	February	201	
	March	207	
	April	146	
	May	132	
	June	144	
	July	159	
	August	144	
	September	146	
	October	145	
	November	211	
	December	174	

The first injury outcome is the *injury cost* that describes the level of financial implications incurred by the employers due to the injury. As shown in Table 4.3, the dataset originally featured the *injury nature* outcome which describes the principal physical characteristics of the injury. The *injury nature* categorization scheme is consistent with that of the Bureau of Labor Statistics' standardized Occupational Injury and Illness Classification Manual (OIICM) (Bureau of Labor Statistics 2012). Subsequently, the *injury nature* categories were assessed based on the worker compensation lost work-time claims data and the injuries in the dataset were classified accordingly into high- and low-cost injuries, as shown in Table 4.3 (Rosecrance et al. 2011; Davis and Stern 2012; Liao et al. 2015; CPWR 2018; Gholizadeh et al. 2018). The injury counts pertaining to the resulting *injury cost* outcome are summarized in Table 4.4. The second injury outcome is the *body part* which describes the part of the body directly affected by the injury. As shown in Table 4.5, in the adopted dataset, the *body part* outcome was originally comprised of 27 subcategories which were also recorded consistent with the OIICM. These subcategories were subsequently classified into five OIICM categories. The classification carried out on the two injury outcomes within the current study was performed in order to simplify the predictive modelling process and to enhance the practical use of the algorithm outcomes.

As shown in Table 4.4, the *injury cost* outcome is classified into two classes including high-cost injury (46.3%) and low-cost injury (53.7%). As shown in Table 4.5, the *body part* outcome is classified into five classes including head & neck (27.6%), trunk (22.7%), upper extremities (29.7%), lower extremities (10.3%), and multiple areas (9.7%). As discussed, the prediction of these outcome classes is called a classification problem, where the *injury cost* problem is a binary classification problem and the *body part* problem is a multi-class classification problem (i.e., having more than two classes).

Table 4.3: Injury nature outcome categories and injury counts

OIICM category	**OIICM subcategory***	**Subcategory injury count**	**Injury cost**	**Category injury count**
2.1.2.1.1 Traumatic injuries to bones, nerves, spinal cord	Fracture/broken bones	402	High cost	402
2.1.2.1.2 Traumatic injuries to muscles, tendons, ligaments, joints, etc.	Strain/sprain	481	Low cost	509
	Dislocation	28	Low cost	
2.1.2.1.3 Open wounds	Amputation/crushing	356	High cost	570
	Laceration	192	Low cost	
	Puncture	22	Low cost	
2.1.2.1.4 Surface wounds and bruises	Bruising/contusion	147	Low cost	147
2.1.2.1.5 Burns and corrosions	Fire burn	21	Low cost	31
	Chemical burn	10	Low cost	
2.1.2.1.6 Intracranial injuries	Head trauma	159	Low cost	159
2.1.2.1.9 Other traumatic injuries and disorders	Asphyxiation/drowning	96	High cost	163
	Electrocution	64	High cost	
	Poisoning	3	Low cost	
Total	–	**1,981**	–	**1,981**

Note. * The attribute in the original data set.

Table 4.4: Injury cost outcome categories and injury counts

Category	**Injury count**
High-cost injury	918
Low-cost injury	1063
Total	**1,981**

Table 4.5: Body part outcome categories and injury counts

OIICM category	OIICM subcategory*	Subcategory injury count	Category injury count
2.2.2.1 Head & 2.2.2.2 Neck	Head	452	546
	Neck	67	
	Face	25	
	Eye(s)	2	
2.2.2.2.3 Trunk	Rib(s)	131	450
	Internal chest organ(s)	131	
	Lung(s)	49	
	Abdomen	41	
	Back	36	
	Heart	33	
	Hip	23	
	Liver	4	
	Kidney(s)	2	
2.2.2.4 Upper extremities	Finger(s)	367	588
	Hand(s)	100	
	Shoulder(s)	47	
	Arm(s)	43	
	Wrist(s)	14	
	Forearm(s)	13	
	Elbow(s)	4	
2.2.2.5 Lower extremities	Leg(s)	99	204
	Feet	57	
	Lower leg(s)	30	
	Thigh(s)	10	
	Knee(s)	8	
2.2.2.8 Multiple areas	Whole body	180	193
	Multiple body parts	13	
Total	–	**1,981**	**1,981**

Note. * The attribute in the original data set.

4.3.2 Class Imbalance

The *injury cost* classes are relatively balanced/equally distributed (see Table 4.4), while the *body part* outcome carries some class imbalance (see Table 4.5). Specifically, on the latter, the underrepresented classes include lower extremities and multiple areas. Class imbalance presents challenges to ML algorithms during training, where typically the final ML algorithms would perform well for the majority classes but neglect the minority classes (Tixier 2016a; Kang and Ryu 2019). In the current study, accurately predicting the rare classes was equally important as predicting the common ones. To address the class imbalance issue, the oversampling with replacement method was used for the *body part* problem instead of the stratified splitting method (described earlier) which was used for the balanced *injury cost* problem. In the oversampling with replacement method, the training set is created through a random sample containing more cases from the minority classes than what would have been normally obtained by the stratified splitting method (Sun et al. 2007; Poh et al. 2018). To achieve this, each *body part* class was assigned a probability that is inversely proportional to its count, for example, head & neck (588/546), trunk (588/450), upper extremities (588/588), lower extremities (588/204), and multiple body parts (588/193). The training set was put together by random drawing with replacement from the entire dataset based on these probabilities until the classes were equally represented. Rebalancing the class distributions allowed the underrepresented classes to become more important to the ML algorithms during training in order to produce final algorithms that perform well across all classes.

4.3.3 Algorithm Training

Two ensemble algorithms were developed; one to predict each injury outcome class from the same set of injury precursor categories. Central to the ensemble approach, the five base ML classification algorithms were trained independently (for each of the two problems) and then used collectively to select the best/final outcome class. During training, each algorithm's hyperparameters were tuned using a cross-validated grid search as described earlier, and the tuned hyperparameters were the ones used to ultimately train the algorithm. The tuned sets of hyperparameters and the corresponding optimal accuracies for each algorithm across both problems are reported in Table 4.6.

Table 4.6: Algorithms tuned hyperparameters

Algorithm	**Hyperparameters**	**Injury cost problem**		**Body part problem**	
		Tuned value	**Optimal accuracy***	**Tuned value**	**Optimal accuracy***
DT	*cp* Tree size	0.0082 10	78.42%	0.0061 9	72.38%
RF	n_{tree} k_{try}	80 3	79.36%	110 4	76.77%
SVM	Kernel function *C* γ	RBF 8 0.005	79.77%	RBF 16 0.005	75.21%
ANN	No. hidden layers Activation function *h* Number of epochs Weight decay	1 sigmoid 36 500 0.3	80.07%	1 sigmoid 22 500 0.3	77.08%

Note. * Optimal accuracies are results of 10-fold cross-validation averages.

Regarding DT, the results of the ***cp*** tuning showed that the optimal ***cp*** for the *injury cost* problem was 0.008 resulting in a tree size of 10, and for the *body part* problem was 0.006 resulting in a tree size of 9. Concerning RF, the $\boldsymbol{n_{tree}}$ and $\boldsymbol{k_{try}}$ tuning results for the *injury cost* problem are demonstrated in Figure 4.5. As shown in the figure, the range of $\boldsymbol{n_{tree}}$ values was set from 5-500 with a step size of 10, and $\boldsymbol{k_{try}}$ values of 1, 2, 3 and 4 were explored. The optimal combination of hyperparameters for the *injury cost* and *body part* problems was ($\boldsymbol{n_{tree}}$ = 80, $\boldsymbol{k_{try}}$ = 3), and ($\boldsymbol{n_{tree}}$ = 110 and $\boldsymbol{k_{try}}$ = 4), respectively. With regards to SVM, the $\boldsymbol{C}$ and $\boldsymbol{\gamma}$ tuning results for the *injury cost* problem are demonstrated in Figure 4.6. The figure shows a 3D plot and a contour plot of the change in cross-validation accuracy as a function of $\boldsymbol{C}$ and $\boldsymbol{\gamma}$. As the range of $\boldsymbol{C}$ values was significantly wide, its axis was represented as the transformed range of $\log_2 C$. Searching through $\boldsymbol{C}$ values within the range [2^{-10}, 2^{4}] and $\boldsymbol{\gamma}$ values within the range [0.005, 0.05], good SVM classifiers were found for intermediate values of $\boldsymbol{C}$ and $\boldsymbol{\gamma}$, where wider boundary classifiers (i.e., smaller $\boldsymbol{\gamma}$ values) can be made stricter by increasing the importance of misclassifying training cases (i.e., larger $\boldsymbol{C}$ values). The optimal SVM hyperparameter combination was found to be ($\boldsymbol{C}$=8, $\boldsymbol{\gamma}$=0.005) for the *injury cost* problem and ($\boldsymbol{C}$=16, $\boldsymbol{\gamma}$=0.005) for the *body part* problem.

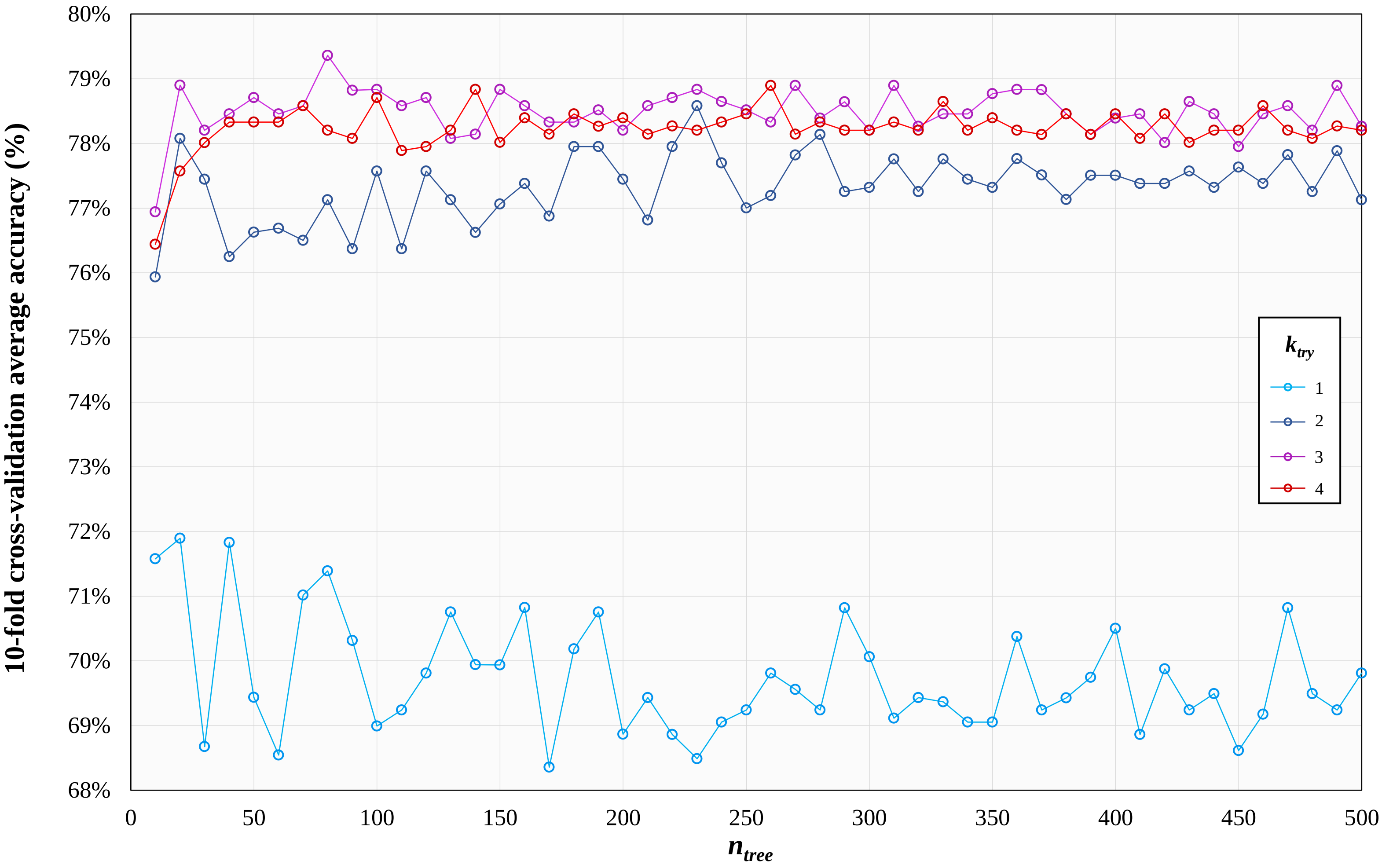
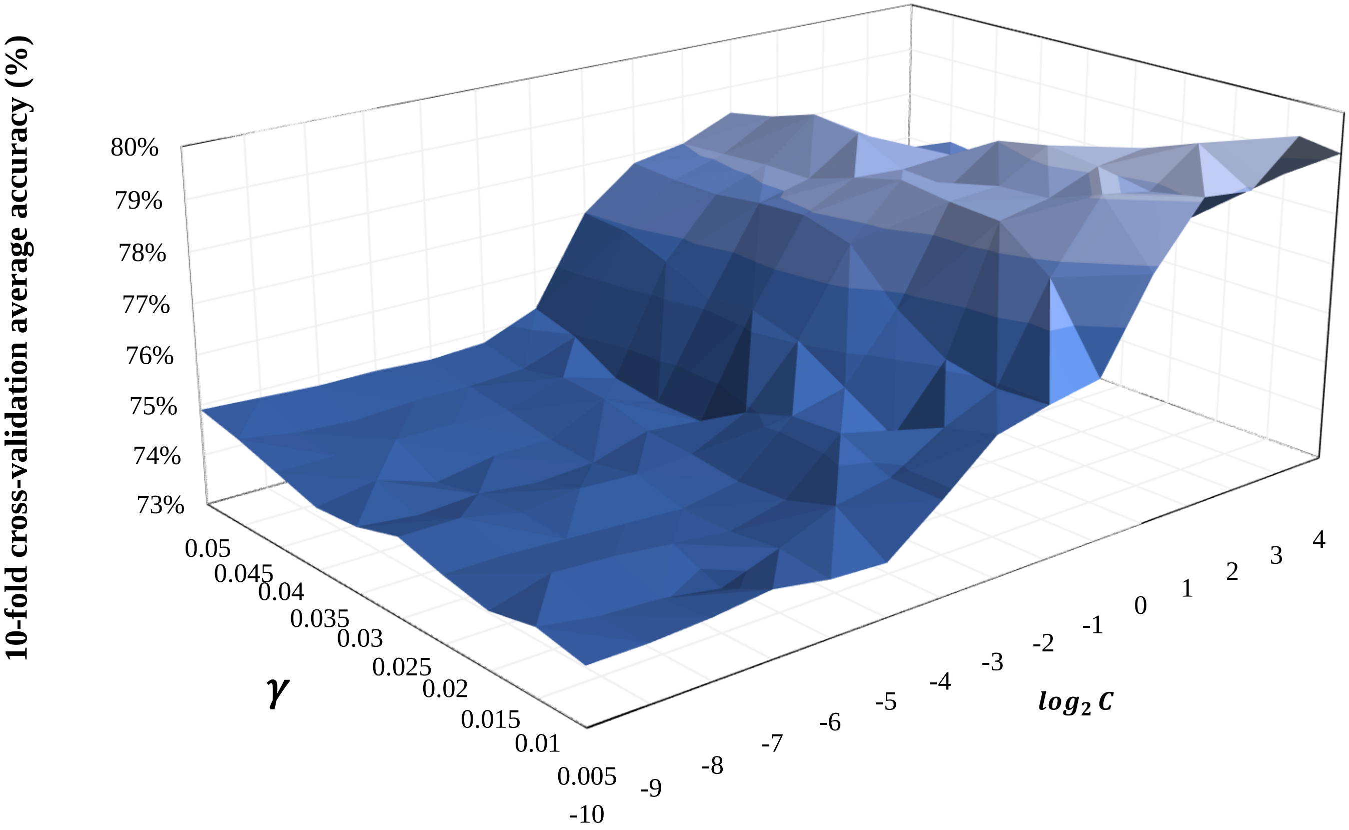
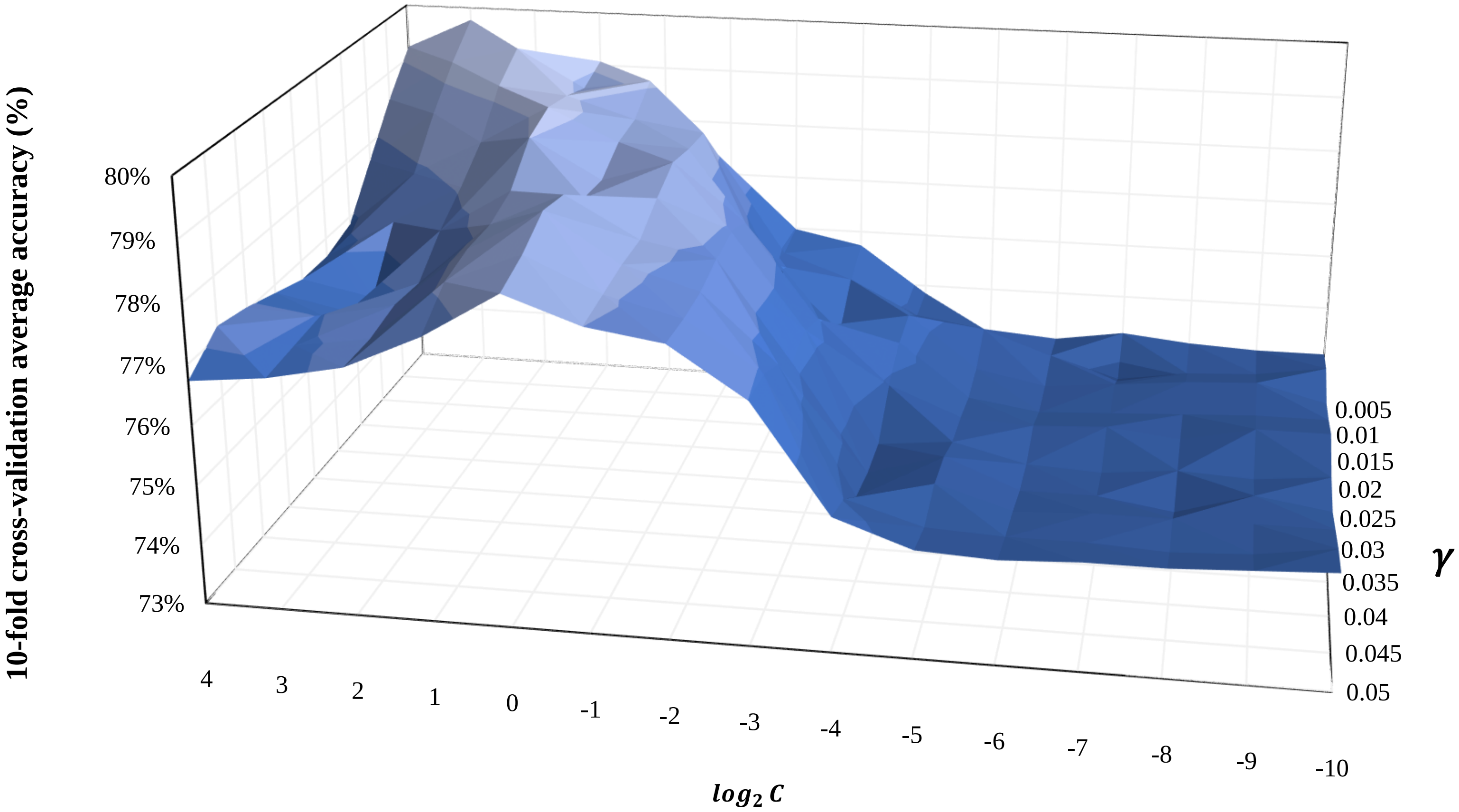


Figure 4.5: RF hyperparameter tuning using cross-validated grid search (injury cost problem): 10-fold cross-validation average accuracy under different values of n_{tree} and k_{try}



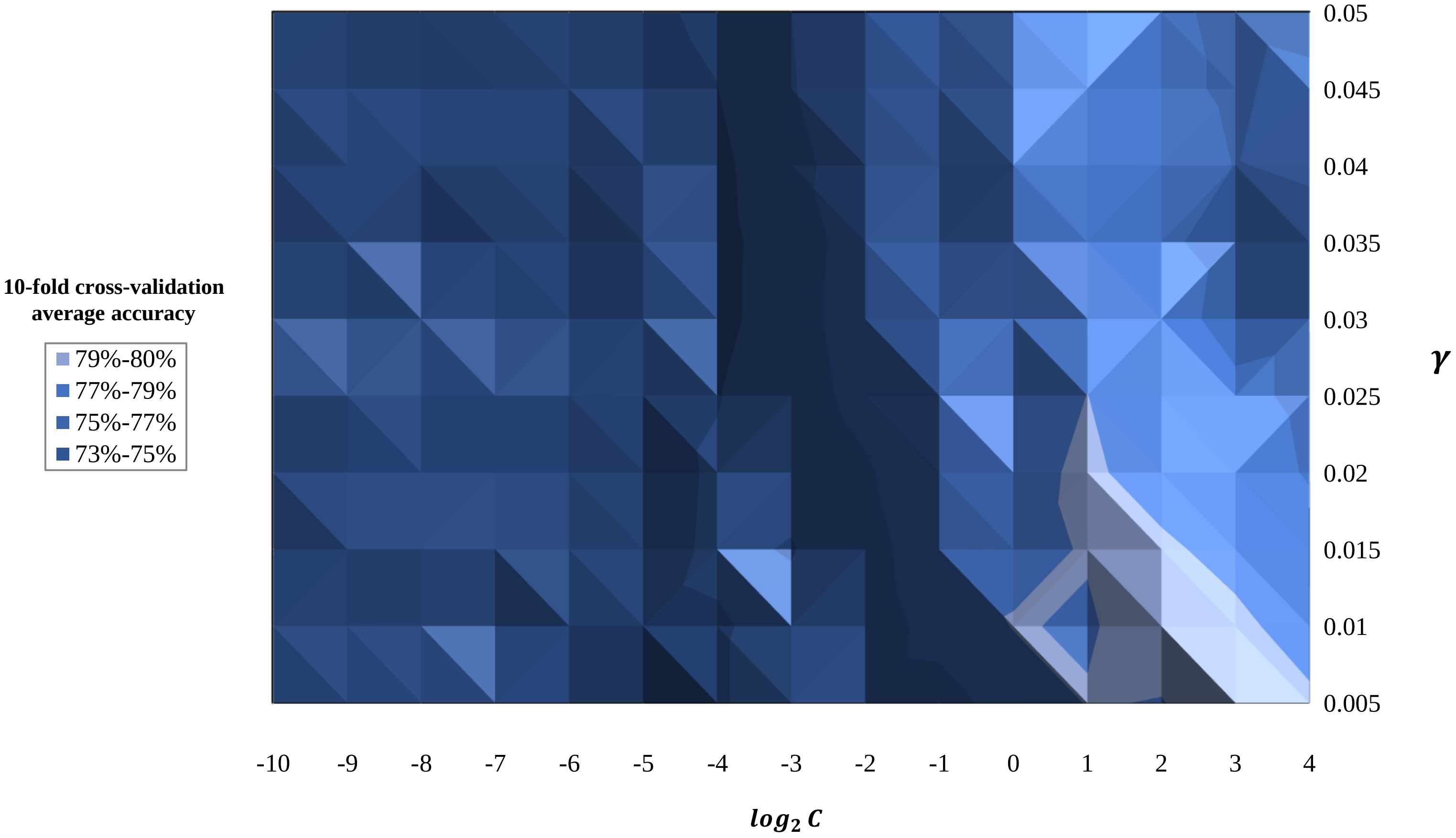
(a)

Figure 4.6: SVM hyperparameter tuning using cross-validated grid search (injury cost problem): 10-fold cross-validation average accuracy under different values of C and γ represented as a) 3D plot angle A; b) 3D plot angle B; and c) contour plot



(b)

Figure 4.6: SVM hyperparameter tuning using cross-validated grid search (injury cost problem): 10-fold cross-validation average accuracy under different values of C and γ represented as a) 3D plot angle A; b) 3D plot angle B; and c) contour plot



(c)

Figure 4.6: SVM hyperparameter tuning using cross-validated grid search (injury cost problem): 10-fold cross-validation average accuracy under different values of *C* and *γ* represented as a) 3D plot angle A; b) 3D plot angle B; and c) contour plot

Throughout the ANN training, the number of epochs or training iterations was set to 500. For ANN, the five categorical injury precursor attributes were encoded to numeric attributes by mapping each to binary vectors based on its underlying categories, thus yielding a total of 53 new input attributes. As such, the maximum ***h*** was determined as 2×53 + 1 = 107 hidden neurons. Figure 4.7 shows the ***h*** tuning results for the *injury cost* problem within the range of 1 and 110 hidden neurons. Too large ***h*** can cause overfitting because too many weights can cause the ANN to match the training data too closely. Indeed, from the figure, the cross-validation average accuracy showed trends of increase with the increase in ***h*** until an optimal value of 36 neurons, beyond which slight overfitting trends started to emerge. The optimal ***h*** for the body part problem was found to be 22. Often ANNs with weight values that are too large can also lead the algorithm to overfitting to the set of training patterns. A weight decay was used (Zur et al. 2009; Nakamura and Hong 2019), where after each epoch/iteration, the weights were multiplied by a factor (i.e., slightly less than one) to prevent the weights from growing too large. By subtracting the decay multiplied by the weight from the original weight, the ANN can be prevented from approaching the global minimum error on the training data. Through manual tuning, the decay values were set to 0.3 for both problems. Finally, based on the ***S*** values discussed next, the ensemble algorithms were developed using confidence-weighted average voting, as discussed earlier.

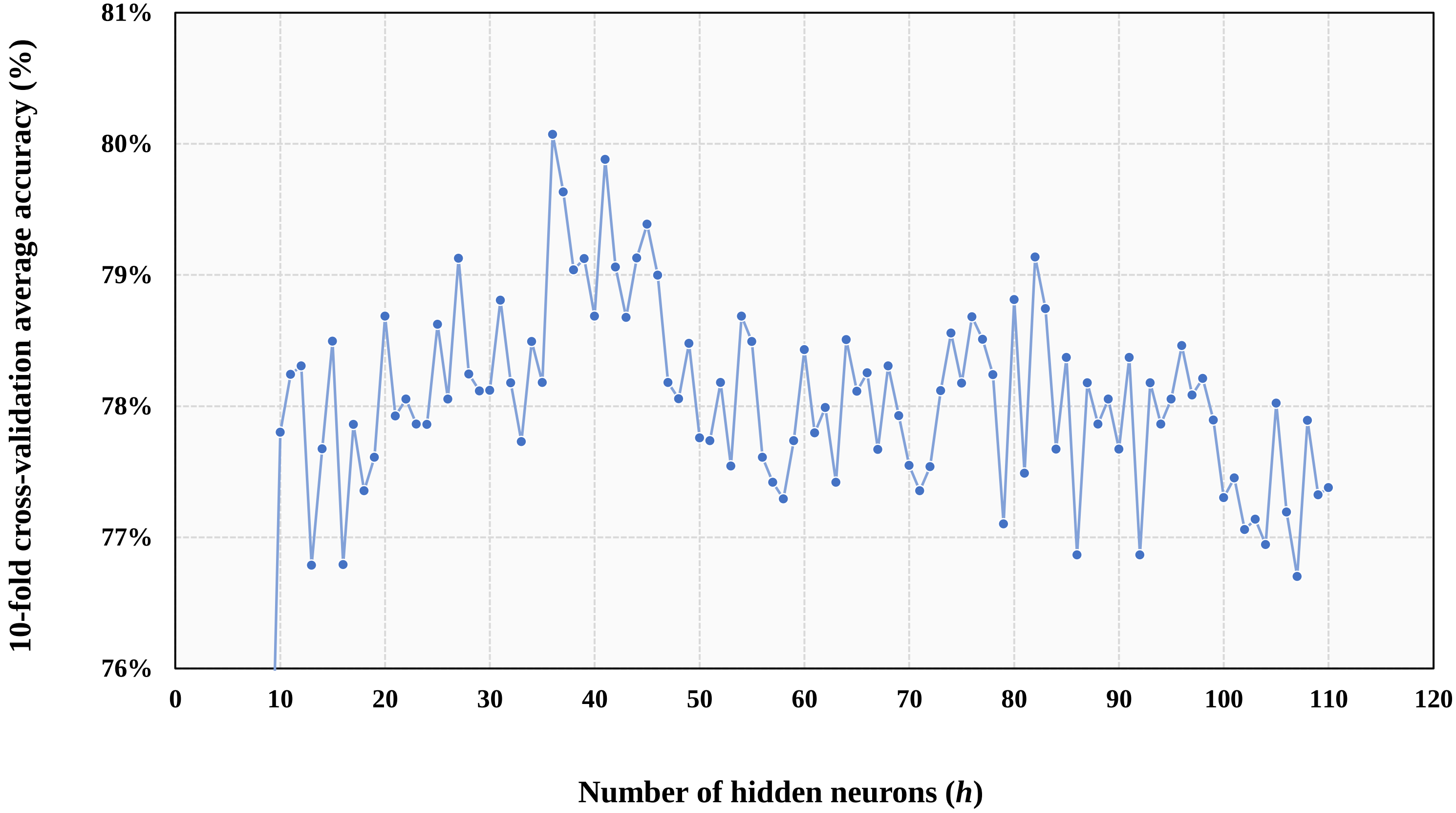


Figure 4.7: ANN hyperparameter tuning (injury cost problem):
10-fold cross-validation average accuracy under different values of *h*

4.3.4 Algorithm Evaluation

Using the tuned hyperparameters, the trained ML algorithms were evaluated based on their ability to replicate the testing set. The evaluation results for the *injury cost* and *body part* problems are reported in Tables 4.7 and 4.8, respectively. In these tables, the algorithms are ranked based on their average performance score (*S*). For the *injury type* problem, the best performing algorithm was ANN (*S*=79.51%) followed by SVM (78.80%), whereas for the *body part* problem, ANN also ranked highest (80.15%) but was followed by RF (79.72%). The ensemble algorithms' aggregation power was also assessed and were found to outperform their base algorithm counterparts for both *injury cost* (81.82%) and *body part* (81.91%) problems. To enable an assessment of the algorithms' generalization capabilities, a 10-fold cross-validation procedure was subsequently applied to the entire dataset, the accuracy results of which are presented in Figure 4.8. Across both problems (injury cost, body part), the ensemble algorithm demonstrated the best generalization performance exhibited through its highest maximum (84.45%, 80.78%), minimum (80.39%, 76.16%), median (82.27%, 78.94%) and average (82.33%, 78.58%) accuracy values, as well as the smallest max-min range (4.06%, 4.62%) over the 10 folds. The latter finding supports that the ensemble algorithms produce less variance and yield similar results if applied repeatedly to distinct datasets. Overall, although the differences in performance between the ensemble algorithm and the base algorithms were not very large, the evaluation results confirm that an ensemble algorithm can improve classification performance relative to that of base algorithms in terms of both accuracy and generalizability.

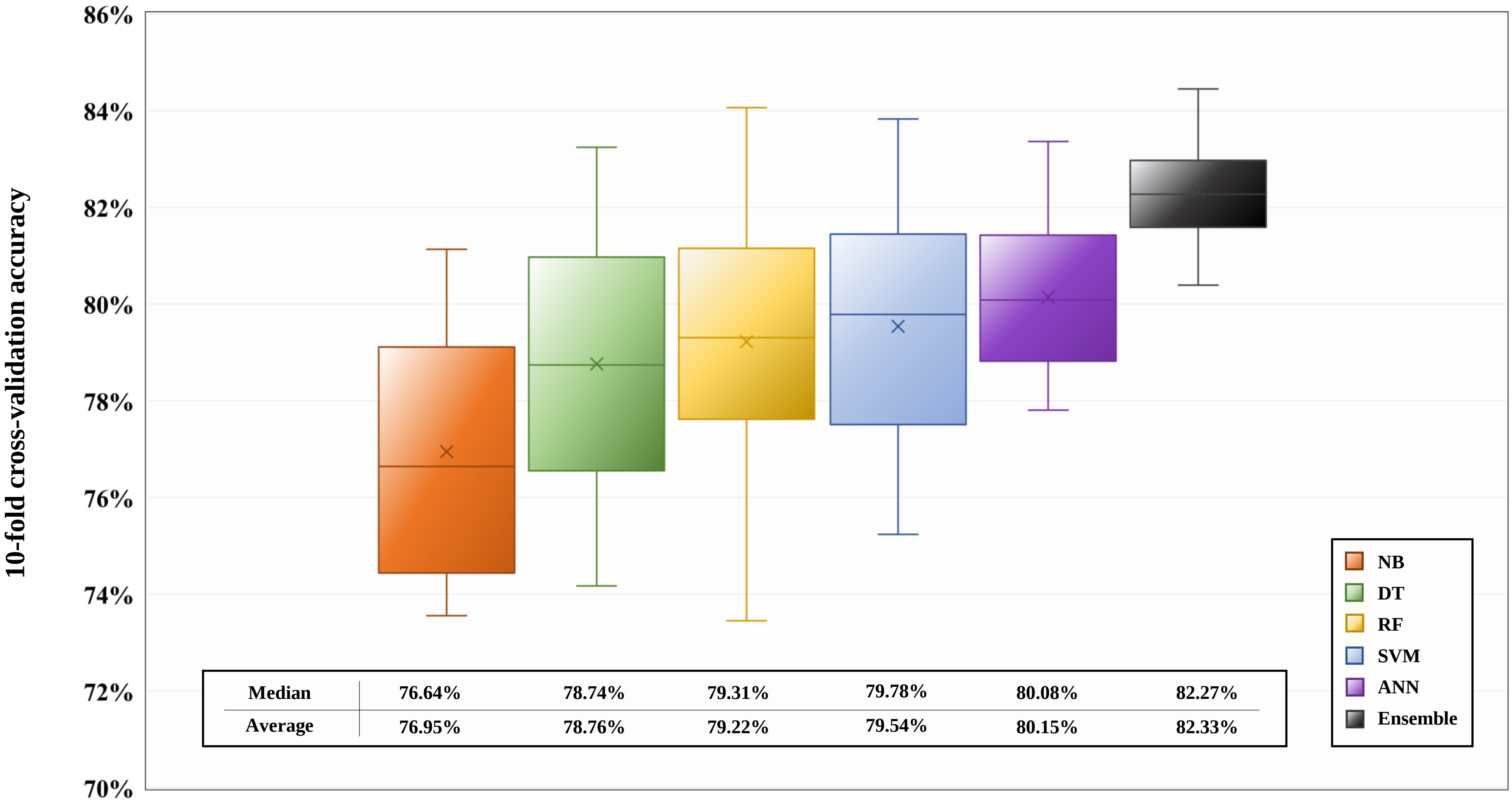
Table 4.7: Algorithm evaluation results (injury type problem)

	Accuracy	Precision	Sensitivity	Specificity	F-measure	S	Ranking
Ensemble	82.62%	78.89%	85.33%	80.28%	81.98%	81.82%	–
ANN	80.35%	76.24%	83.70%	77.46%	79.79%	79.51%	1
SVM	79.60%	76.41%	80.98%	78.40%	78.63%	78.80%	2
RF	79.09%	74.63%	83.15%	75.59%	78.66%	78.23%	3
DT	78.34%	74.50%	80.98%	76.06%	77.60%	77.50%	4
NB	77.33%	72.60%	82.07%	73.24%	77.04%	76.45%	5

Table 4.8: Algorithm evaluation results (body part problem)

	Accuracy	Precision	Sensitivity	Specificity	F-measure	S	Ranking
Ensemble	78.84%	79.01%	78.84%	93.95%	78.93%	81.91%	–
ANN	76.83%	76.93%	76.83%	93.29%	76.88%	80.15%	1
RF	76.57%	76.45%	76.57%	92.49%	76.51%	79.72%	2
SVM	75.31%	75.30%	75.31%	92.20%	75.31%	78.69%	3
NB	73.05%	73.42%	73.05%	92.41%	73.23%	77.03%	4
DT	72.29%	73.23%	72.29%	90.30%	72.76%	76.18%	5

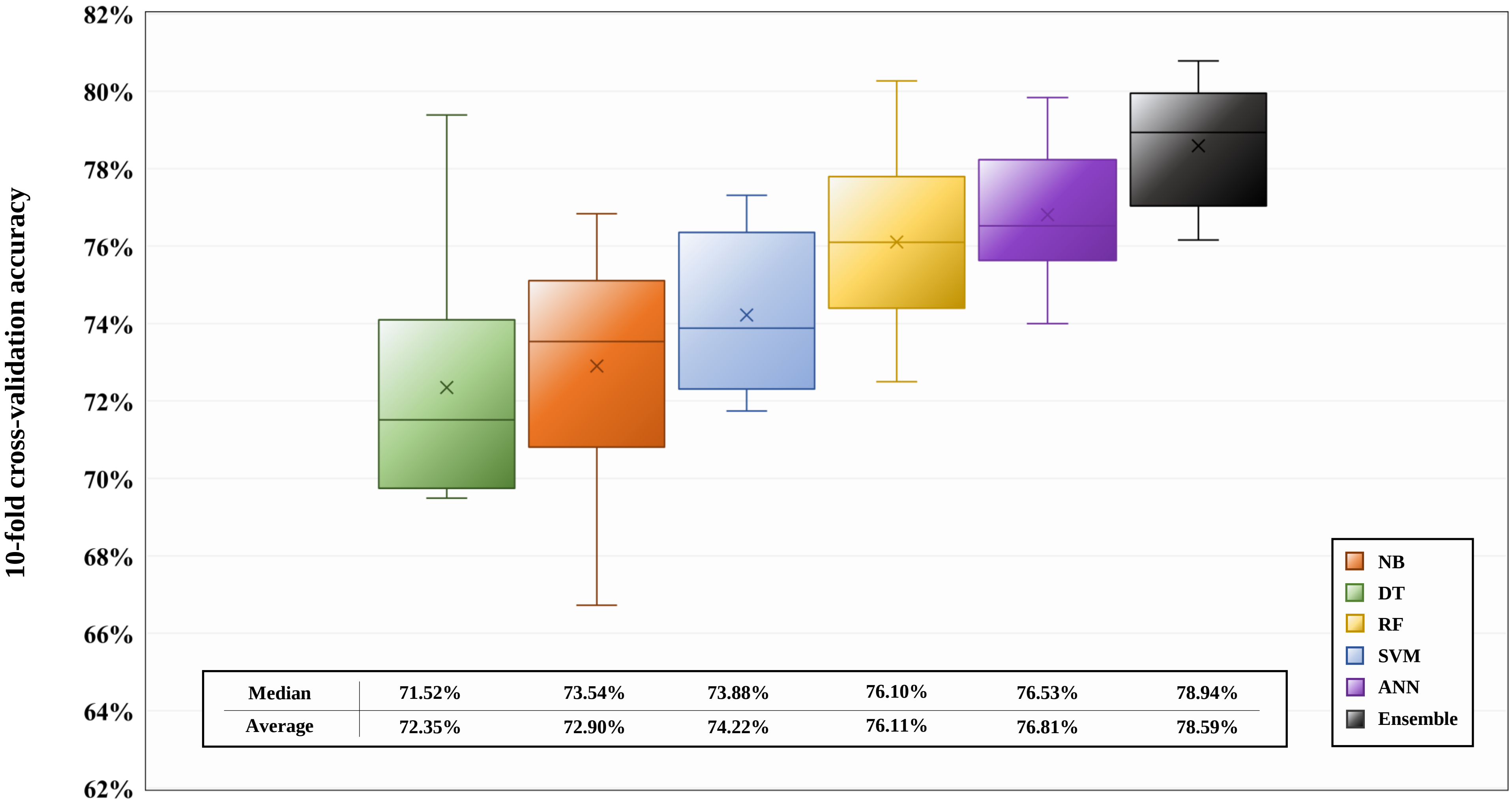
Note. The values presented are class averages.



Median	76.64%	78.74%	79.31%	79.78%	80.08%	82.27%
Average	76.95%	78.76%	79.22%	79.54%	80.15%	82.33%

(a)

Figure 4.8: Algorithm evaluation results of 10-fold cross-validation on entire dataset: a) injury cost problem; and b) body part problem



Median	71.52%	73.54%	73.88%	76.10%	76.53%	78.94%
Average	72.35%	72.90%	74.22%	76.11%	76.81%	78.59%

(b)

Figure 4.8: Algorithm evaluation results of 10-fold cross-validation on entire dataset: a) injury cost problem; and b) body part problem

4.3.5 Site Risk Model

The two ensemble algorithms developed herein to predict injury cost and body part can be used in conjunction with the algorithm developed by Gondia et al. (2021) which predicts only the site risk level. These three algorithms can be deployed to drive a site risk model that supports proactive safety management and real-time decision-making, as shown in Figure 4.9, where input, hidden and output layers would comprise the model.

The input layer can be presented in a checklist form (e.g., left side of Figure 4.9), with pre-job safety inspections and audits serving as the means of collecting work setting characteristics and injury precursor information such as worksite-, work means- and worker potential hazards. The hidden layer (engine) of the model comprises the ML ensemble prediction algorithms that would have been previously trained to learn the relationships between the work setting characteristics/potential injury precursors and the injury outcomes based on a historical dataset, as explained above in detail. Based on this training, the ensemble algorithms can generate predictions of different injury and safety risk leading indicators as per the inputted precursor information from the input layer. The output layer can be presented as a visual interface (e.g., right side of Figure 4.9) that showcases both real-time and proactive values of these predictions across different sites, presented as prediction probabilities of site risk level, injury cost, and body parts most likely to become affected should an accident occur.

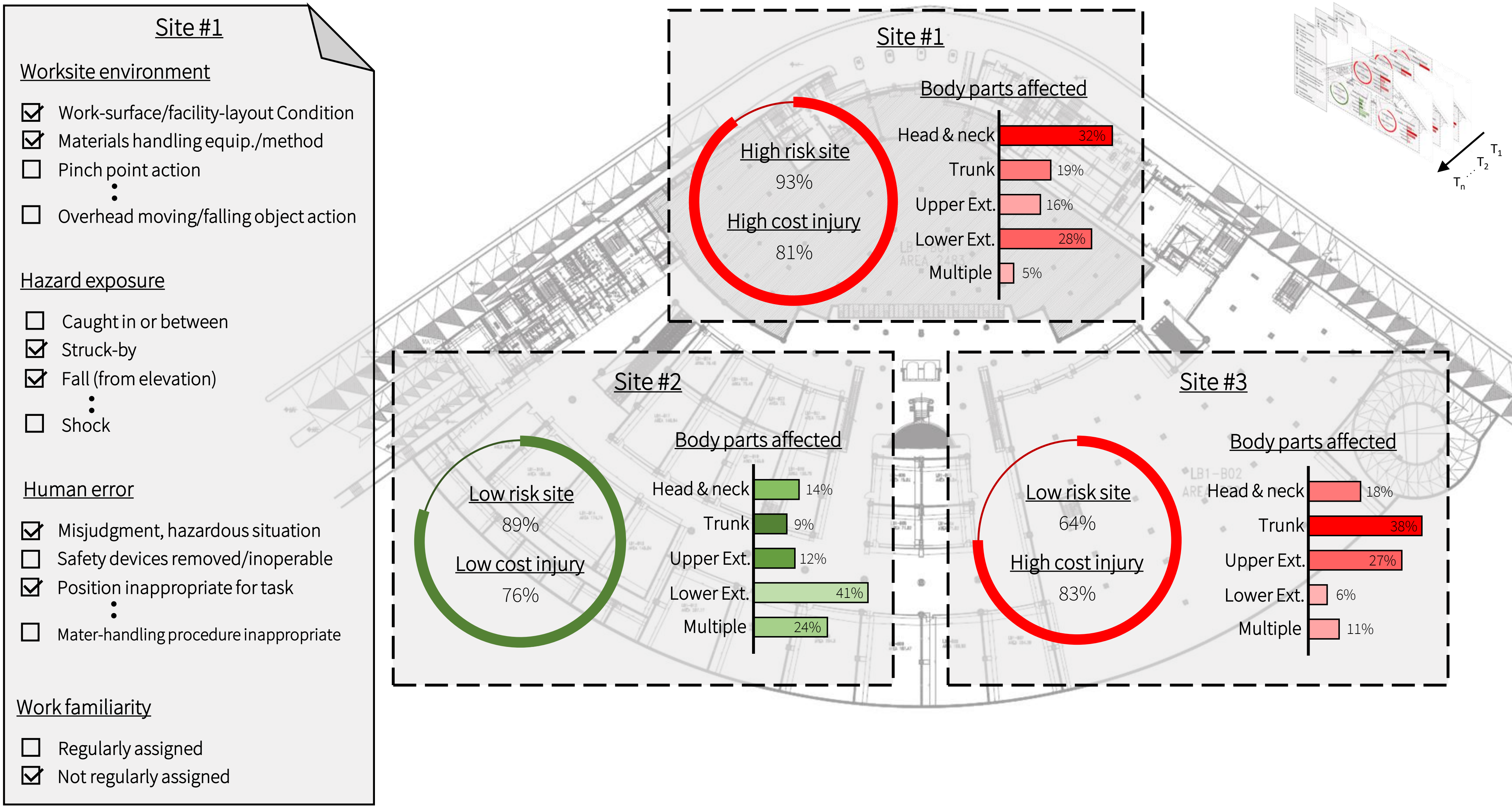


Figure 4.9: Site risk model for proactive safety management and real-time decision making

These leading indicators can indicate which sites exhibit work settings that suggest they are at a heightened risk of injury occurrences and what the financial implications could look like.

4.3.6 Practical Use

It is useful to provide an example of how the model can be practically used by a company that requires a safety oversight/forecast across multiple of its worksites which are scheduled to have work packages starting in October, for instance (see Figure 4.9). Within the considered application, towards the end of August and during preparations for such work packages, site inspections can be carried out and the checklist method can be used to mark the availability of one or more potential incident precursor categories pertaining to the spatial working conditions of the intended site (worksite environment). Through the knowledge of the nature and technical requirements of the work packages, any potential availability of precursor categories associated with the planned work execution methods can also be marked (hazard exposure). Through the knowledge of the assigned workforce's experience and motivation, potential precursor categories for human error and work familiarity can be identified and marked. The month representing prevailing weather conditions will be marked as October in this demonstration example.

These marked sets of precursors are then automatically fed as *inputs* into the ensemble algorithms embedded within the *hidden* layer which would *output* three leading indicators. These leading indicators classify the site into a high- or low-risk site, with the potential of high- or low-cost injuries, and denote which body parts are the most likely to become affected in the case of an incident. The same can be performed across all relevant sites, thus enabling a spatial vision of potential worksite risk levels, as shown in Figure 4.9. With every temporal change (e.g., T_1 , T_2, .. T_n), there can exist a different setting of precursors, and so with timely precursor information collection, the platform can be used as a spatio-temporal tool in the sense described above, providing ongoing decision support in the form of updated/real-time forecasts of leading indicators every month or every work package handover. These indicators can be used to better plan a worksite in time and/or space. For example, sites flagged as high-risk sites and/or prone to high-cost injuries can ultimately be prioritized for additional management attention in the form of more in-depth inspections and intervention actions—all in a proactive manner prior to work commencement. On the worker level, workers can be informed of such risk level classification and most likely affected body parts should an injury occur. The body part leading indicators can also help better plan pre-work planning and safety meetings. For example, a forecasted high probability of an upper extremity injury can heighten targeted discussions around the use of proper gloves for the work; and likewise, a high probability of a head and neck injury can encourage discussions on proper headwear.

Such a model can help company decision-makers and site safety managers 1) accurately predict and proactively intervene to prevent incidents and injuries; 2) prioritize the use of limited safety resources to areas that will make the most impact; 3) enhance their social responsibility goals by communicating action-items and forecasts between management and workers to create a culture of continuous improvement; and 4) save costs by realizing significant reductions in workers' compensation premiums. While the model layers in Figure 4.9 are specific to the demonstration application and dataset considered in the current study (i.e., the ensemble algorithms were trained on a specific dataset of construction injuries described by specific precursors and outcomes), the described ensemble approach can be extended to develop other full end-to-end models that are based on different safety-related applications and datasets within any setting (e.g., manufacturing facility, heavy industry, oil & gas, utility company and office space). An ensemble approach’s good generalization capabilities (e.g., Figure 4.8) support its reliability when applied to different safety-related domains and datasets.

4.4 CONCLUSIONS

Injury statistics position construction among the most dangerous industries in the world, yet safety improvement in the industry has decelerated in the last decade. The current study develops construction site risk models that can generate predictions of safety risk leading indicators across different zones and over project lifecycles. The models are driven by ensemble ML algorithms that, given sets of injury precursors, can predict injury outcomes that act as worksite safety risk leading indicators for supporting proactive actions and real-time monitoring. The developed ensemble algorithms comprised five base algorithms, namely naïve Bayes, decision trees, random forests, support vector machines and artificial neural networks. The training and hyperparameter tuning of each base algorithm were based on cross-validated grid search techniques to prevent overfitting by ensuring the algorithm's accuracy was not constrained to the training data only and help yield a tuned algorithm that generalizes better. When employed, the ensemble algorithm aggregates the predictions of the base algorithms through confidence-weighted average voting. The advantage of the ensemble approach relies on combining the learning strengths of its base algorithms to better shape the final algorithm for capturing systems with complex relationships as those between injury precursors and injury outcomes/leading indicators.

A demonstration application was presented using a construction dataset from the OSHA injury and illness cases database, where ensemble algorithms were trained to predict injuries' financial implications and body parts most likely affected as leading indicators. The algorithm evaluation results supported that the ensemble algorithm could improve classification accuracy and generalizability relative to those of base algorithms across both prediction problems. Therefore, a demonstration was introduced to show how these trained and validated ensemble algorithms can be deployed as part of a site risk model capable of supporting safety managers with proactive and real-time updated leading indicator predictions of worksite safety risks across various sites and timeframes (e.g., site risk level classification, potential injury financial implications, and body parts most likely to be affected) which can inspire key injury-preventive decision making. Extended utilization of such models can bring practical benefits to empower construction companies' safety management strategies in the form of: 1) accurately predicting and proactively intervening to prevent incidents and injuries; 2) prioritizing the use of limited safety resources to areas with high-risk levels; 3) communicating action-items and forecasts between management and workers to create a culture of continuous improvement; and 4) saving costs by realizing significant reductions in workers' compensation premiums.

4.5 ACKNOWLEDGMENTS

The authors are grateful to the financial support of the Natural Science and Engineering Research Council of Canada (NSERC) CaNRisk-CREATE program, and the INTERFACE Institute and the INViSiONLab, both of McMaster University.

4.6 REFERENCES

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

Akram, R., Thaheem, M. J., Nasir, A. R., Ali, T. H., & Khan, S. (2019). Exploring the role of building information modeling in construction safety through science mapping. Safety Science, 120, 456-470.

Alkaissy, M., Arashpour, M., Ashuri, B., Bai, Y., & Hosseini, R. (2020). Safety management in construction: 20 years of risk modeling. Safety science, 129, 104805.

Amiri, M., Ardeshir, A., & Zarandi, M. H. F. (2017). Fuzzy probabilistic expert system for occupational hazard assessment in construction. Safety science, 93, 16-28.

Arditi, D., & Tokdemir, O. B. (1999). Comparison of case-based reasoning and artificial neural networks. Journal of computing in civil engineering, 13(3), 162-169.

Arditi, D., Oksay, F. E., & Tokdemir, O. B. (1998). Predicting the outcome of construction litigation using neural networks. Computer‐Aided Civil and Infrastructure Engineering, 13(2), 75-81.

Ayhan, B. U., & Tokdemir, O. B. (2020). Accident analysis for construction safety using latent class clustering and artificial neural networks. Journal of Construction Engineering and Management, 146(3), 04019114.

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

Behm, M. (2005). Linking construction fatalities to the design for construction safety concept. Safety science, 43(8), 589-611.

Bhowmik, T. K. (2015). Naive bayes vs logistic regression: theory, implementation and experimental validation. Ibero-American Journal of Artificial Intelligence, 18(56), 14-30.

Breiman, L. (1996). Bagging predictors. Machine learning, 24(2), 123-140.

Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. Belmont, CA: Wadsworth. International Group, 432, 151-166.

Bureau of Labor Statistics (2012), Occupational Injury and Illness Classification Manual Version 2.01, U.S. Department of Labor, https://www.bls.gov/iif/oiics_manual_2010.pdf

Bureau of Labor Statistics (2021). Survey of Occupational Injuries and Illnesses Data. Retrieved April 2, 2021 from https://www.bls.gov/iif/soii-data.htm#dafw.

Chen, Q., & Jin, R. (2013). Multilevel safety culture and climate survey for assessing new safety program. Journal of Construction Engineering and Management, 139(7), 805-817.

Chen, Y., McCabe, B., & Hyatt, D. (2018). A resilience safety climate model predicting construction safety performance. Safety science, 109, 434-445.

Cheng, C. W., Leu, S. S., Cheng, Y. M., Wu, T. C., & Lin, C. C. (2012). Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. Accident Analysis & Prevention, 48, 214-222.

Chou, J. S., & Lin, C. (2013). Predicting disputes in public-private partnership projects: Classification and ensemble models. Journal of Computing in Civil Engineering, 27(1), 51-60.

Choudhry, R. M., Fang, D., & Ahmed, S. M. (2008). Safety management in construction: Best practices in Hong Kong. Journal of professional issues in engineering education and practice, 134(1), 20-32.

Choudhry, R. M., Fang, D., & Lingard, H. (2009). Measuring safety climate of a construction company. Journal of construction Engineering and Management, 135(9), 890-899.

Choudhry, R. M., Fang, D., & Mohamed, S. (2007). Developing a model of construction safety culture. Journal of management in engineering, 23(4), 207-212.

Cooper, M. D., & Phillips, R. A. (2004). Exploratory analysis of the safety climate and safety behavior relationship. Journal of safety research, 35(5), 497-512.

CPWR - The Center for Construction Research, and Training. (2018). The construction chart book: The US construction industry and its workers.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. Mathematics of control, signals and systems, 2(4), 303-314.

Davis, J., & Stern, D. (2012). Workers Compensation Claim Frequency: 2012 Update. NCCI Research Brief. Boca Raton, FL: National Council on Compensation Insurance Inc.

Dedobbeleer, N., & Béland, F. (1991). A safety climate measure for construction sites. Journal of safety research, 22(2), 97-103.

Esmaeili, B., & Hallowell, M. (2012). Attribute-based risk model for measuring safety risk of struck-by accidents. In Construction Research Congress 2012: construction challenges in a flat world (289-298).

Fang, D., Chen, Y., & Wong, L. (2006). Safety climate in construction industry: A case study in Hong Kong. Journal of construction engineering and management, 132(6), 573-584.

Feng, Y., Teo, E. A. L., Ling, F. Y. Y., & Low, S. P. (2014). Exploring the interactive effects of safety investments, safety culture and project hazard on safety performance: An empirical analysis. International Journal of Project Management, 32(6), 932-943.

Fiore, A., Quaranta, G., Marano, G. C., & Monti, G. (2016). Evolutionary polynomial regression–based statistical determination of the shear capacity equation for reinforced concrete beams without stirrups. Journal of Computing in Civil Engineering, 30(1), 04014111.

Gandomi, A. H., & Roke, D. A. (2015). Assessment of artificial neural network and genetic programming as predictive tools. Advances in Engineering Software, 88, 63-72.

Gerassis, S., Martín, J. E., García, J. T., Saavedra, A., & Taboada, J. (2017). Bayesian decision tool for the analysis of occupational accidents in the construction of embankments. Journal of construction engineering and management, 143(2), 04016093.

Gholizadeh, P., Esmaeili, B., & Memarian, B. (2018). Evaluating the Performance of Machine Learning Algorithms on Construction Accidents: An Application of ROC Curves. In Construction Research Congress (CRC 2018), ASCE, New Orleans, Louisiana.

Glendon, A. I., & Litherland, D. K. (2001). Safety climate factors, group differences and safety behaviour in road construction. Safety science, 39(3), 157-188.

Goh, Y. M., & Ubeynarayana, C. U. (2017). Construction accident narrative classification: An evaluation of text mining techniques. Accident Analysis & Prevention, 108, 122-130.

Gondia, A., Ezzeldin, M., & El-Dakhakhni, W. (2021). Forthcoming. Machine Learning-based Decision Support Framework for Construction Injury Severity Prediction and Risk Mitigation. Journal of Risk and Uncertainty in Engineering Systems.

Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2019). Machine Learning Algorithms for Construction Projects Delay Risk Prediction. Journal of Construction Engineering and Management, 146(1), 04019085.

Guo, B. H., & Yiu, T. W. (2016). Developing leading indicators to monitor the safety conditions of construction projects. Journal of management in engineering, 32(1), 04015016.

Guo, B. H., Yiu, T. W., & González, V. A. (2015). Identifying behavior patterns of construction safety using system archetypes. Accident Analysis & Prevention, 80, 125-141.

Guo, H., Yu, Y., Xiang, T., Li, H., & Zhang, D. (2017). The availability of wearable-device-based physical data for the measurement of construction workers' psychological status on site: From the perspective of safety management. Automation in Construction, 82, 207-217.

Hallowell, M. R., Hinze, J. W., Baud, K. C., & Wehle, A. (2013). Proactive construction safety control: Measuring, monitoring, and responding to safety leading indicators. Journal of construction engineering and management, 139(10), 04013010.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Hecht-Nielsen, R. (1987). Kolmogorov's mapping neural network existence theorem. In Proceedings of the international conference on Neural Networks (Vol. 3, 11-14). IEEE Press New York.

Ho, T. K. (1998). The Random Subspace Method for Constructing Decision Forests, IEEE T. Pattern. Anal., 20, 832–844.

Hossein Alavi, A., Hossein Gandomi, A., Mollahassani, A., Akbar Heshmati, A., & Rashed, A. (2010). Modeling of maximum dry density and optimum moisture content of stabilized soil using artificial neural networks. Journal of Plant Nutrition and Soil Science, 173(3), 368-379.

Hou, S., & Ramani, K. (2007). Classifier combination for sketch-based 3D part retrieval. Computers & Graphics, 31(4), 598-609.

Hsueh, S. L., Huang, C. F., & Tseng, C. Y. (2013). Using Data Mining Technology to Explore Labor Safety Strategy-A Lesson from the Construction Industry. Pakistan Journal of Statistics, 29(5).

Huang, L., Wu, C., Wang, B., & Ouyang, Q. (2018). Big-data-driven safety decision-making: a conceptual framework and its influencing factors. Safety science, 109, 46-56.

Hwang, S., Seo, J., Jebelli, H., & 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.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112, p. 18). New York: springer.

Jin, R., Zou, P. X., Piroozfar, P., Wood, H., Yang, Y., Yan, L., & Han, Y. (2019). A science mapping approach based review of construction safety research. Safety science, 113, 285-297.

John, G. H., & Langley, P. (2013). Estimating continuous distributions in Bayesian classifiers. arXiv preprint arXiv:1302.4964.

Kakhki, F. D., Freeman, S. A., & Mosher, G. A. (2019). Evaluating machine learning performance in predicting injury severity in agribusiness industries. Safety science, 117, 257-262.

Kang, K., & Ryu, H. (2019). Predicting types of occupational accidents at construction sites in Korea using random forest model. Safety Science, 120, 226-236.

Kim, T., & Chi, S. (2019). Accident case retrieval and analyses: Using natural language processing in the construction industry. Journal of Construction Engineering and Management, 145(3), 04019004.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In Ijcai (Vol. 14, No. 2, 1137-1145).

Kulkarni, P., Londhe, S., & Deo, M. (2017). Artificial neural networks for construction management: a review. Journal of Soft Computing in Civil Engineering, 1(2), 70-88.

Lander, F., Nielsen, K. J., & Lauritsen, J. (2016). Work injury trends during the last three decades in the construction industry. Safety science, 85, 60-66.

Le, Q. T., Pedro, A., & Park, C. S. (2015). A social virtual reality based construction safety education system for experiential learning. Journal of Intelligent & Robotic Systems, 79(3), 487-506.

Liao, C. W., & Chiang, T. L. (2015). The examination of workers' compensation for occupational fatalities in the construction industry. Safety science, 72, 363-370.

Lippmann, R. (1987). An introduction to computing with neural nets. IEEE ASSP magazine, 4(2), 4-22.

Liu, B., Blasch, E., Chen, Y., Shen, D., & Chen, G. (2013, October). Scalable sentiment classification for big data analysis using naive bayes classifier. In 2013 IEEE international conference on big data (99-104). IEEE.

Mangalathu, S., & Jeon, J. S. (2018). Classification of failure mode and prediction of shear strength for reinforced concrete beam-column joints using machine learning techniques. Engineering Structures, 160, 85-94.

Marin, L. S., Lipscomb, H., Cifuentes, M., & Punnett, L. (2019). Perceptions of safety climate across construction personnel: Associations with injury rates. Safety science, 118, 487-496.

Martínez-Aires, M. D., Lopez-Alonso, M., & Martinez-Rojas, M. (2018). Building information modeling and safety management: A systematic review. Safety science, 101, 11-18.

Marucci-Wellman, H. R., Corns, H. L., & Lehto, M. R. (2017). Classifying injury narratives of large administrative databases for surveillance—A practical approach combining machine learning ensembles and human review. Accident Analysis & Prevention, 98, 359-371.

McDonald, D., Price, M. N., Goodrich, J., Nawrocki, E. P., DeSantis, T. Z., Probst, A., ... & Hugenholtz, P. (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. The ISME journal, 6(3), 610-618.

Mearns, K., Whitaker, S. M., & Flin, R. (2003). Safety climate, safety management practice and safety performance in offshore environments. Safety science, 41(8), 641-680.

Meyer, D. (2001). Support vector machines. Porting R to Darwin/X11 and Mac OS X, 1, 23.

Misiurek, K., & Misiurek, B. (2017). Methodology of improving occupational safety in the construction industry on the basis of the TWI program. Safety science, 92, 225-231.

Moayed, F. A., & Shell, R. L. (2011). Application of artificial neural network models in occupational safety and health utilizing ordinal variables. Annals of occupational hygiene, 55(2), 132-142.

Mohamed, S. (2002). Safety climate in construction site environments. Journal of construction engineering and management, 128(5), 375-384.

Mohammadi, A., Tavakolan, M., & Khosravi, Y. (2018). Factors influencing safety performance on construction projects: A review. Safety science, 109, 382-397.

Moon, H., Dawood, N., & Kang, L. (2014a). Development of workspace conflict visualization system using 4D object of work schedule. Advanced Engineering Informatics, 28(1), 50-65.

Moon, H., Kim, H., Kim, C., & Kang, L. (2014b). Development of a schedule-workspace interference management system simultaneously considering the overlap level of parallel schedules and workspaces. Automation in Construction, 39, 93-105.

Moreira, M. W., Rodrigues, J. J., Oliveira, A. M., Saleem, K., & Neto, A. (2016, December). Performance evaluation of predictive classifiers for pregnancy care. In 2016 IEEE Global Communications Conference (GLOBECOM) (1-6). IEEE.

Nakamura, K., & Hong, B. W. (2019). Adaptive weight decay for deep neural networks. IEEE Access, 7, 118857-118865.

Nath, N. D., Akhavian, R., & Behzadan, A. H. (2017). Ergonomic analysis of construction worker's body postures using wearable mobile sensors. Applied ergonomics, 62, 107-117.

Navada, A., Ansari, A. N., Patil, S., & Sonkamble, B. A. (2011, June). Overview of use of decision tree algorithms in machine learning. In 2011 IEEE control and system graduate research colloquium (37-42). IEEE.

Nitsche, P., Stütz, R., Kammer, M., & Maurer, P. (2014). Comparison of machine learning methods for evaluating pavement roughness based on vehicle response. Journal of Computing in Civil Engineering, 28(4), 04014015.

Niu, Y., Lu, W., Xue, F., Liu, D., Chen, K., Fang, D., & Anumba, C. (2019). Towards the “third wave”: An SCO-enabled occupational health and safety management system for construction. Safety science, 111, 213-223.

Occupational Safety and Health Administration (OSHA) (2019). Injuries Illnesses, and Fatalities. Retrieved February 17, 2019 from https://www.bls.gov/iif/oshoiics.htm.

Occupational Safety and Health Administration (OSHA) (2021). Recommended Practices for Safety and Health Programs. Retrieved April 6, 2021 from https://www.osha.gov/safety-management.

Olson, D. L., Delen, D., & Meng, Y. (2012). Comparative analysis of data mining methods for bankruptcy prediction. Decision Support Systems, 52(2), 464-473.

Park, C. S., & Kim, H. J. (2013). A framework for construction safety management and visualization system. Automation in Construction, 33, 95-103.

Patel, D. A., & Jha, K. N. (2015). Neural network approach for safety climate prediction. Journal of management in engineering, 31(6), 05014027.

Poh, C. Q., Ubeynarayana, C. U., & Goh, Y. M. (2018). Safety leading indicators for construction sites: A machine learning approach. Automation in construction, 93, 375-386.

Ripley, B. D. (2002). Modern applied statistics with S. springer.

Rish, I. (2001, August). An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, 41-46).

Rivas, T., Paz, M., Martín, J. E., Matías, J. M., García, J. F., & Taboada, J. (2011). Explaining and predicting workplace accidents using data-mining techniques. Reliability Engineering & System Safety, 96(7), 739-747.

Rosecrance, J., Butler, L., Schwatka, N., & AEP, M. (2011). The role of age on the cause, type, nature and cost of construction injuries. CPWR Small Grant Final Report.

Sacks, R., Perlman, A., & Barak, R. (2013). Construction safety training using immersive virtual reality. Construction Management and Economics, 31(9), 1005-1017.

Sacks, R., Rozenfeld, O., & Rosenfeld, Y. (2009). Spatial and temporal exposure to safety hazards in construction. Journal of construction engineering and management, 135(8), 726-736.

Sacks, R., Whyte, J., Swissa, D., Raviv, G., Zhou, W., & Shapira, A. (2015). Safety by design: dialogues between designers and builders using virtual reality. Construction Management and Economics, 33(1), 55-72.

Sakhakarmi, S., Park, J., & Cho, C. (2019). Enhanced machine learning classification accuracy for scaffolding safety using increased features. Journal of construction engineering and management, 145(2), 04018133.

Salas, R., & Hallowell, M. (2016). Predictive validity of safety leading indicators: Empirical assessment in the oil and gas sector. Journal of construction engineering and management, 142(10), 04016052.

Seong, H., Son, H., & Kim, C. (2018). A comparative study of machine learning classification for color-based safety vest detection on construction-site images. KSCE Journal of Civil Engineering, 22(11), 4254-4262.

Shen, X., & Marks, E. (2016). Near-miss information visualization tool in BIM for construction safety. Journal of construction engineering and management, 142(4), 04015100.

Somvanshi, M., Chavan, P., Tambade, S., & Shinde, S. V. (2016, August). A review of machine learning techniques using decision tree and support vector machine. In 2016 international conference on computing communication control and automation (ICCUBEA) (1-7). IEEE.

Son, H., Kim, C., Hwang, N., Kim, C., & Kang, Y. (2014). Classification of major construction materials in construction environments using ensemble classifiers. Advanced Engineering Informatics, 28(1), 1-10.

Sun, Y., Kamel, M. S., Wong, A. K., & Wang, Y. (2007). Cost-sensitive boosting for classification of imbalanced data. Pattern Recognition, 40(12), 3358-3378.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016a). Application of machine learning to construction injury prediction. Automation in construction, 69, 102-114.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016b). Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports. Automation in Construction, 62, 45-56.

Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2017). Construction safety clash detection: identifying safety incompatibilities among fundamental attributes using data mining. Automation in Construction, 74, 39-54.

Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE transactions on neural networks, 10(5), 988-999.

Villanova, M. P. (2014). Attribute-based risk model for assessing risk to industrial construction tasks (Doctoral dissertation, University of Colorado at Boulder).

Wang, P., Wu, P., Wang, J., Chi, H. L., & 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.

Waziri, B. S., Bala, K., & Bustani, S. A. (2017). Artificial neural networks in construction engineering and management. International Journal of Architecture, Engineering and Construction, 6(1), 50-60.

Yang, K., Ahn, C. R., Vuran, M. C., & Aria, S. S. (2016). Semi-supervised near-miss fall detection for ironworkers with a wearable inertial measurement unit. Automation in Construction, 68, 194-202.

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

Zhang, P., Wu, H. N., Chen, R. P., & Chan, T. H. (2020). Hybrid meta-heuristic and machine learning algorithms for tunneling-induced settlement prediction: A comparative study. Tunnelling and Underground Space Technology, 99, 103383.

Zhou, H., Li, X., Huang, D., Yang, Y., & Ren, F. (2011). Voting-based ensemble classifiers to detect hedges and their scopes in biomedical texts. IEICE TRANSACTIONS on Information and Systems, 94(10), 1989-1997.

Zhou, Y., Li, S., Zhou, C., & Luo, H. (2019). Intelligent approach based on random forest for safety risk prediction of deep foundation pit in subway stations. Journal of Computing in Civil Engineering, 33(1), 05018004.

Zhou, Y., Su, W., Ding, L., Luo, H., & Love, P. E. (2017). Predicting safety risks in deep foundation pits in subway infrastructure projects: support vector machine approach. Journal of Computing in Civil Engineering, 31(5), 04017052.

Zohar, D. (1980). Safety climate in industrial organizations: theoretical and applied implications. Journal of applied psychology, 65(1), 96.

Zou, P. X. (2011). Fostering a strong construction safety culture. Leadership and Management in Engineering, 11(1), 11-22.

Zou, Y., Kiviniemi, A., & Jones, S. W. (2017). Retrieving similar cases for construction project risk management using Natural Language Processing techniques. Automation in construction, 80, 66-76.

Zur, R. M., Jiang, Y., Pesce, L. L., & Drukker, K. (2009). Noise injection for training artificial neural networks: A comparison with weight decay and early stopping. Medical physics, 36(10), 4810-4818.

Chapter 5:
DYNAMIC NETWORKS FOR RESILIENCE-DRIVEN MANAGEMENT OF INFRASTRUCTURE PROJECTS

ABSTRACT

The importance of ensuring the resilience (rapid adaption to and recovery from disruptions) of infrastructure projects in modern societies can be hardly overstated. However, using currently available tools, managing such projects continues to be challenging because of their intrinsic complexities and dynamic spatiotemporal interdependencies. In this respect, the objective of the current study is to develop a novel framework for resilience-driven management of infrastructure projects. By-design, the framework can ensure project resilience through mitigating the risk of complex interdependence-induced vulnerabilities and subsequent cascade disruptions of project performance. The framework adopts a dynamic network approach to model and analyze spatiotemporal contractor interdependence within infrastructure project sites. Subsequently, the framework harnesses the power of metaheuristic optimization techniques to proactively detect interdependence-induced vulnerabilities and rapidly adapt contractor networks accordingly. Such adaption will then reflect on the work schedule—enhancing the overall project resilience to possible performance cascade disruptions. Finally, to demonstrate the applicability of the developed

framework, a several hundred million dollars power infrastructure overhaul project of high strategic importance was considered. By examining complex infrastructure projects through a network-level lens, the framework provides project managers with key managerial insights and deepened understanding of their projects' underlying interdependence-induced vulnerabilities and possible shortcomings of preset risk mitigation strategies. Overall, the current study empowers managers through its *resilient-by-design* approach to infrastructure projects in order to absorb, recover from, and adapt to disruptive events persistently triggered by the project's dynamic risk environment.

5.1 INTRODUCTION

Public infrastructure systems (e.g., power, water/wastewater, telecommunication, and transportation) function as arteries of modern urban communities as they provide vital services to meet societal, economic, and political needs (Di Maddaloni and Davis 2018). Construction, operation, overhaul/refurbishment, retrofit and expansion projects of such systems typically: 1) have long schedules and large budgets; 2) require substantial resources, extensive work scopes and diverse specialized expertise; 3) spread spatially over a large geographical area; 4) result in significant socio-economic impacts; and 5) attract high private- and public sector engagements (Sun and Zhang 2011; Flyvbjerg 2014). In this respect, managing such infrastructure projects is perceived to be particularly challenging and risky due to their inherent complex interdependence and their dynamic (time-dependent) nature (Aritua et al. 2009; Mok et al. 2015). As a result, projects' inability to meet basic targets of duration, budget, benefits realization, and subsequently stakeholder satisfaction, has been well recognized (Yeo 1995; Han et al. 2009; Cantarelli et al. 2012; Eriksson et al. 2017; McKinsey Global Institute 2017).

A key facet of infrastructure projects' inherent complexity lies in their different ***task technical complexities*** as their associated work scopes typically require high degrees of technical know-how, specialized skillsets, and significant multidisciplinary collaborative efforts (Luo et al. 2016). As such, this technical complexity leads to ***contractor-related interdependence***, where the sites of the project typically accommodate numerous specialized contractor crews from different disciplines (e.g., mechanical, electrical, fire protection, steel fabricator, etc.), each with their own interests, constraints and uncertainties. These crews also work simultaneously (and not interchangeably) on heavily interdependent, successive and overlapping tasks over extended periods (Chester and Hendrickson 2005; Jarkas 2017). Contractor-related interdependence also influences ***performance-related interdependence***, where project performance objectives are also highly intertwined. Specifically, unless the project tasks are properly coordinated (i.e., contractor-related interdependence is adequately managed and controlled), execution errors, quality deficits, delays in approval, and subsequently, other interdependence-induced project cascade performance disruptions (**CPD**), as shown in Figure 5.1, are likely to spread throughout (Serrador and Turner 2015; Eriksson et al. 2017; Gondia et al. 2020). For illustration purposes, Figure 5.1a represents several key CPD as components of a "small-world network", where all network components are interconnected. Therefore, different disruptions have the potential to induce others, in a cascade manner possibly extending to all project components, as illustrated in Figure 5.1b. For instance, immediate execution errors by one contractor in the network may not only affect those contractors within the closest task-dependent proximity but might also trigger multiple additional CPD to other contractors within the network. Such errors typically require multiple reworks—hindering productivity

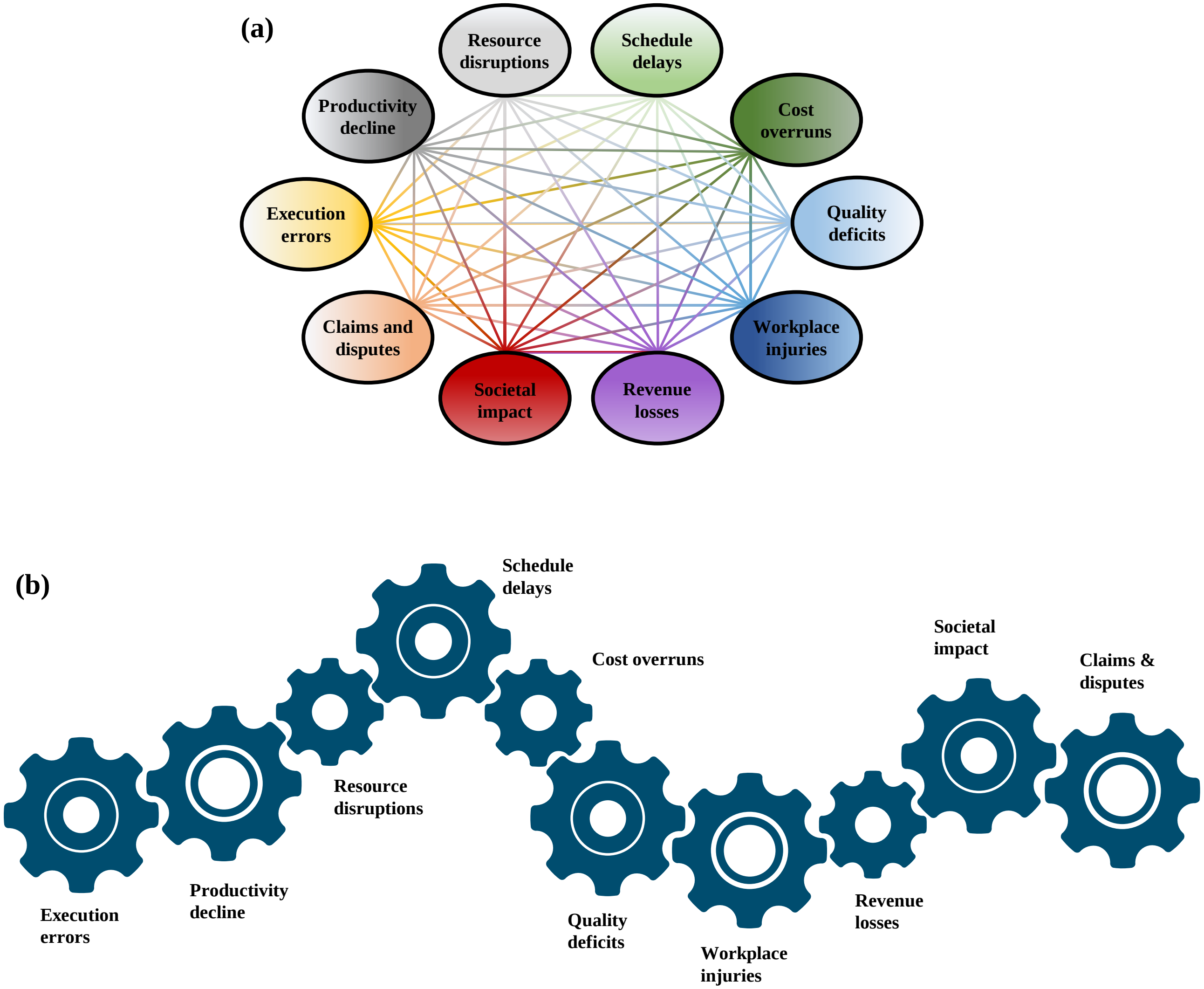
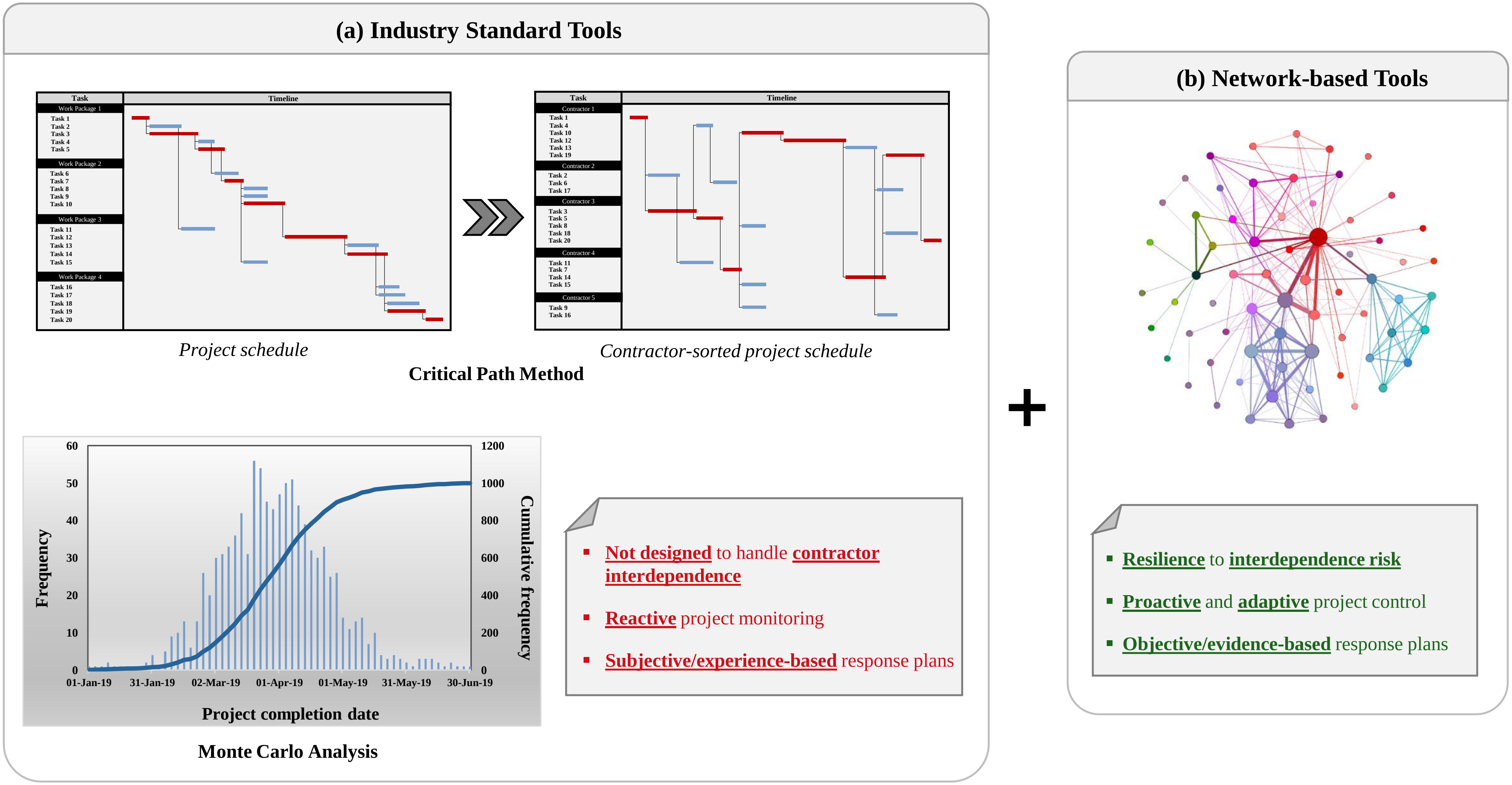


Figure 5.1: Project performance disruptions' (a) interdependent nature and (b) cascade potential

and disrupting resource allocations to multiple contractors and tasks, and subsequently incurring schedule delays and cost overruns (Larsen et al. 2015). As they attempt to adapt to such CPD, project managers typically resort to accelerating progress through compressing schedule and/or crashing tasks, which may also have a negative impact on the finished work quality and the safety of the workers— inducing further project complications (Nepal et al. 2006; Love at al. 2016). Such disruptions may ultimately result in delayed infrastructure project completion and operation-readiness, leaving governments and other stakeholders with lost revenues on project capital, and subsequently provoking negative societal perceptions and public controversies (Ndekugri et al. 2008; Di Maddaloni and Davis 2018). Such consequences, in turn, spark tensions between project stakeholders, where unresolved conflicts give rise to legal claims and disputes which have become increasingly common in infrastructure projects (Yates and Epstein 2006; Mehany et al. 2018).

Industry standard and commercially available software tools, widely used for project management and control, typically employ the critical path method (**CPM**) and the Monte Carlo analysis (see Figure 5.2a). The former is customarily used for planning, scheduling and controlling of projects task durations, resource allocations and costs, whereas the latter is employed to enable further probabilistic modelling and risk analyses to the former’s outcomes in order to quantify the levels of reliability associated with such outcomes (Nasir et al. 2003;



(a) Industry Standard Tools

Project schedule

Contractor-sorted project schedule

Critical Path Method

Monte Carlo Analysis

- **Not designed** to handle **contractor interdependence**
- **Reactive** project monitoring
- **Subjective/experience-based** response plans

+

(b) Network-based Tools

- **Resilience** to **interdependence risk**
- **Proactive** and **adaptive** project control
- **Objective/evidence-based** response plans

Figure 5.2: Project management and control tools: Complementing (a) Industry standard tools with (b) Network-based tools

Tomczak and Jaśkowski 2020). Despite providing valuable insights to support the management of small- and medium-sized projects, these current tools suffer from **three main underlying limitations** when adopted to manage complex infrastructure projects, as outlined in Figure 5.2a. ***First***, such tools are not particularly designed to visualize and analyze contractor-related interdependence. Specifically, commercial CPM tools, such as Microsoft Project or Oracle Primavera P6, enable visualizing interdependence at the project task-level by connecting user-defined predecessor and successor tasks within project schedules, as shown in Figure 5.2a. Such tools also allow tasks to be sorted by contractor responsibility (i.e., through grouping all tasks assigned to a particular contractor), as shown in Figure 5.2a, which can be used to understand work schedules and scope of work (overall duration involved) for specific contractors within the project. However, focusing on interdependence, representing such a contractor-sorted schedule in the form of a very large sheet introduces too many details (taskbars) to remain comprehensible (Lu and AbouRizk 2000; Aritua et al. 2009; Tomczak and Jaśkowski 2020). Subsequently, such sorting does not lend itself to concise visualization or analyses of aggregated and/or quantified measures describing interdependence between contractors, which in turn can reveal specific critical contractors and interdependent contractor collaborations. In terms of performance-related interdependence, such tools are also unable to recognize the full extent of interrelationships between CPD, where they only optimize performance with respect to achieving a single objective (i.e., minimum completion duration or optimum resource usage levels), thus disregarding the influences of other performance measures such as, for example, quality and safety (Chassiakos and Sakellaropoulos 2005; Ipsilandis 2007). ***Second***, such tools can reveal predictions of project completion durations and budgets only following periodic schedule/budget updates, occasional scope changes or infrequent project risk level reassessments (Hazir 2015; Project Management Institute 2017). As a result, such tools are mainly useful to assist project managers in periodic/progressive project monitoring which typically allows for only reactive response plans against performance disruptions as they arise (Avlijaš 2019). ***Third***, such tools leave project managers dependent solely on their own (subjective) expertise to find, prepare and execute response plans for project task and contractor coordination in order to rectify any oversimplified predictions of project performance disruptions revealed through periodic updates (Goldstein 2006; Sadeghi et al. 2010). However, decision-making based solely on subjective- and experience-driven reasoning often does not lend itself to systematic strategic project management or to develop systemic robust solutions. As such, such an approach to project management often leads to uncoordinated response plans with resulting lost functionalities and slow project performance recoveries (Loizou and French 2012). Notwithstanding the value and indispensability of current analysis tools, the above three limitations raise the need to complement such tools with additional project resilience-focused tools —which can be achieved through adopting more capable complex systems simulation approaches (Figure 5.2b).

A multidisciplinary concept, ***resilience*** of a system, organization or community in the face of disruptive events denotes the system's ability to: 1) absorb the impacts of such disruptions through prior identification of systemic vulnerabilities and proactive preparedness ; 2) adapt to such events by mobilizing risk management strategies aimed at preserving continued performance; and 3) rapidly recover from such events and restore the pre-disruption performance state through fully operationalizing the developed response strategies (Barker et al. 2013; Hernandez-Fajardo and Dueñas-Osorio 2013; Wilkinson et al. 2016; Hariri-Ardebili 2018). Considering the above, the *resilient-by-design* project management approach, proposed in the current study, aims at enhancing the project's ability to absorb systemic risks attributed to contractor-interdependence-induced vulnerabilities and develop adaptive solutions against the resulting CPD, thus facilitating a rapid restoration of the most important set of project performance objectives. In this fulfillment, approaches related to complex dynamic network theory (**CDNT**) offer an ideal suite of tools to model, analyze and understand dynamic interdependencies in complex systems (both over space and time) and subsequently adapt their behaviors (Barker and Haimes 2009; Gong et al. 2017; Fu et al. 2018) to mitigate possible cascade (systemic) risks.

5.2 STUDY GOAL AND OBJECTIVES

The goal of the current study is to develop a resilience-driven management framework that ensures project resilience by-design through the adoption of network analyses and manipulations to complement the current industry-standard tools (see Figure 5.2b). Specifically, this framework aims at enhancing infrastructure projects' ability to rapidly overcome possible CPD through providing project managers with proactive, dynamically adaptive and objective response plans with respect to contractor and task re-coordination. Within this goal, this study focuses on achieving the following three key objectives: 1) ***modelling*** the complex dynamic ***interdependencies*** of contactors' tasks, both spatially (i.e., based on physical site location) and temporally (i.e., as the project schedule progresses), using CDNT approaches; 2) analyzing the behaviors of these networks and ***proactively*** assessing possible CPD; and 3) facilitating data-driven ***adaptive*** (self-organized) network recovery through re-coordinated contractor tasks to ensure project resilience against CPD.

In fulfillment of the stated research goal and objectives, the current study is organized into five main sections (Sections 5.3 to 5.7). In Section 5.3, an understanding of CDNT and relevant CDNT-based measures are provided, after which how corresponding project-level managerial insights can be drawn from such measures are highlighted. In Section 5.4, the resilience-driven infrastructure project management framework for adaptive contractor re-coordination against

CPD using a CDNT approach is presented, and the procedures related to the framework’s five underlying tools are described. Finally, to demonstrate the implementation of the developed framework, an actual large-scale infrastructure project is considered. In that respect, Section 5.5 features a description of the project and its strategic importance of compliance with performance targets, followed by a demonstration of the framework tools and procedures to model and analyze the project data using the afore-described CDNT-based measures. Subsequently, in Section 5.6, the results of the analysis are described to reveal several insights that further both the comprehensive and granular understandings of the project in terms of key contractor influences, namely their interdependence-induced (collaboration) vulnerabilities, challenging months, and critical work packages—insights that would typically not have been revealed using solely available industry-standard tools. As a closure, Section 5.7 discusses some overall generalizable insights that can be inferred from the developed framework, aggregated by project stage (i.e., from commencement to completion) and are categorized by insights gained by different project stakeholders such as contractors and project managers among others.

5.3 COMPLEX DYNAMIC NETWORK MEASURES

CDNT facilitates modelling and visualization of the dynamic relationships and complex interdependencies and behaviors within spatiotemporal systems through network layouts of nodes and links (Sarkar and Moore 2006; Zhu and Mostafavi 2015; Fu et al. 2018; Duan and Ayyub 2020). This section provides an understanding of some relevant network-based measures and highlights how these measures facilitate drawing unique project-level managerial insights.

Within the context of the current study, nodes represent contractors bound by a specific project-based collaboration, while links represent these contractor interdependencies if they collaborate on tasks within the same spatiotemporal setting. The networks considered herein are both: a) *undirected*, since task interdependence between any pair of contractors affects them (concurrently) in a two-way/mutual relationship; and b) *weighted*, since the level/magnitude of task interdependence can vary widely among the contractors on site. Furthermore, since actual infrastructure project contractors' interactions vary within the project site (space) locations and project lifecycle (time) stages, the spatio-temporal network approach adopted in the current study is key to examine the project in its evolution through multiple spatiotemporal stages (McGee et al. 2019; Lu et al. 2020).

Once the contractor interdependence network is constructed, numerical analyses are key to better understand the relationships (interdependence) between the different network nodes (contractors) and hence to uncover critical vulnerabilities. There is a wide range of complex network analytic measures, including those specific to network components (nodes) and those that are systemic (network-level). For ***node-specific*** measures, *centralities* present insightful measures that describe the relative importance of nodes in the network by assessing how connected such nodes are to other nodes. Several types of centrality measures exist all with their own unique characteristics, including:

- **Weighted degree centrality (WDC)** (Dueñas-Osorio et al. 2007; Park et al. 2010; Gong et al. 2017; Xue et al. 2018; Ezzeldin and El-Dakhakhni 2019): Representing the total summation of weights of the links connected to the underlying node. This type of centrality quantifies the node's ability to influence other network nodes. Since it reflects direct influence, within infrastructure project management, this centrality is key for capturing the criticality of specific contractors pertaining to the magnitude of interdependent shared tasks within the network.

- **Betweenness centrality (BC)** (Park et al. 2010; Gong et al. 2017; Schröpfer et al. 2017; Yu et al. 2017; Xue et al. 2018): Reflecting the number of shortest paths between connected pairs of network nodes that pass through the underlying node. This type of centrality characterizes how important a node is in connecting other pairs of nodes. In project management, such centrality can be used to capture the importance of a contractor in terms of controlling the flow of work on-site through task continuity (i.e., tasks received from or handed-over to other contractors).

- **Closeness centrality (CC)** (Park et al. 2010; Kao et al. 2017; Xue et al. 2018; Ezzeldin and El-Dakhakhni 2019): Denoting the number of links on the shortest paths from the underlying node to all other nodes in the network. This type of centrality essentially portrays how relatively close the node is to the rest of other nodes. In project management, this centrality can be used to represent the importance of contractors in terms of the magnitude of direct and indirect involvements they may have with other contractors' work scopes.

- **Eigenvector centrality (EC)** (Estrada and Knight 2015; Gong et al. 2017; Kao et al. 2017): Quantifying the extent to which a node is connected to other influential (high centrality) nodes. This type of centrality is used to investigate how connected the node is to very- versus not very important nodes. In project management, this centrality can be used if there is prior management knowledge that highly central contractors are only exclusively participating with a few other contractors.

Unlike those that are node-specific, ***network-level*** measures are related to the overall network structure, and the most relevant of these measures to infrastructure project management include:

- **Network density (ND)** (Park et al. 2010; Wehbe et al. 2016; Kereri and Harper 2019): Indicating the ratio of the actual number of links in the network to their maximum possible number. This measure, which reflects how well connected and cohesive a network is, has values ranging from zero to one, where zero denotes the network nodes are completely unconnected and one indicates full connectivity. In infrastructure projects, this measure can be used to assess the level of involvement among unique contractors' work scopes.

- **Average weighted degree centrality (AWDC)** (Estrada and Knight 2015; Wehbe et al. 2016; Lu et al. 2020): Normalizing the summation of the weighted degree centrality values for all network nodes by their number, reflecting how fast disruptions can cascade throughout a network. In project management, this measure can be employed to gain a deeper understanding of contractors' site-interdependence, thus reflecting the network-level magnitude of specific task interdependencies and thus their vulnerability to CPD.

5.4 FRAMEWORK ARCHITECTURE

Figure 5.3 shows the tools comprising the proposed resilience-driven infrastructure project management framework for adaptive contractor re-coordination against CPD. The framework encompasses the following five tools: 1) Interdependence quantification and snapshot network modelling; 2) Dynamic network modelling; 3) Dynamic network centrality analyses; 4) Network measure versus project key performance indicator (**KPI**) correlation analyses; and 5) Adaptive (self-organized) network development.

5.4.1 Project Schedule Pre-processing

The data source on which the framework tools are applied is the project schedule which typically encapsulates abstract yet valuable interdependence data. In this respect, the schedule first needs to undergo some key data pre-processing steps to prepare the interdependence data to be incorporated in the network-based tools. As stated earlier, contractor interdependencies within the current focus are understood as contractor collaborations on tasks within the same spatiotemporal setting. As such, one network layout (snapshot) is produced for each group of contractors working together within a specific project site (spatial setting) and for a definite time-window (temporal setting). Therefore, tasks in the project schedule need to be segmented by site location and time-window and further sorted by each contractor assignment.

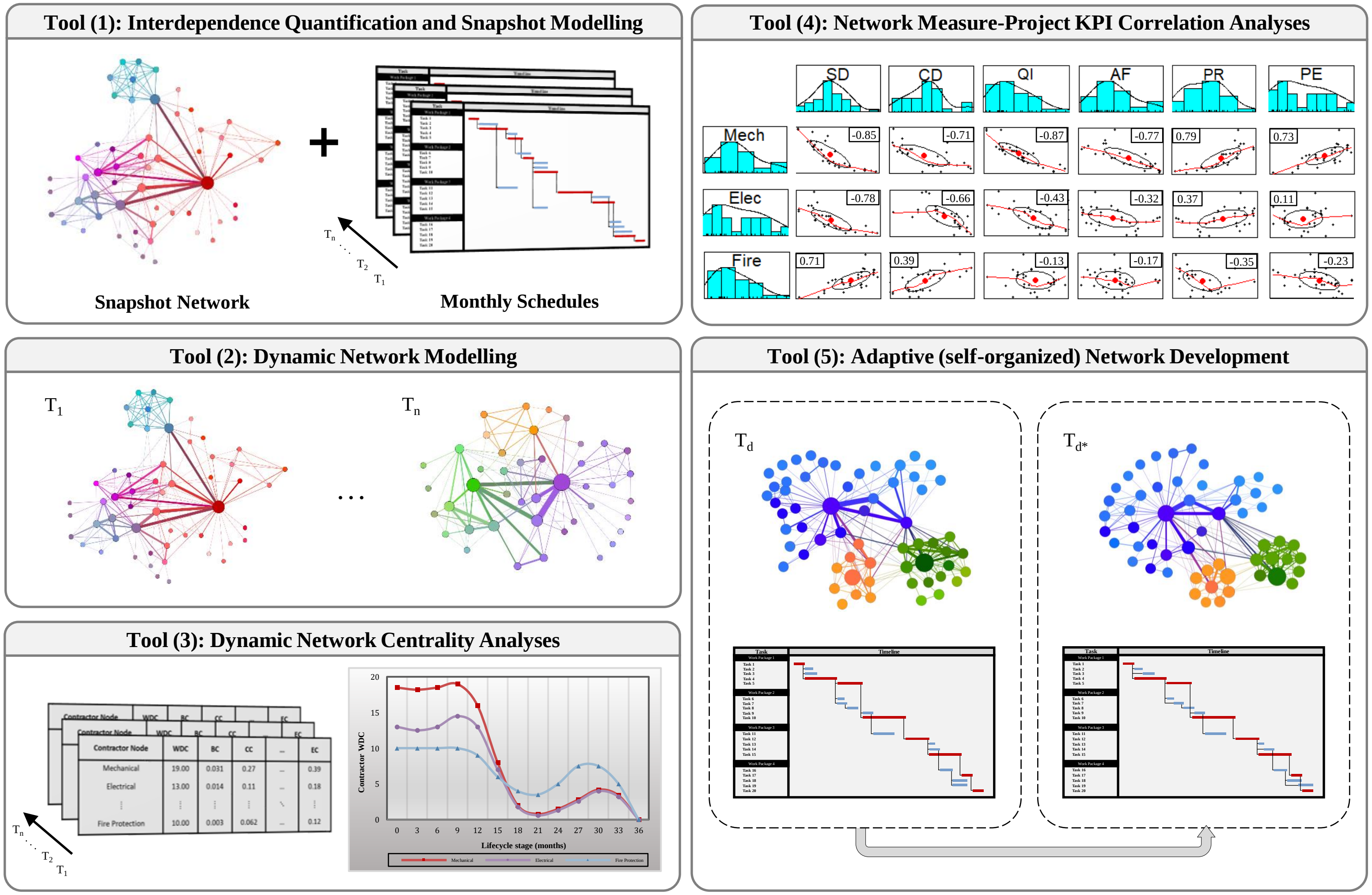
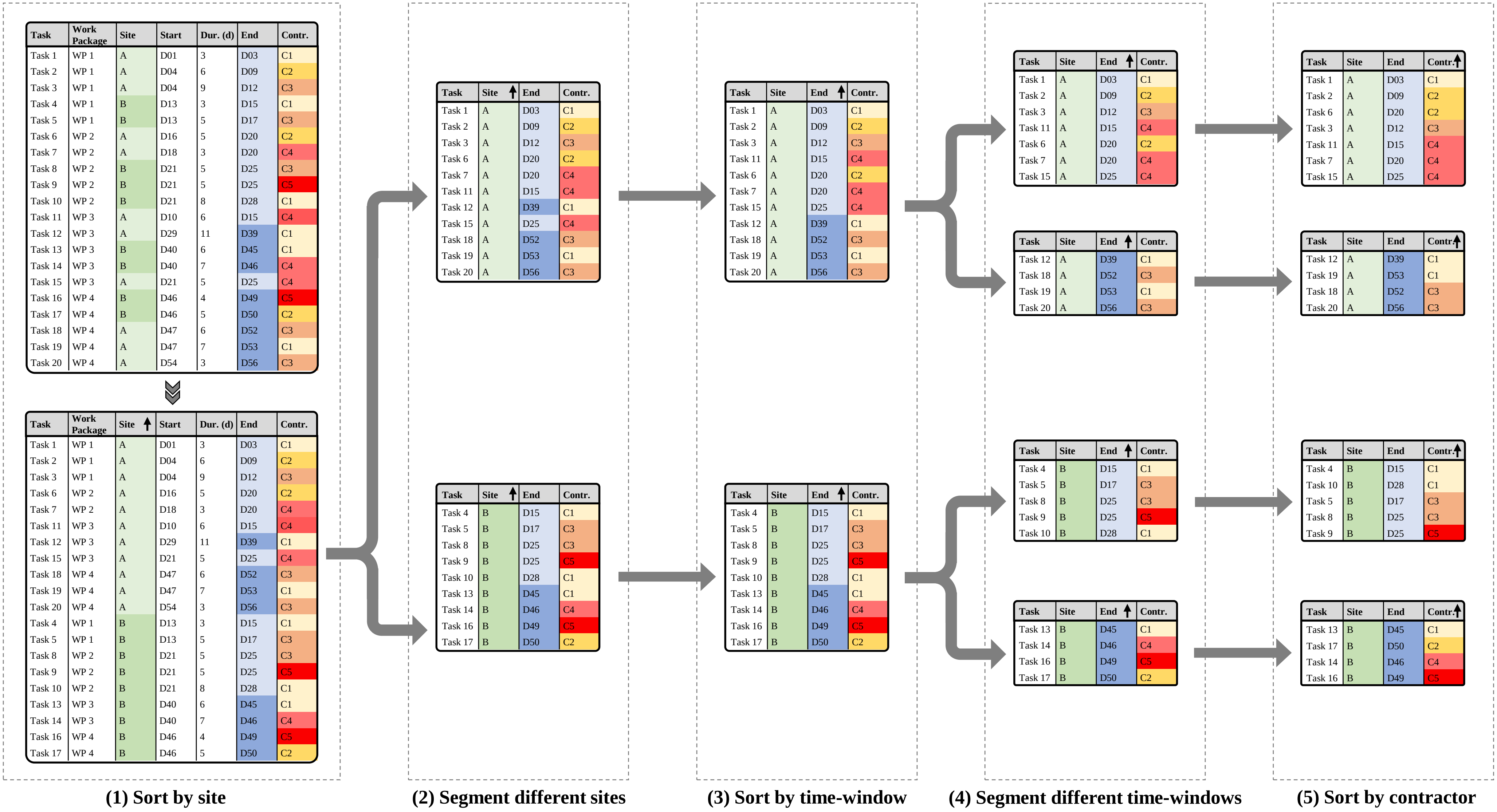


Figure 5.3: Infrastructure project resilience-driven management framework for adaptive contractor re-coordination against CPD

The time-window represents a time interval during which contractor interdependencies are aggregated into one snapshot. The size of the time-window by which the project lifecycle is segmented can be selected as, for instance, a yearly quarter, month, week, or any other time interval based on the project duration and purpose of the analysis. Although reducing the size of the time-window increases temporal resolution and thus yields more network snapshots, managerial judgment is needed to ensure that each snapshot aggregates all contractor interdependencies during the selected time-window. For example, a too-small time-window can yield a snapshot with no network structure and hence little meaningful project insights. For demonstration, the current study employs monthly time-windows throughout the project lifecycle as will be shown later.

The project schedule segmentation can be carried out using two types of tools. First, commercial CPM tools (e.g., Microsoft Project or Oracle Primavera P6) allow for grouping and sorting project tasks based on different (and more than one) criteria/attributes. Although the default grouping is based on the work package (e.g., Figure 5.2a), tasks can also be grouped by site location (grouping level 1) and then by contractor responsibility (e.g., Figure 5.2a) (grouping level 2). Subsequently, these tools' filtering capabilities can be utilized to segment the full schedule by each site and by each time-window (e.g., month) through the filter controls. Finally, the different schedule segments can be exported in a tabular format. Alternatively, standard numerical computing or spreadsheet tools can be used to achieve the same schedule segmentation procedure on the full schedule in a tabular format, as demonstrated in Figure 5.4.

For demonstration, Figure 5.4 shows the Gantt chart from Figure 5.2a in a tabular format, where each task is attributed by its work package, site location where it is performed, start date (e.g., D04 means the fourth day since the project start), duration (e.g., in days), end date and assigned contractor (e.g., C1 to C5). The procedure can be described in five stages. In Stage 1, the tasks in the table are sorted by the site to create a clear distinction between the tasks performed on the two project sites, A and B, selected in this example. In Stage 2, tasks in each site are segmented in separate tables. For Stage 3, tasks in each table are sorted by their end date to distinguish between different time-windows (e.g., in months). In this example, tasks ending in the second month of the project are considered falling within the second time-window. In Stage 4, tasks in each time-window are segmented in separate tables. Finally, in Stage 5, tasks in each table are sorted by contractor to group contractor-specific tasks together. Ultimately, each of the four segmented tables at the end of the procedure is used to generate four corresponding network layouts—two monthly snapshots per site.



Task	Work Package	Site	Start	Dur. (d)	End	Contr.
Task 1	WP 1	A	D01	3	D03	C1
Task 2	WP 1	A	D04	6	D09	C2
Task 3	WP 1	A	D04	9	D12	C3
Task 4	WP 1	B	D13	3	D15	C1
Task 5	WP 1	B	D13	5	D17	C3
Task 6	WP 2	A	D16	5	D20	C2
Task 7	WP 2	A	D18	3	D20	C4
Task 8	WP 2	B	D21	5	D25	C3
Task 9	WP 2	B	D21	5	D25	C5
Task 10	WP 2	B	D21	8	D28	C1
Task 11	WP 3	A	D10	6	D15	C4
Task 12	WP 3	A	D29	11	D39	C1
Task 13	WP 3	B	D40	6	D45	C1
Task 14	WP 3	B	D40	7	D46	C4
Task 15	WP 3	A	D21	5	D25	C4
Task 16	WP 4	B	D46	4	D49	C5
Task 17	WP 4	B	D46	5	D50	C2
Task 18	WP 4	A	D47	6	D52	C3
Task 19	WP 4	A	D47	7	D53	C1
Task 20	WP 4	A	D54	3	D56	C3

Task	Work Package	Site ↑	Start	Dur. (d)	End	Contr.
Task 1	WP 1	A	D01	3	D03	C1
Task 2	WP 1	A	D04	6	D09	C2
Task 3	WP 1	A	D04	9	D12	C3
Task 6	WP 2	A	D16	5	D20	C2
Task 7	WP 2	A	D18	3	D20	C4
Task 11	WP 3	A	D10	6	D15	C4
Task 12	WP 3	A	D29	11	D39	C1
Task 15	WP 3	A	D21	5	D25	C4
Task 18	WP 4	A	D47	6	D52	C3
Task 19	WP 4	A	D47	7	D53	C1
Task 20	WP 4	A	D54	3	D56	C3
Task 4	WP 1	B	D13	3	D15	C1
Task 5	WP 1	B	D13	5	D17	C3
Task 8	WP 2	B	D21	5	D25	C3
Task 9	WP 2	B	D21	5	D25	C5
Task 10	WP 2	B	D21	8	D28	C1
Task 13	WP 3	B	D40	6	D45	C1
Task 14	WP 3	B	D40	7	D46	C4
Task 16	WP 4	B	D46	4	D49	C5
Task 17	WP 4	B	D46	5	D50	C2

(1) Sort by site

Task	Site ↑	End	Contr.
Task 1	A	D03	C1
Task 2	A	D09	C2
Task 3	A	D12	C3
Task 6	A	D20	C2
Task 7	A	D20	C4
Task 11	A	D15	C4
Task 12	A	D39	C1
Task 15	A	D25	C4
Task 18	A	D52	C3
Task 19	A	D53	C1
Task 20	A	D56	C3

Task	Site ↑	End	Contr.
Task 4	B	D15	C1
Task 5	B	D17	C3
Task 8	B	D25	C3
Task 9	B	D25	C5
Task 10	B	D28	C1
Task 13	B	D45	C1
Task 14	B	D46	C4
Task 16	B	D49	C5
Task 17	B	D50	C2

(2) Segment different sites

Task	Site	End ↑	Contr.
Task 1	A	D03	C1
Task 2	A	D09	C2
Task 3	A	D12	C3
Task 11	A	D15	C4
Task 6	A	D20	C2
Task 7	A	D20	C4
Task 15	A	D25	C4
Task 12	A	D39	C1
Task 18	A	D52	C3
Task 19	A	D53	C1
Task 20	A	D56	C3

Task	Site	End ↑	Contr.
Task 4	B	D15	C1
Task 5	B	D17	C3
Task 8	B	D25	C3
Task 9	B	D25	C5
Task 10	B	D28	C1
Task 13	B	D45	C1
Task 14	B	D46	C4
Task 16	B	D49	C5
Task 17	B	D50	C2

(3) Sort by time-window

Task	Site	End ↑	Contr.
Task 1	A	D03	C1
Task 2	A	D09	C2
Task 3	A	D12	C3
Task 11	A	D15	C4
Task 6	A	D20	C2
Task 7	A	D20	C4
Task 15	A	D25	C4

Task	Site	End ↑	Contr.
Task 12	A	D39	C1
Task 18	A	D52	C3
Task 19	A	D53	C1
Task 20	A	D56	C3

Task	Site	End ↑	Contr.
Task 4	B	D15	C1
Task 5	B	D17	C3
Task 8	B	D25	C3
Task 9	B	D25	C5
Task 10	B	D28	C1

Task	Site	End ↑	Contr.
Task 13	B	D45	C1
Task 14	B	D46	C4
Task 16	B	D49	C5
Task 17	B	D50	C2

(4) Segment different time-windows

Task	Site	End	Contr. ↑
Task 1	A	D03	C1
Task 2	A	D09	C2
Task 6	A	D20	C2
Task 3	A	D12	C3
Task 11	A	D15	C4
Task 7	A	D20	C4
Task 15	A	D25	C4

Task	Site	End	Contr. ↑
Task 12	A	D39	C1
Task 19	A	D53	C1
Task 18	A	D52	C3
Task 20	A	D56	C3

Task	Site	End	Contr. ↑
Task 4	B	D15	C1
Task 10	B	D28	C1
Task 5	B	D17	C3
Task 8	B	D25	C3
Task 9	B	D25	C5

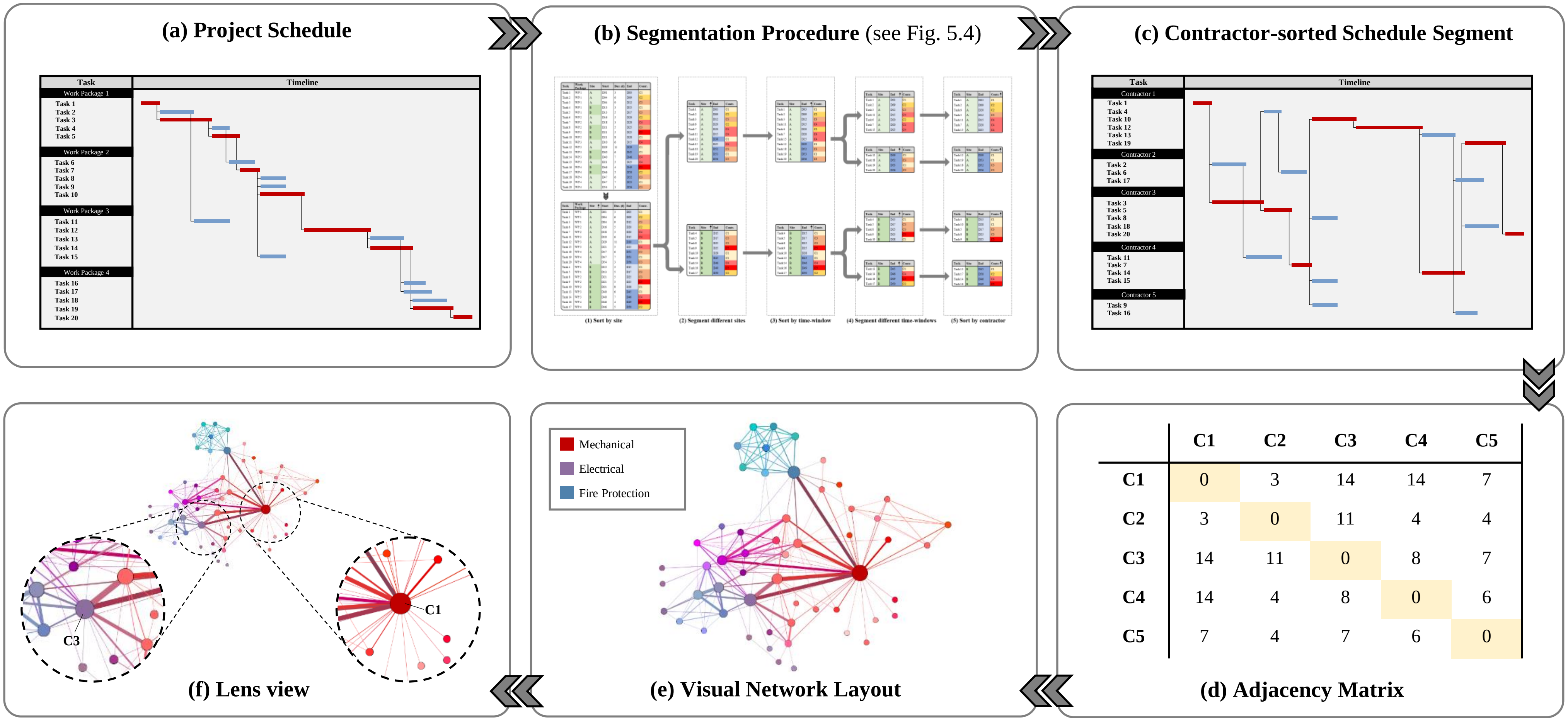
Task	Site	End	Contr. ↑
Task 13	B	D45	C1
Task 17	B	D50	C2
Task 14	B	D46	C4
Task 16	B	D49	C5

(5) Sort by contractor

Figure 5.4: Project schedule segmentation procedure

5.4.2 Tool (1): Interdependence Quantification and Snapshot Network Modelling

As a first step in the framework, different contractor snapshot network layouts are generated for each site location and every time-window (i.e., for each segment), as shown in Figure 5.5. For demonstration, it is assumed that the contractor-sorted schedule in Figure 5.5c pertains to the same site and month (i.e., is one segment). Within each segment, the magnitude of interdependence between any pair of contractors is numerically quantified through the time in days (longevity) of these contractors working simultaneously with interrelated tasks within the same site location. These numerical quantifications are then used to develop an adjacency matrix (Newman et al. 2006), as shown in Figure 5.5d, which describes the magnitude of task interdependence (i.e., the weight of links) between each contractor (i.e., node) pairs in the network. Such an adjacency matrix can be derived from the schedule segment in Figure 5.5c (or in tabular format) using standard numerical computing or spreadsheet tools based on the start and end dates of contractor-specific tasks. For example, consider the first row of the matrix in Figure 5.5d. Contractor C1 worked together interdependently with C2 for 3 days, with C3 for 14 days, with C4 for 14 days, and with C5 for 7 days, and so on for other contractor pairs. The matrix is also symmetrical with zeroes in the diagonal since contractors working with themselves is meaningless.



	C1	C2	C3	C4	C5
C1	0	3	14	14	7
C2	3	0	11	4	4
C3	14	11	0	8	7
C4	14	4	8	0	6
C5	7	4	7	6	0

Figure 5.5: Tool (1): Interdependence quantification and snapshot network modelling

This adjacency matrix can be further transformed into a visual monthly contractor network layout, as shown in Figure 5.5e. Different software packages are available for network modelling, visualization, and subsequent analysis (e.g., Ucinet, Pajek, Net Draw, and Gephi). Within these tools, Gephi was used in the current study as it is more user friendly, is an open-source software program, possess powerful visualization capabilities, supports the customization of network layout features of color and size for elegant network representation, and provides metrics including those introduced earlier, (e.g., WDC, BC, ND, etc.) for further analyses (Apostolato 2013; Wehbe et al. 2016; Faysal and Arifuzzaman 2018). By importing the adjacency matrix in .xlsx format into Gephi version 0.9.2, a corresponding network layout is generated such as the one showed in Figure 5.5e. This layout contains 55 nodes and is only presented for demonstration purposes as it would typically be the result of a 55×55 adjacency matrix (compared to that in Figure 5.5d) and a much larger schedule (compared to that in Figure 5.5c). For context, one node in the network layout would essentially represent one of the five contractors (C1 to C5) in the adjacency matrix, as shown in the magnified lens view in Figure 5.5f. Different visual network layout features, as node size and color, can be customized to represent different measures to maximize the utility of the network layout. For example, in Fig 5e, node sizes represent each contractor's scope of work that is measured as the number of working days within the considered monthly segment. Specifically, contractors with larger scopes are expressed by larger node sizes and are placed in the more central areas of the network layout, whereas contractors with smaller scopes are expressed by smaller node sizes and are placed in the more peripheral areas of the network layout. Node colors represent the different contractor disciplines/professions, where for demonstration, the legend in Figure 5.5e comprises of three disciplines, namely mechanical, electrical, and fire protection. In the network layout, contractors of the same discipline are marked with the same color as per the legend and are clustered together. In addition to this, node color shades represent varying contractor influences toward the project work. From the multiple types of centrality measures introduced earlier, WDC is selected for further demonstration hereafter. The WDC of node *i* can be calculated from the adjacency matrix as (Park et al. 2010; Xue et al. 2018; Ezzeldin and El-Dakhakhni 2019):

$$WDC_i = \sum_{j=1}^{n} w_{ij} \tag{5.1}$$

where w_{ij} is the weight/magnitude of interdependence between contractor *i* and all other contractors, and *n* is the number of contractors (nodes) in the network. Contractors with higher WDC values are expressed by darker node color shades, and vice versa. Those nodes with high WDC values indicate contractors with tasks that are highly correlated and interdependent with the tasks of other contractors. As such, the contractors with high WDC present *bottlenecks* on the site, thus possessing high potentials of triggering or amplifying CPD if they underperform their tasks. Beyond the scope of work and WDC used above for demonstration purposes, nodes sizes, colors, and color shades can be customized interchangeably to portray rankings/scales of other measures, such as BC, CC, and EC, as necessary for different application purposes.

5.4.3 Tool (2): Dynamic Network Modelling

For every spatial and temporal change in the project, there transpires a different arrangement of tasks assigned to either the same or a different group of collaborating contractors and thus a different adjacency matrix and corresponding (snapshot) network layout. Therefore, as shown in Figure 5.6, dynamic networks are generated from the snapshot network layouts developed from Tool (1), with one network for each time-window per site. The monthly time-windows selected for the current demonstration are represented as month T_1 to T_n hereafter. As can be inferred from the demonstration example shown in Figure 5.6, the group of contractors working during the first month, T_1, are not the same group working during the final month, T_n. As such, for a project with a specific site location and work scopes that are scheduled for three years, for example, a dynamic network of 36 monthly layouts can be generated (i.e., $T_n = T_{36}$). If quarterly (every three months) time-windows are selected instead, a corresponding dynamic network of 12 quarter layouts can be generated (i.e., $T_n = T_{12}$). These dynamic networks represent the different groups of contractors and their varying interdependencies across this site location and over the project lifecycle.

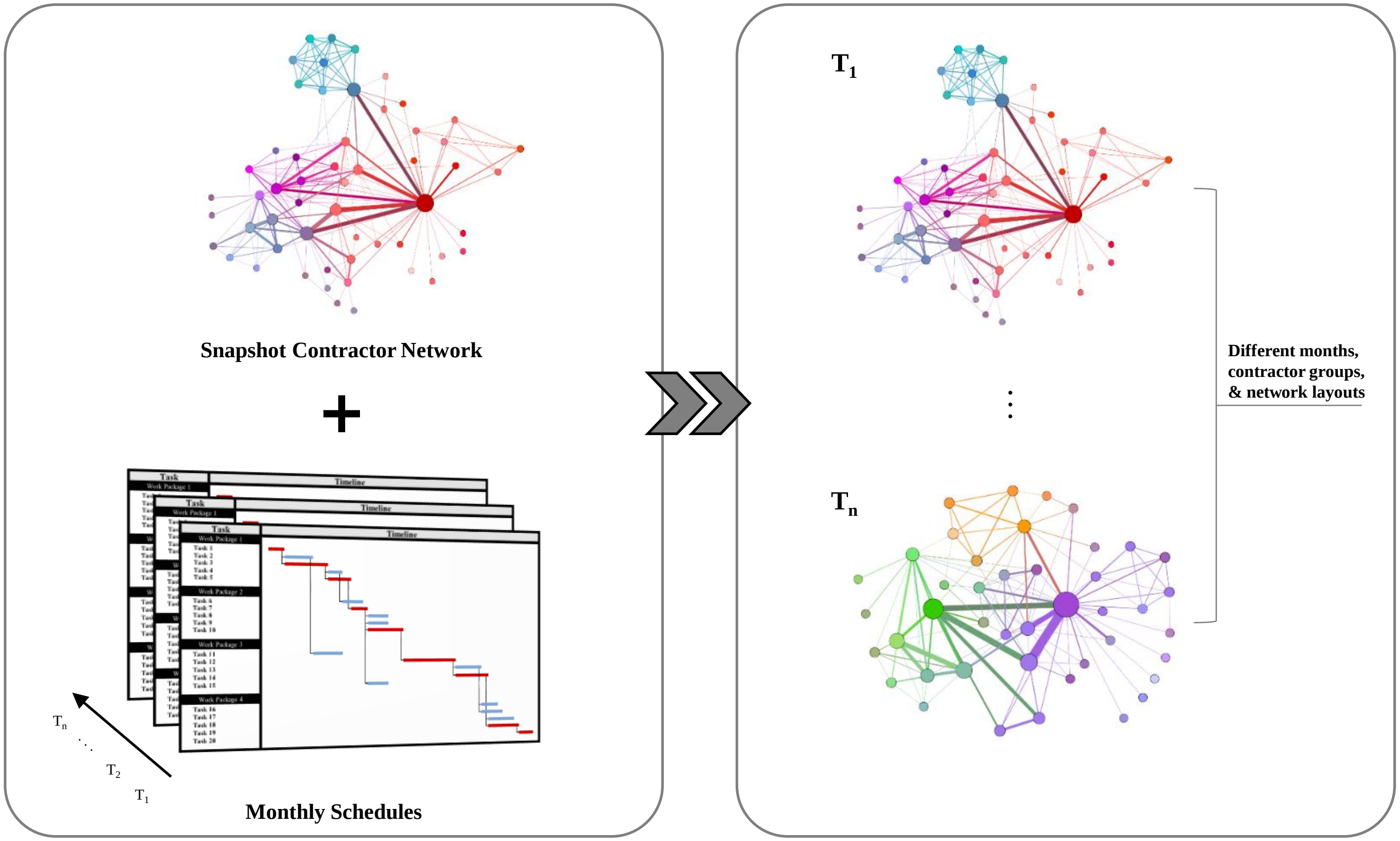
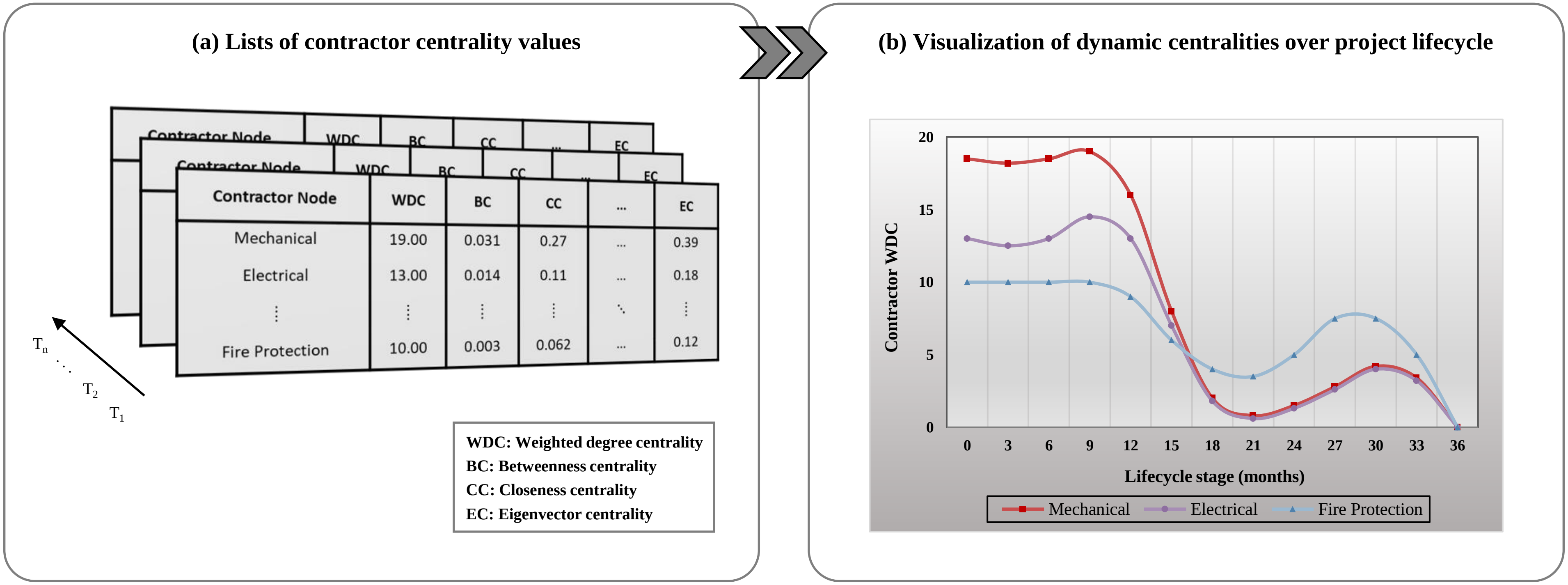


Figure 5.6: Tool (2): Dynamic network modelling

5.4.4 Tool (3): Dynamic Network Centrality Analyses

Dynamic centrality analyses identify the most influential contractors on each site location and during different months of the project, which subsequently facilitates the visualization of how these influences vary with time over the entire project lifecycle. For any network layout, Gephi facilitates exporting a list of the computed node centrality values in .xlsx format (e.g., WDC, BC, CC, and EC), which can be carried out for every monthly snapshot. For example, Figure 5.7a demonstrates a list of centralities for three contractors in the network (one from each discipline listed earlier) during every month of the project. This in turn enables aggregating these values into one master list and visualizing these contractors' varying influences throughout the entire project lifecycle, as shown in Figure 5.7b. For instance, it can be observed from the figure that, during the first month of the project, the mechanical contractor is the most influential on the site, followed by the electrical and fire protection contractors, respectively. The same figure also reveals that these three contractors possess higher WDC values, and are thus more influential, during the initial stages of the project compared to the later stages. The greater the influence of a contractor during a specific month, the greater their potential to cause disruptions to other contractors' performances due to their highly interdependent tasks with such contractors.



(a) Lists of contractor centrality values

Contractor Node	WDC	BC	CC	...	EC
Mechanical	19.00	0.031	0.27	...	0.39
Electrical	13.00	0.014	0.11	...	0.18
⋮	⋮	⋮	⋮	⋱	⋮
Fire Protection	10.00	0.003	0.062	...	0.12

WDC: Weighted degree centrality
BC: Betweenness centrality
CC: Closeness centrality
EC: Eigenvector centrality

(b) Visualization of dynamic centralities over project lifecycle

Figure 5.7: Tool (3): Dynamic network centrality analyses

5.4.5 Tool (4): Network Measure-Project KPI Correlation Analyses

The proposed resilience-driven project management approach focuses on ensuring rapid recovery from and adaption to multiple CPD. As such, different metrics against which to measure network resilience are needed. In this respect, the fourth tool involves tracking the project's performance at the end of every month (i.e., for every network snapshot) and evaluating the different corresponding project key performance indicators (**KPIs**). The current study considers six KPIs (Leon et al. 2017; Castillo et al. 2018) which are: 1) Schedule deviation (**SD**); 2) Cost deviation (**CD**); 3) Quality index (**QI**); 4) Accident frequency (**AF**); 5) Productivity rate (**PR**); and 6) Planning effectiveness (**PE**). Particularly, lower values of SD, CD, QI, and AF, and higher values of PR and PE are desirable and indicate satisfactory project performance.

- $\text{Schedule deviation (SD)} = (\text{Scheduled advance} - \text{Actual advance})/\text{Scheduled advance}$

- $\text{Cost deviation (CD)} = (\text{Actual cost} - \text{Budgeted cost})/\text{Budgeted cost}$

- $\text{Quality index (QI)} = \text{Number of rework orders}/\text{Work hours}$

- $\text{Accident frequency (AF)} = \text{Number of recordable accidents}/\text{Work hours}$

- $\text{Productivity rate (PR)} = \text{Budgeted labor cost}/\text{Actual labor cost}$

- $\text{Planning effectiveness (PE)} = \text{Completed tasks}/\text{Scheduled tasks}$

These tracked KPIs are then correlated with their respective network layouts to assess the effectiveness of scheduling specific tasks to different groups of contractors working within the same space and during the same time. This in turn is performed by plotting the six KPI values, for every network snapshot, against the WDC values for the different contractors, as presented in Figure 5.8. These plots can be developed using standard computing tools, as the R programming language and RStudio IDE (R Core Team 2013) which are used herein. For example, Figure 5.8 can be reproduced using the *pairs.panels* function within the *psych* package (Revelle 2020) or using the *ggpairs* function from the *GGally* package (Schloerke et al. 2021). From these plots, correlation scores (ranging from +1 to -1) are provided to describe the strength and direction of the relationship of the plotted values. These correlation scores uncover which task arrangements, high-influence contractors and contractor collaborations contribute to the satisfactory or poor project performance based on different indicators. For instance, it can be observed from Figure 5.8 that the greater the mechanical contractor's influence on the project site (i.e., the more involved and the greater control this contractor has over the monthly tasks), the better the overall project performance. This observation is exemplified through this contractor's WDC values' strong relationships with decreasing values of SD, CD, QI and AF, and increasing values of PR and PE.

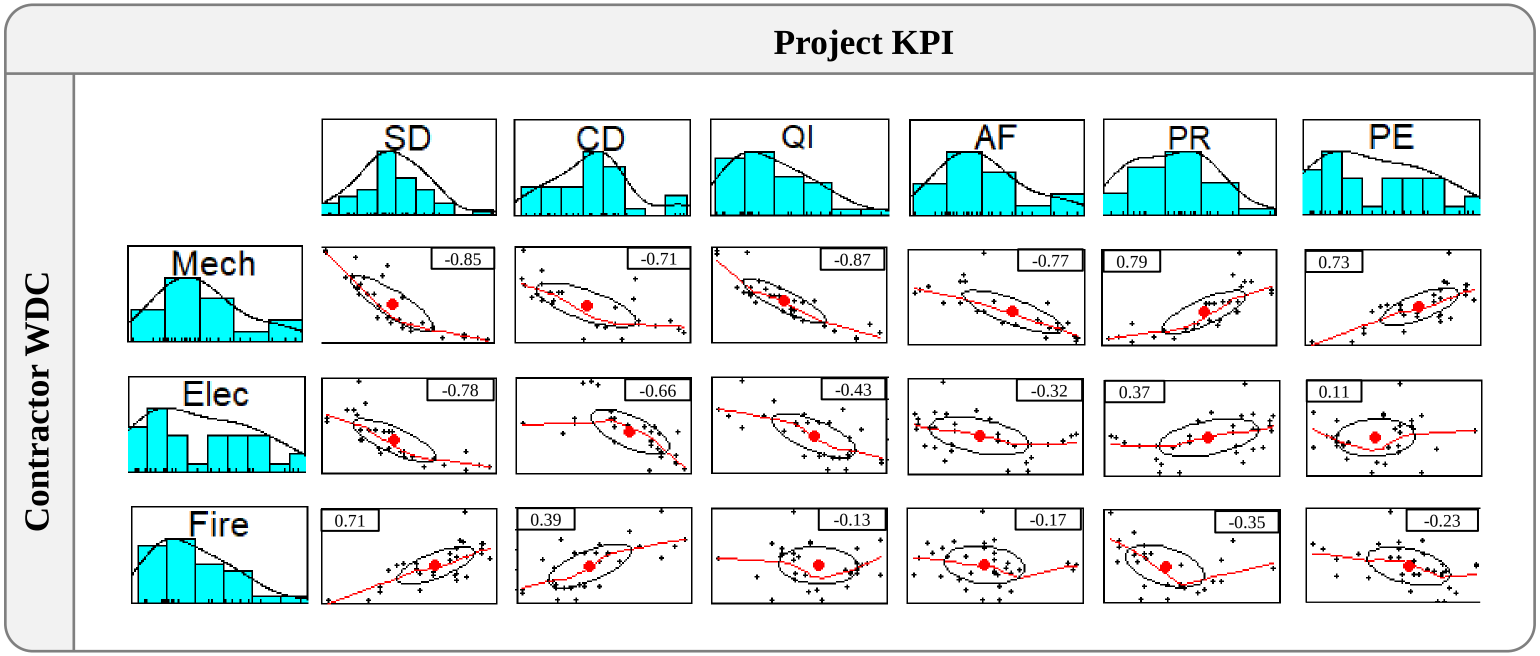
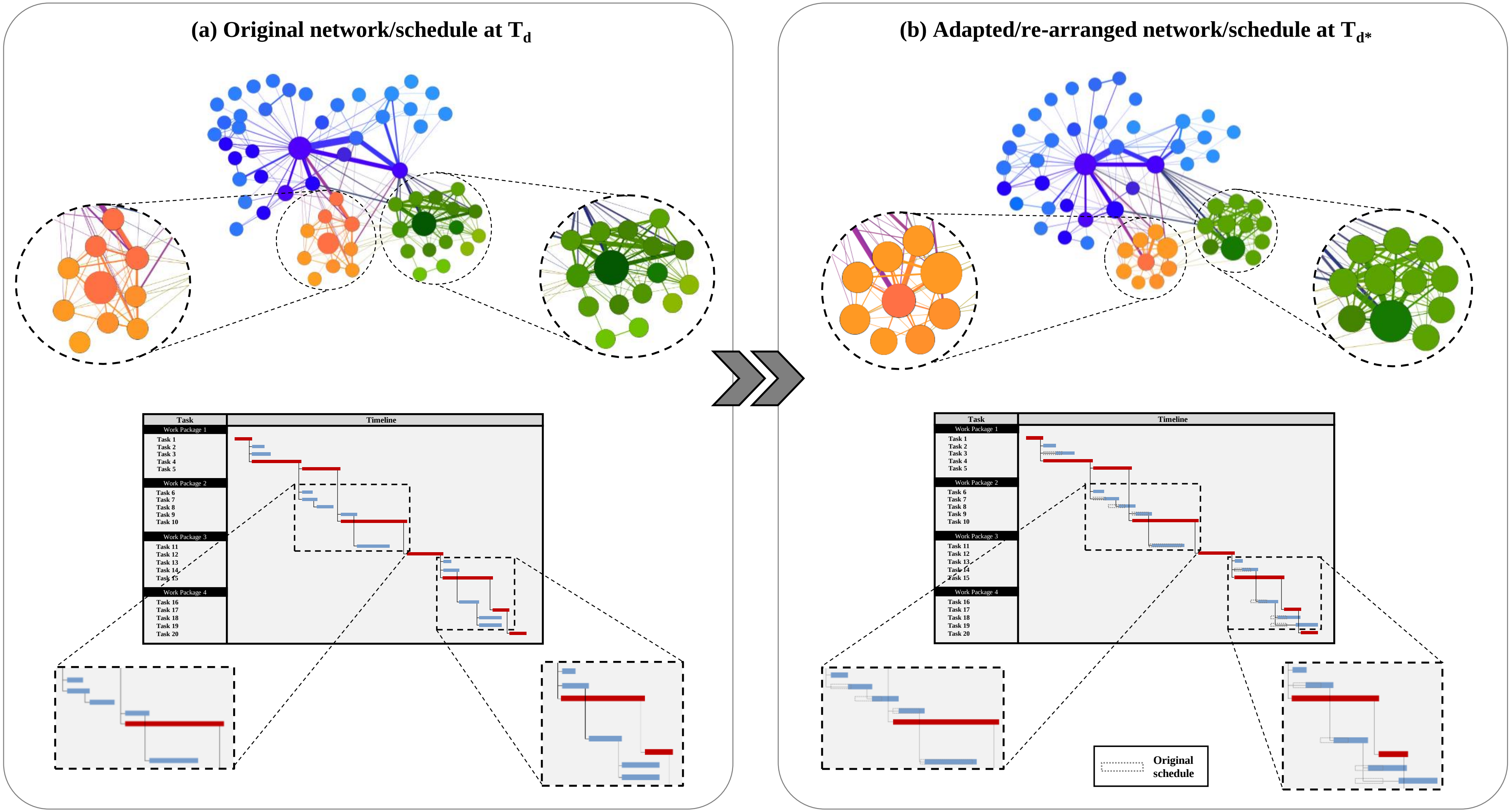


Figure 5.8: Tool (4): Network measure-project KPI correlation analyses

5.4.6 Tool (5): Adaptive (self-organized) Network Development

The fifth and final tool is designed as a twofold process. The ***first*** process involves proactively (at early project stages) foreseeing when (e.g., in which month, T_d) CPD may occur and which KPI(s) may be affected. This prediction can be reached through utilizing each month's contractors' WDC values [obtained from Tool (3)] to analytically extrapolate each month's expected KPI values from these contractors' corresponding correlation plots [generated by Tool (4)]. From the extrapolated KPI values (e.g., SD, CD, QI, etc.), a disrupted month, T_d, can be detected and may also be expected to experience unsatisfactory performance in one or more KPI(s), thus signaling the type of managerial intervention needed in response to such a disruptive event(s).

The ***second*** process involves analytically evaluating all T_d's possible decision alternatives and returning the optimal response plan (see Figure 5.9). Specifically, metaheuristic schedule optimization techniques may be employed to: 1) understand the internal structure of T_d's original schedule to define the different tasks' time intervals and float durations, contractors' and resources' inactive slack durations, and underlying sets of logical-, resource-, technological- and organizational constraints; 2) iteratively search all possible solutions of task and contractor rearrangements bound by the above-learned durations and constraints; 3) convert each solution into a corresponding network layout and apply WDC analysis to each network; 4) assess each solution's extrapolated KPIs; and



(a) Original network/schedule at T_d

(b) Adapted/re-arranged network/schedule at T_{d*}

Figure 5.9: Tool (5): Adaptive (self-organized) network development

5) ultimately return the optimal solution (self-organized network) that renders the most favorable KPI values. The metaheuristic search will thus equip project managers with an objective and evidence-supported means for rapidly detecting the optimal response plan to face possible CPD. The response plan yields an adapted (self-organized) contractor network layout, T_{d*}, which can be used as a strategic countermeasure to replace the disrupted T_d network to rapidly restore a desired infrastructure project performance state. For example, the network layout pertaining to T_{d*} in Figure 5.9b can have different (adapted) node layouts compared to that of the layout of T_d in Figure 5.9a. This adapted layout reflects the re-arranged tasks in T_{d*} schedule in Fig 5.9b. By moving tasks within the boundaries of their floats and constraints, critical vulnerable work interdependencies between pairs of contractors can be avoided – interdependencies known to induce CPD in the past. It is important to observe how WDC values for contractors would decrease in the network layout in Figure 5.9b compared to that in Figure 5.9a, reflected by lighter color scales. As such, by decreasing the WDC of specific nodes in the network layout, interdependence-induced vulnerabilities can be avoided between contractors decreasing the chance of possible CPD. Finally, as shown in Figure 5.9b, this optimized/adapted network layout can be further reinstated into a new arrangement of (re-coordinated) scheduled tasks, which may facilitate contractor re-coordination insights to project managers in preparedness against the threats of any further CPD extending to subsequent months. This can be observed through the re-

arrangement of tasks in Figure 5.9b—providing project managers with new proposed schedules.

5.5 Framework Demonstration Application: a Hydroelectric Power Generation Infrastructure Overhaul Project

The developed framework was introduced to one of the largest power generation corporations in North America, that owns and operates over a hundred power generation stations of nuclear, hydroelectric, wind, gas, and biomass sources. Every year, the corporation generates hundreds of TWh of power and billions of dollars in revenue and contributes billions of dollars into the continent's GDP. To demonstrate its utility, the framework was applied in an industry setting to a large-scale overhaul project on one of the corporation's hydroelectric power generation stations.

5.5.1 Project Description

The studied station houses six identical hydroelectric power generation units which are part of a synchronized outage cycle designed to ensure continued safe and reliable long-term operations of the station and compliance with regulatory requirements. This outage cycle is synchronized so that five units always remain operational and only one undergoes a planned outage and thus overhaul, with the process repeated sequentially. A typical unit overhaul was studied herein and included refurbishment and rehabilitation work packages that aimed at restoring the unit's original performance, extending its lifespan, preventing its components' breakdowns, and reducing unforeseen forced outages and associated costs. Each unit overhaul project spans across a 13-month lifecycle period and consists of four main work packages as scheduled in Figure 5.10: ***components disassembly***, ***pre-assembly works***, ***components off-site refurbishment*** and ***components assembly and commissioning***. The project's main working location is the station site, where the units and their components are situated; however, work also spreads across several other locations remote to this site. A total of 14 contractors collaborated within the project and their anonymized IDs, contract types, technical disciplines, and working locations are presented in Table 5.1. Specifically, nine of these contractors (M1 to M4, E1 to E3 and X1 to X2) were in-house contractors, whereas five contractors (C1 to C5) were procured as external vendors and specialty contractors. As shown in the table, contractors M1 to M4, E1 to E3 and X1 to X2 denote mechanical, electrical and machining contractors, respectively, where contractors within the same discipline offer different skill sets, capabilities and roles, as necessitated by project requirements. These contractors were all positioned on the station site except for X2 who worked from an off-site location (machining yard). Also visible from the table, contractors C1 to C4, specialized in the refurbishment of power system components, were situated at their own remote locations for the

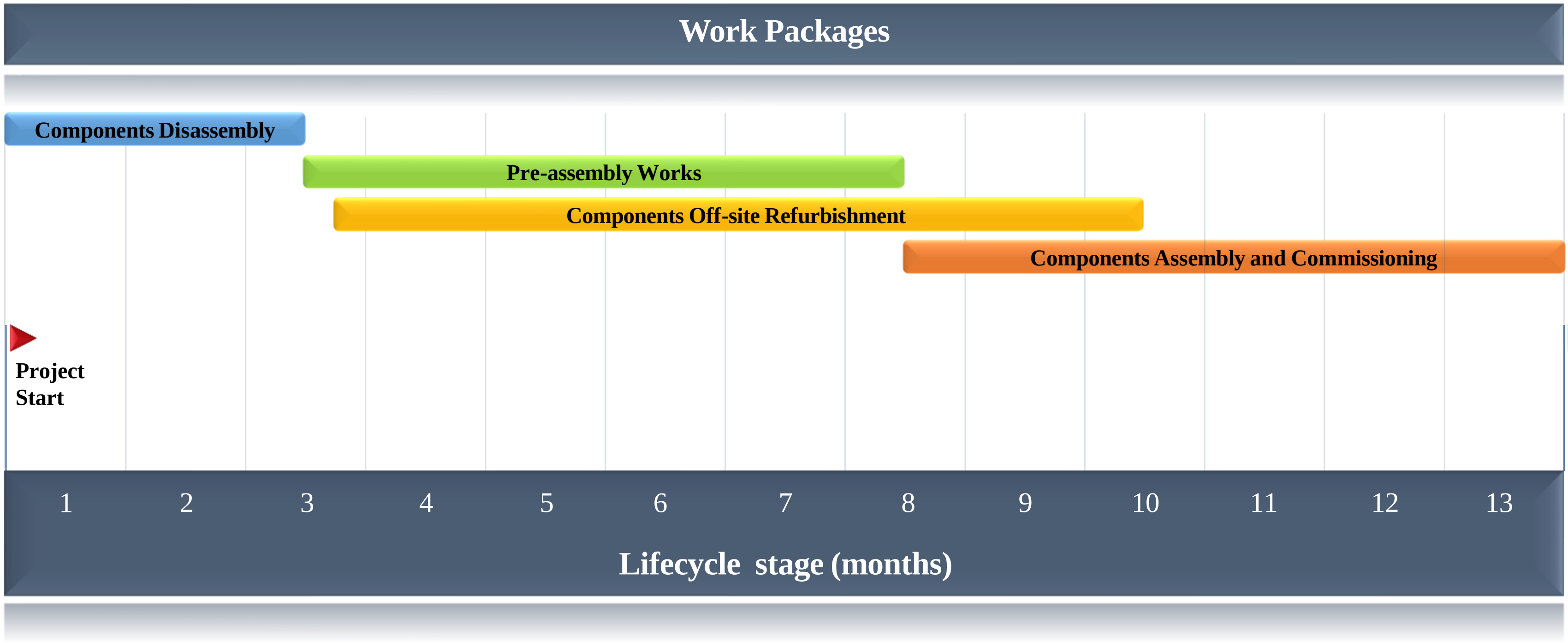


Figure 5.10: Hydroelectric power generation station typical unit overhaul project schedule

off-site refurbishment work package and the subsequent shipping of components back to the station site. In this respect, the external crane handling specialty contractor, C5, was positioned on the station site to employ the knowledge sets of these external contractors in the assembly and installation of the received components.

Table 5.1: Contractors collaborating within the hydroelectric power generation station overhaul project

Contractor ID	**Type**	**Discipline**	**Location**
M1	In-house contactor	Mechanical	On station site
M2			
M3			
M4			
E1	In-house contractor	Electrical	On station site
E2			
E3			
X1	In-house contractor	Machining	On station site
X2			Remote to station site
C1	External vendor	Power system components refurbishment	Remote to station site
C2			
C3			
C4			
C5	External specialty contractor	Crane operator	On station site

5.5.2 Project Strategic Importance

Notwithstanding the importance of the described overhaul projects in maintaining the station's long-term operations safety and performance, it is important to note that variations to the frequencies, timings and durations of planned outages under the station's outage cycle result in temporal variability in the company's financial results, including an impact on revenue and operations, maintenance and administrative expenses. In this context, a single overhaul is not only faced with *intra*-project interdependence but also *inter*-project interdependence. As such, the importance of a well-managed overhaul project that successfully coordinates between the complex work packages and ensures effective collaborations of the contractors to achieve the desired performance targets (e.g., duration and budget) is critical. Because of this criticality, and because of the repetitive/identical nature of the overhaul projects across the units which rendered the managerial insights gained from one project directly transferable to the others, this type of project was considered suitable to demonstrate the application and utility of the developed framework.

5.5.3 Project Network Analysis

Available project data included project schedules and contractors' assignments to tasks, whereas intermittently tracked KPIs were unavailable due to data restrictions imposed by the corporation. As such, the current demonstration focuses on applying Tools (1) to (3) of the framework, while applying the remaining tools was beyond the scope of the current demonstration. Snapshot network modelling was primarily applied whilst considering the entire project lifecycle of 13 months as a single segment, as shown in Figure 5.11. This network contains all 14 participating contractors and presents an all-encompassing overview of their collaborations and relationships formed throughout the entire project lifecycle. However, given the 13-month duration of the project lifecycle, monthly segments were found to be suitable for implementing the subsequent dynamic (monthly) network modelling, as shown in Figures 5.12 (Months 1 to 7) and 5.13 (Months 8 to 13). Since links in the current study represent interdependent tasks bound by both spatial and temporal settings, as defined earlier, all networks can be divided into two visibly distinct areas (e.g., see Figure 5.11), namely: 1) the ***core of the network***, which exhibits only the contractors working on the station site and their interdependencies (links); and 2) the ***periphery of the network***, where the remote contractors' nodes are placed with no links connected. When visualizing the networks, the node sizes, link weights and node colors were each set to portray different project insights to maximize the practical use of each network plot. Specifically, node sizes represent each contractor's scope of work measured as the number of working days within the project or the considered month. Link weights represent the magnitude of direct task interdependence that exists locally between two contractors. Node colors represent the magnitude of overall task interdependencies that exist globally between any one contractor and all others sharing links.

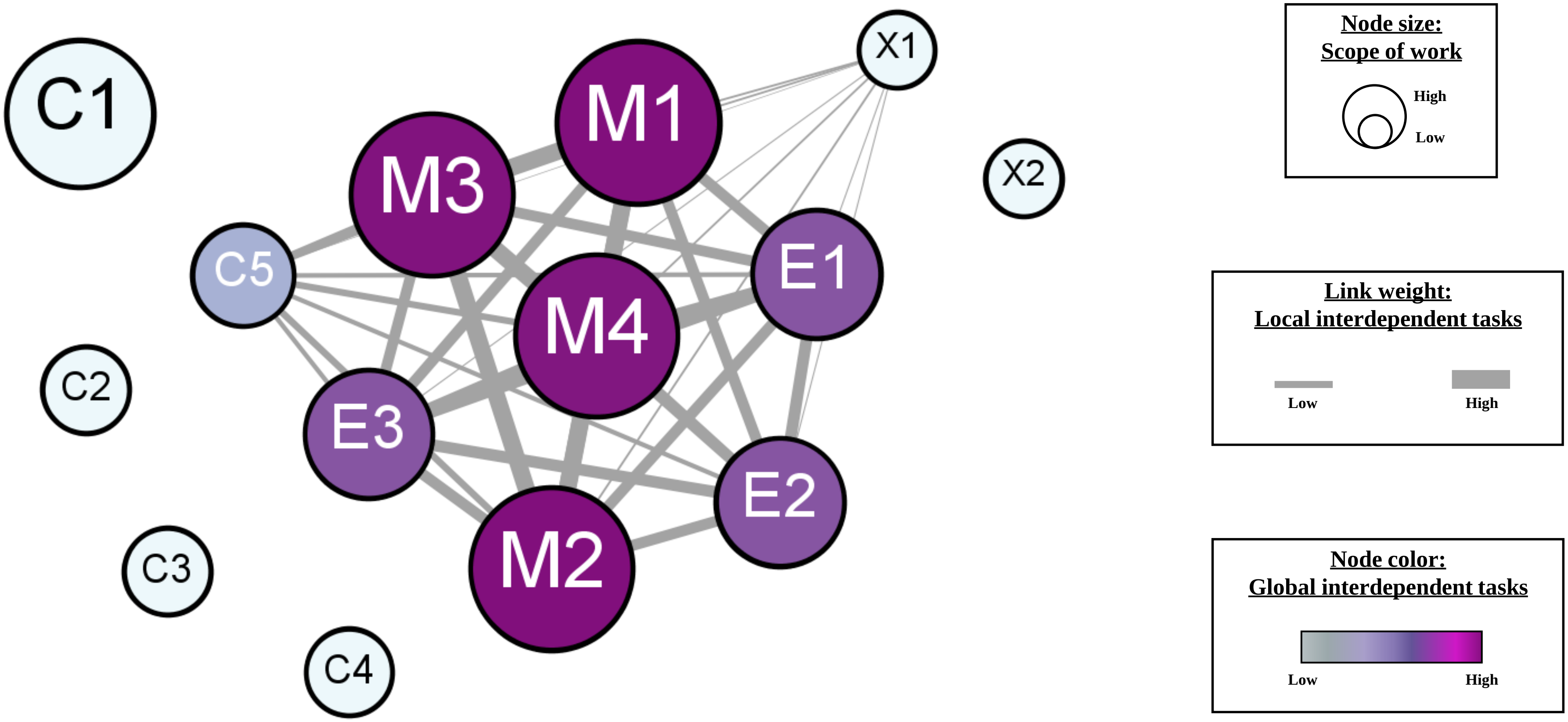


Figure 5.11: Snapshot network modelling representing contractors' scope of work and interdependent tasks for entire project lifecycle (13 months)

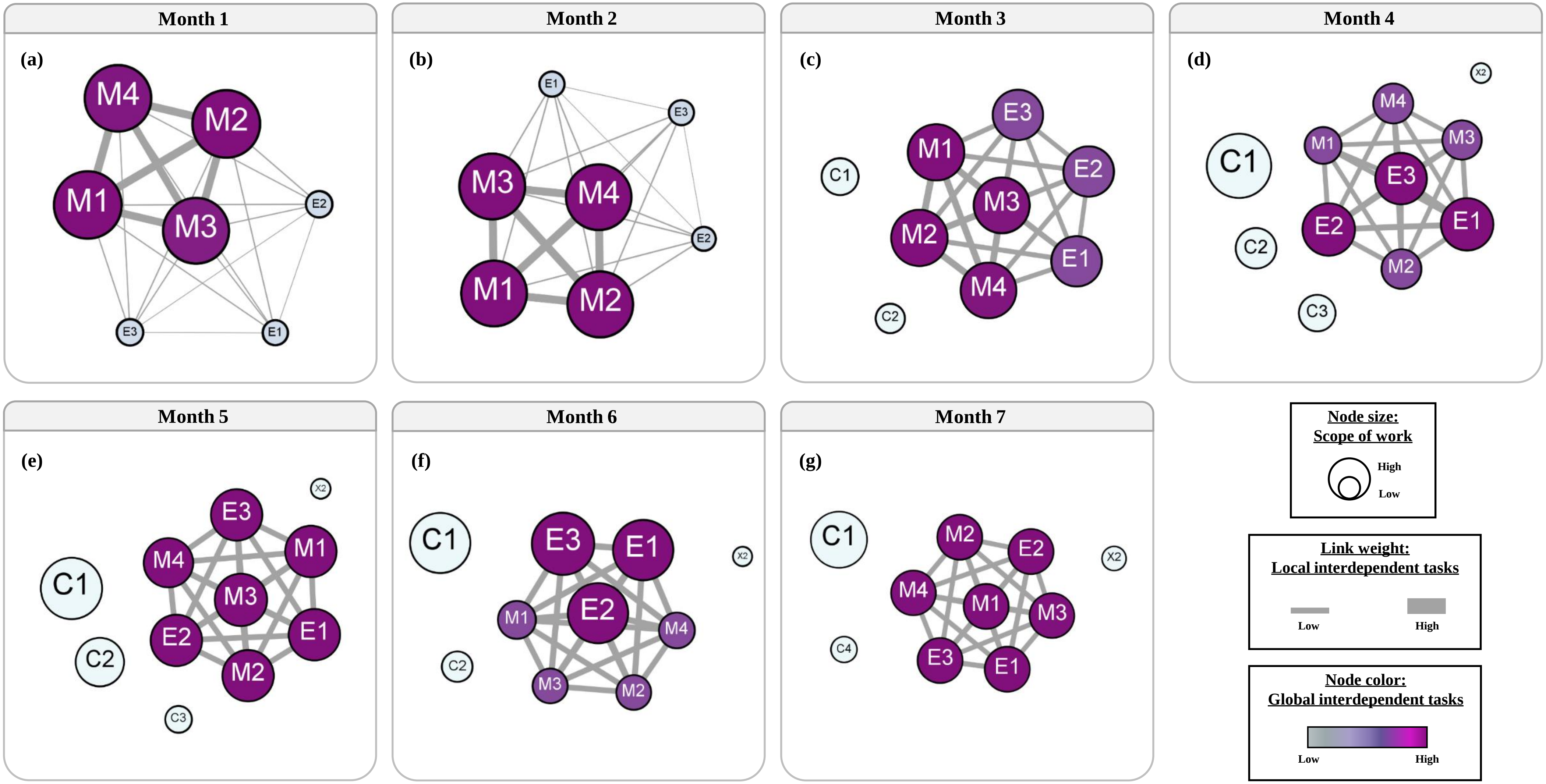


Figure 5.12: Dynamic (monthly) network modelling representing contractors' scope of work and interdependent tasks over project lifecycle stages months 1-7

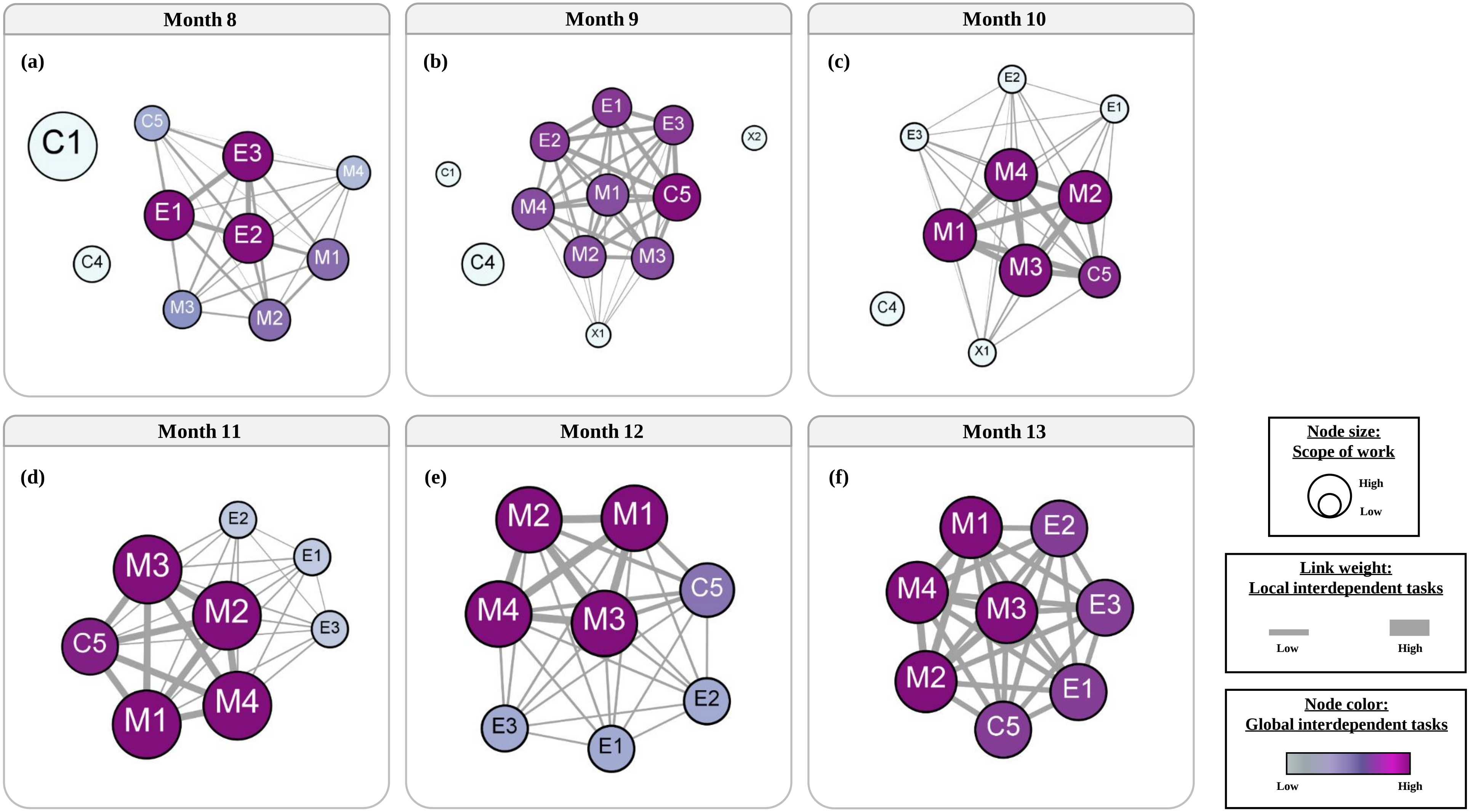


Figure 5.13: Dynamic (monthly) network modelling representing contractors' scope of work and interdependent tasks over project lifecycle stages months 8-13

While network modelling enables the visual inspection of the overall network layouts and provides initial findings of existing interdependencies/interactions, network measure analyses offer a deeper interpretation. As summarized in Table 5.2, network-level measures, including the number of nodes and links, network density (**ND**) and average weighted degree centrality (**AWDC**) were computed to reflect the changes in the network layouts over the project lifecycle monthly stages. ND and AWDC can be computed as per Equations (5.2) and (5.3), respectively (Park et al. 2010; Estrada and Knight 2015; Xue et al. 2018).

$$ND = l/(n \times (n-1)/2) \tag{5.2}$$

$$AWDC = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}{n} \tag{5.3}$$

where *l* is the number of network links, *n* is the number of network nodes, and w_{ij} is the weight/magnitude of interdependence between contractor i and j. Furthermore, two sets of dynamic WDC analyses were implemented. The first simulates the scope of work over project life cycle stages and is shown through Figures 5.14 and 5.15. Figure 5.14 is exclusive to the contractors working on the station site location and Figure 5.15 is specific to those working remotely. The second dynamic WDC analysis simulates interdependent tasks over lifecycle stages and is shown in Figure 5.16 to include only contractors working on-site.

Table 5.2: Network-level measures over lifecycle stages of the hydroelectric power generation station overhaul project

	Lifecycle stage (months)													**Entire lifecycle**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	
No. nodes representing on-site contractors	7	7	7	7	7	7	7	8	9	9	8	8	8	**9**
No. nodes representing remote contractors	0	0	2	4	4	3	3	2	3	1	0	0	0	**5**
No. links	21	21	21	21	21	21	21	28	36	36	28	28	28	**36**
Network Density (ND)	1	1	1	1	1	1	1	1	1	1	1	1	1	**1**
Average weighted degree centrality	159.40	172.84	351.51	353.69	405.70	420.09	411.48	211.35	396.13	298.79	283.02	277.73	468.41	**371.69**

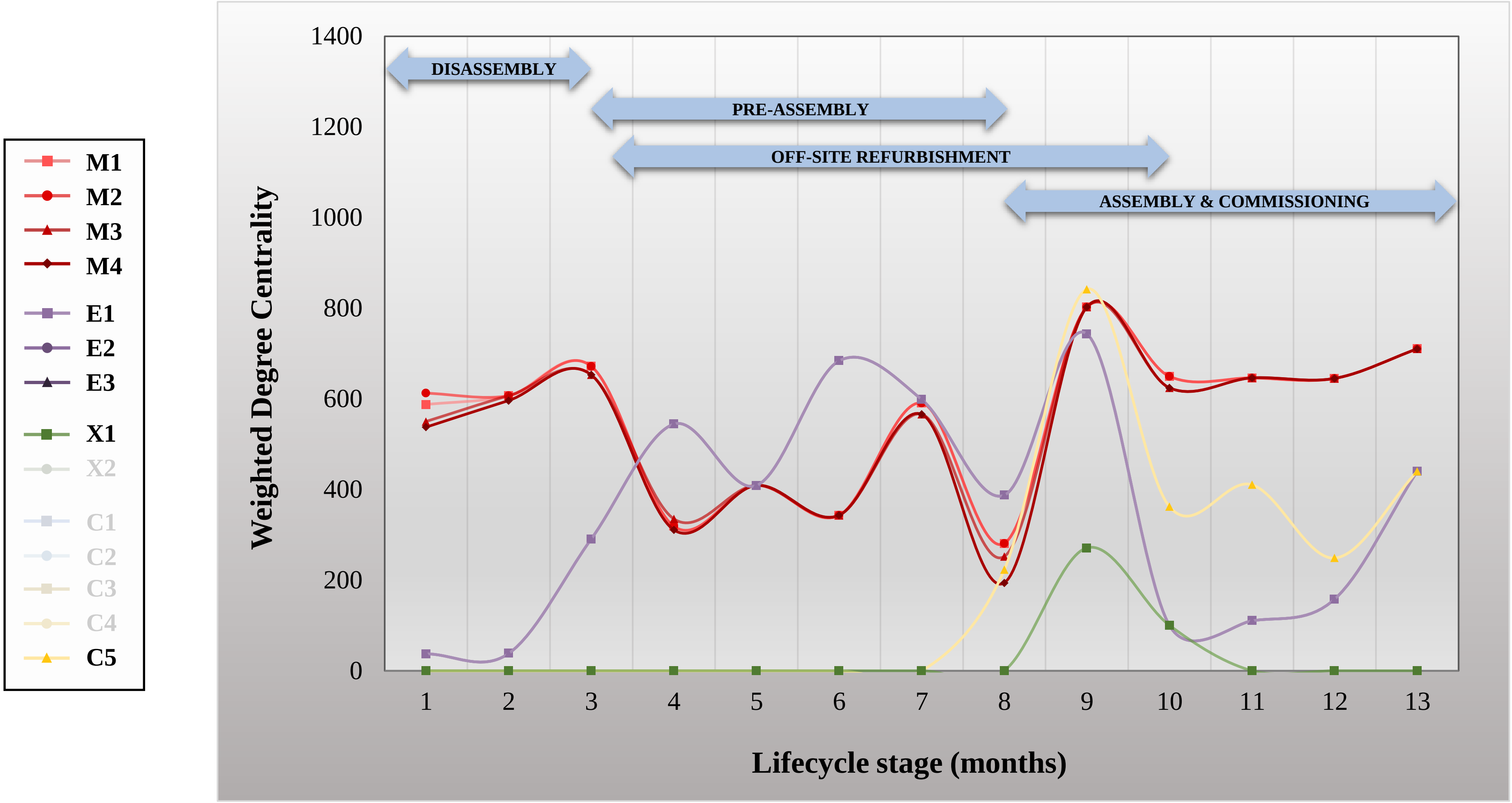


Figure 5.14: Dynamic WDC analysis simulating scope of work over project lifecycle stages for contractors working on project site location

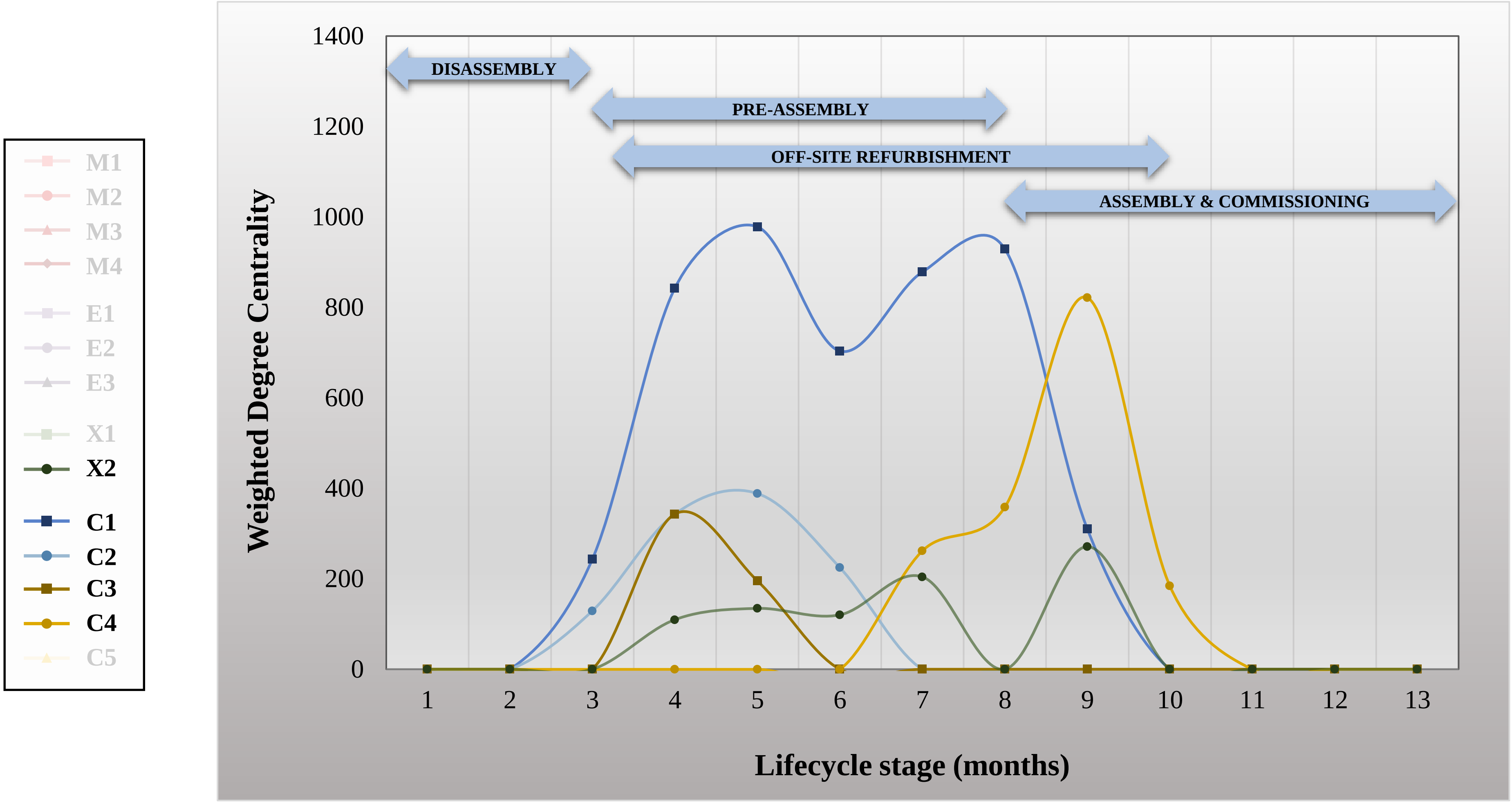


Figure 5.15: Dynamic WDC analysis simulating scope of work over project lifecycle stages for contractors working remote to project site location

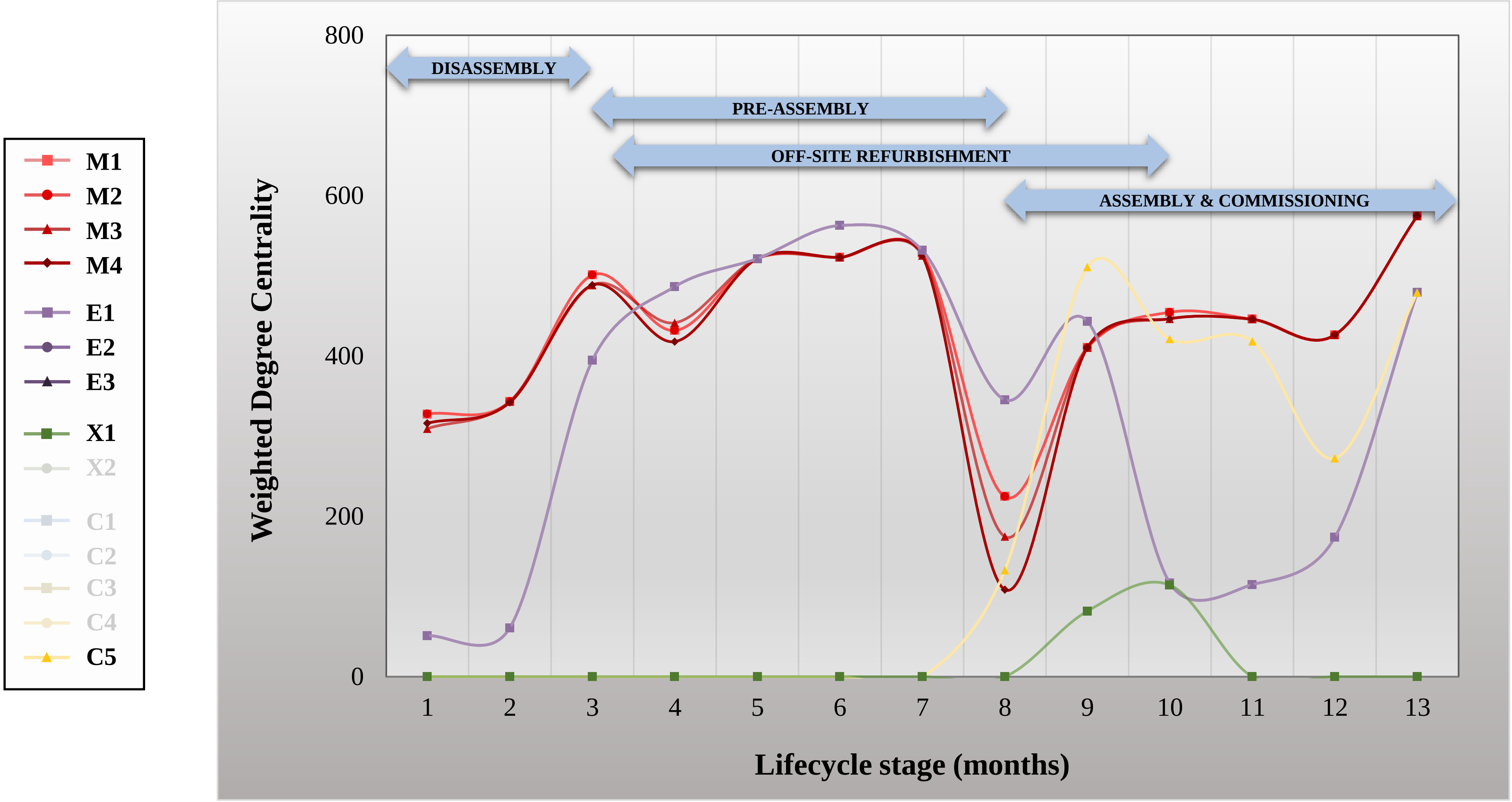


Figure 5.16: Dynamic WDC analysis simulating interdependent tasks over project lifecycle stages for contractors working on project site location

In Figures 5.14 to 5.16, the WDCs, provided for each month and every contractor, are connected using smoothed lines to only facilitate visual interpretation of the WDCs dynamics over time rather than to imply continuity per se.

5.6 DISCUSSION OF ANALYSIS RESULTS

5.6.1 Project Overview

From Figure 5.11, it can be observed that the overall project was centralized around the efforts of the four mechanical contractors M1 to M4 and, to a slightly lesser extent, the three electrical contractors E1 to E3. This finding is not surprising given the nature of such an overhaul project. In addition, the external vendor C1 was responsible for the largest portion of the scope within the off-site refurbishment works. By looking at the project through a network-level lens, such conclusions can help project managers develop a clearer overview as to the critical players with roles that are vital to the overall project's success.

5.6.2 Month-by-month Analysis

Next, a closer inspection of the project's monthly network layouts, and thus contractor collaboration patterns, can be inferred through investigating Table 5.2 and Figures 5.12 to 5.16. Focusing first on the ***network cores***, Table 5.2 indicates that the number of collaborating on-site contractors increased from seven in Months 1-7 to nine in Months 9 and 10, and subsequently remained constant at eight thereafter. It can also be observed from Table 5.2 and from Figures 5.12 and 5.13 that each monthly network is fully connected (i.e., ND = 1, as the total possible number of links exists between the collaborating contractors in the network core), which conveys that all on-site contractors remained interactive and collaborated on at least one task every month. Despite all months having the same ND values, a closer inspection of Table 5.2 shows varying AWDC values which indicates that the contractor/task interdependencies varied month-to-month. These AWDC values gradually increased from 159.40 in Month 1 to 420.09 in Month 6, then fluctuated thereafter before ending with a peak value of 468.41 in the final month, Month 13. Specific months experiencing high AWDC values (i.e., those exceeding the entire lifecycle AWDC threshold of 371.69) were Month 5 (405.70), Month 6 (420.09), Month 7 (411.48), Month 9 (396.13), and Month 13 (468.41). This finding conveys that such months require more focused management efforts to handle such high magnitudes of overall contractors/tasks interdependence and thus vulnerability to CPD—an insight that would not have otherwise been obvious through typical industry standard tools (e.g., through Figure 5.10). Another observation related to contractors/tasks interdependence concluded from Figures 5.12 and 5.13 was that the four mechanical contractors M1 to M4 and three electrical contractors E1 to E3 continuously occupied relatively central positions in the networks throughout the project lifecycle (due to the considered project's specific nature as alluded to earlier), while contractor X1 seemed to have more relatively moderate roles.

Moreover, a consistent finding was that contractors M1 to M4 and contractors E1 to E3 had almost identical work scopes and tasks assignments reflected through their essentially identical WDC values over time, as demonstrated from Figures 5.14 and 5.16. This finding emphasizes the critical need for effective communication and collaboration means to be facilitated within each of these groups. Turning to the ***network peripheries***, Table 5.2 and Figure 5.15 show that there was never a specific month where all five remote contractors participated together, but their roles rather alternated over time. More specifically, remote contractors were engaged until halfway through Month 10 before gradually pulling out, as can be observed from Table 5.2 and Figure 5.15.

5.6.3 Analysis by Work Package

Subsequently, a deeper examination of the project work packages that are shown in Figure 5.10 is discussed. Throughout the ***components disassembly*** in Months 1 and 2, there were high participations (i.e., the scope of work) and task interdependencies between M1 to M4 and only modest roles for E1 to E3 (see Figures 5.12a-b, 5.14 and 5.16). This situation is attributed to the fact that this work package mainly involved tasks of mechanical nature such as the uncoupling, removing and disassembly of several components including the rotor, generator shaft, turbine shaft, carbon shaft seal, control head, runner, diffuser, stub shaft, load cell (following its oil draining), headcover and exciter. Concerning the ***pre-assembly works*** between Months 3 and 8, there were more balanced scope involvements between both contractors M1 to M4 and contractors E1 to E3 (see Figures 5.12 c-g, 5.13a, 5.14 and 5.16) as the collective expertise of both disciplines was required for: 1) preparation of the erection bay that would be used for the shipping of certain components and the in-house refurbishment and assembly of other components; 2) in-house refurbishment of the rotor (including cleaning, ice blasting, cracks repairing and fan repairing), generator shaft, turbine shaft, carbon shaft seal, and control head; and 3) in-situ work as corn blasting the stator, cleaning the control head bearings and ventilating the exciter enclosure.

Regarding the ***components off-site refurbishment*** spanning from Months 3 to 10, contractor C1 was the critical remote contractor in charge of the largest portion of the off-site scope between Months 3 and 9 (see Figures 5.12c-g, 5.13a-b and 5.15) including the refurbishment of the runner, diffuser and stub shaft. Contractor C2 was involved from Months 3 to 6 (see Figures 5.12c-f and 5.15) and assumed responsibility for the refurbishment of the load cell. Contractor C3 was involved in Months 4 and 5 (see Figures 5.12d-e and 5.15) and undertook the refurbishment of the headcover. Contractor C4 was involved from Months 7 to 10 (see Figures 5.12g, 5.13a-c and 5.15) and performed the refurbishment of the exciter. As for the ***components assembly and commissioning*** from Months 8 to 13, contractor C5 interestingly emerged as a critical player (see Figures 5.13a-f, 5.14 and 5.16). Through crane operating and handling efforts, contractor C5 contributed heavily to other contractors (M1 to M4, E1 to E3 and X1) in the assembly, installation, alignment, coupling and securing/locking of all the previously mentioned disassembled components. Month 9 is a prime example of contractor C5's criticality (see Figure 5.13b), where this contractor was responsible for the most scope of work assigned to any contractor in any one month evident from their WDC value of 841.75 (see Figure 5.14). Contractor C5 was also the most heavily interdependent-working contractor within the month apparent from its WDC value of 510.99 (see Figure 5.16). The significant contributions of the crane handing contractor to the tasks of the mechanical and electrical contractors made it crucial to the continuity of the works and presented another unexpected insight that typically would not have been discovered using industry standard tools (e.g., through Figure 5.10).

Finally, although Month 13 had the same number of on-site contractors as the two preceding months (see Table 5.2), this last month had the most overall interdependence evident from its AWDC value (see Table 5.2 and Figure 5.16). By observing the month's network layout in Figure 5.13f, visible changes are seen compared to the layouts of the preceding months. Specifically, all contractors worked closely and interdependently to carry out the testing and commissioning of the installed components and systems according to the regulatory requirements—thereby proving a critical month for the successful close out of the project.

5.6.4 Network Layout Expansion

A final valuable feature is the network layout expansion which facilitates a more granular examination of any monthly network. Project managers can select a month to examine more closely and the expansion tool would provide a more holistic overview of that month. To highlight the tool's added features, the network of Month 8 in Figure 5.13a is expanded to yield the network in Figure 5.17 which provides a more detailed overview of contractors' scope of work and interdependent tasks in that month. Within this expanded layout, nodes contain the number of working days within the month (e.g., contractor C1 worked for 26 days on three levels of parallel activities), whereas links between any pair of contractors show the number of days they collaborated on interdependent tasks.

5.7 MANAGERIAL INSIGHTS

The developed resilient-driven management framework described above and subsequently applied within a practical setting, yields several key managerial insights that can enhance infrastructure project management and control from commencement to completion, serving both contractors, project managers, and other stakeholders. These insights are summarized in the following three subsections according to the project temporal stages.

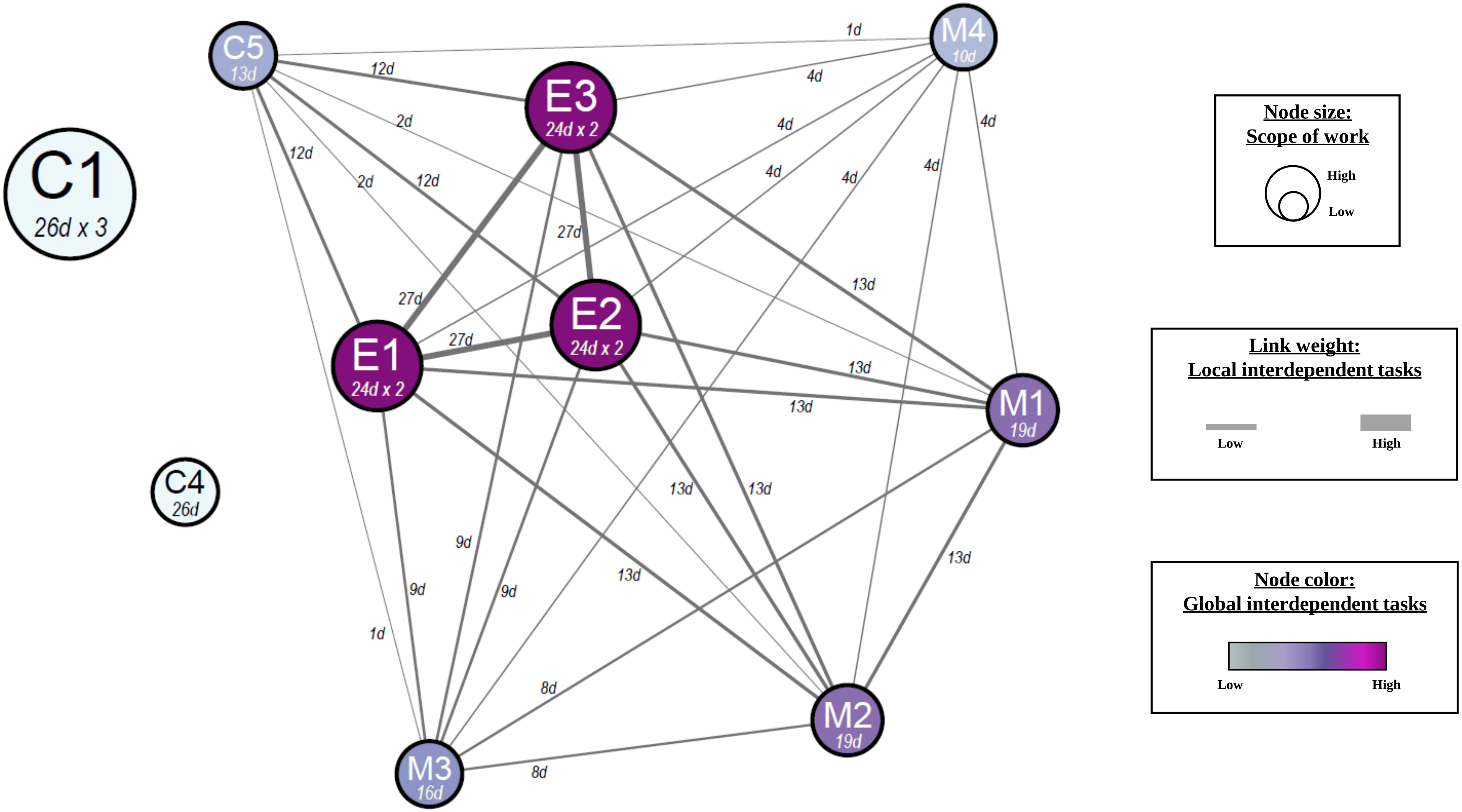


Figure 5.17: Network layout expansion for granular monthly overview of contractors’ scope of work and interdependent tasks in month 8 of the project lifecycle

5.7.1 Preliminary Project Planning

From the preliminary project stage, where strategic planning is key, the framework can equip project managers with a more in-depth assessment of project requirements, potential risks and difficult points in order to reach a clearer understanding of the management approach necessary to ensure project success through resilience to disruptions. For instance, project managers can secure stakeholder satisfaction and realize further savings by originally setting more realistic project durations and budgets from the preliminary stage, and continually monitoring adherence to these objectives during project execution. Moreover, the framework's capabilities in predicting the performance of future lifecycle stages through KPIs can guide resilience-driven schedule improvements. Specifically, if unsatisfactory performance under the current schedules is predicted, project managers are equipped with tools to analytically evaluate decision alternatives so that they can revise schedules until a satisfactory project performance level is attained. In addition, project managers can, through network centrality analyses, identify critical contractors with the highest CPD influences within the entire contractor network over different project lifecycle stages. Project managers can place more emphasis on these contractors from the onset through financial incentives/penalties, cautious resource allocation and timely inspection of finished works in a proactive effort towards the efficient functioning of the entire network.

5.7.2 During-Project Execution

During project execution, where project managers periodically review and adjust performance objectives, the framework can facilitate a more comprehensive project scenario planning to ensure that the updated objectives are compatible with the project's dynamic surrounding environment. However, when unforeseen events/disruptions that warrant sudden changes to the project schedules arise during execution, the need for automated adaptation capabilities becomes even more critical to rapidly adjust previous schedules to recover from such new unexpected situations. In such events, the framework's adaptive capabilities may enable rapid decision-making through generating self-organized networks and subsequently resilience-driven re-coordinated schedules. Such new networks and schedules can facilitate the project's rapid recovery and can prevent further CPDs from extending to subsequent months, thus ensuring project resilience throughout its lifecycle stages. With the aid of the network layout models and analyses, contractors can also visually and analytically communicate the interdependent nature of other contractors with their crews for improved awareness of potential risks within their own immediate networks and thus workflows and performances over different lifecycle stages. Moreover, and guided by the provided visualizations of interdependence between crews over the project lifecycle, contractors can more appropriately organize their crews' workspace allocations, task assignments and task supervisions.

5.7.3 Project Closure

Upon project completion, project managers can revisit the experienced events resolved through the framework tools and subsequently assess their original management approach capabilities pertaining to vulnerability identification, risk awareness and coping strategies. By understanding the shortcomings in their previous management approaches, managers can improve their approaches for future projects as necessary. Project managers can also benefit by incorporating insights, gained from previous projects, into future projects. As such, lessons learned can be reflected in better-informed future decision-making in terms of contractor and subcontractor procurement, task scheduling, and contractors' collaborative assignments within specific site locations and lifecycle stages.

5.8 CONCLUSIONS

Public infrastructure systems are crucial components of modern urban communities as they play major roles in elevating a country's socio-economic status. However, the inherent complexity and interdependence of infrastructure construction/renewal projects have left project performances hindered with multiple forms of disruptions (e.g., schedule delays and cost overruns) that result in long-term consequences such as claims, disputes, and stakeholder dissatisfactions. The key challenge facing the management of infrastructure projects addressed in the current study pertains to the successful coordination of the interdependent tasks assigned to the diverse set of contractors on site and the subsequent need for rapid re-coordination between such contractors in response to sudden disruptive events. In this respect, abstract, yet valuable, interdependence data of tasks and contractors are typically embedded within project schedules but are left unexplored and unexploited due to the limited capacities of the current industry standard tools.

In the current study, the power of complex dynamic network theoretic approaches is harnessed to develop a resilience-driven infrastructure project management framework (summarized in Figure 5.3). The developed framework empowers project managers with the ability to proactively mitigate the systemic risks of their projects' underlying complex interdependence-induced vulnerabilities and rapidly recover from cascade performance disruptions. First, interdependencies between contractors are modelled through a series of snapshot network layouts employing the first tool. The second tool focuses on generating dynamic networks representing the variations of these interdependencies across various site locations and over different project lifecycle stages. Further dynamic analyses are performed using the third tool to identify critical contractors with the highest potentials for disrupting performance in different site locations and lifecycle stages. Subsequently, key project performance indicators, tracked throughout intermittent lifecycle stages, are correlated with the dynamic network layouts using the fourth tool to proactively forecast the impact of collective vulnerabilities of contractors on the overall project performance. Finally, the fifth tool proposes optimization search techniques to generate adaptive (self-organized) network layouts and subsequently alternative coordination strategies between contractors. This tool will thus ultimately enhance the project's resilience against further interdependence-induced cascade performance disruptions.

To demonstrate the application and utility of the developed framework, a large-scale overhaul project of a hydroelectric power generation station was analyzed. The analysis revealed numerous insights that furthered both the comprehensive and granular understandings of the project with respect to key contractor influences, their interdependence-induced vulnerable collaborations, challenging durations and critical work packages undertakings—insights that would typically not have been revealed using available industry standard tools. These valuable insights can thus play crucial roles in steering multiple similar projects to success in terms of conforming to the planned outage cycle durations and budgets, achieving baseload generation supply and surplus, realizing significant revenues, maintaining safe and reliable long-term operations, and adhering to regulatory requirements. Although demonstrated herein on one application, the framework can tackle other more complex and diverse infrastructure projects.

Due to restrictions on KPI data imposed by the corporation, the current study does not extend the demonstrated application to cover the full proposed framework; however, the framework tools and procedures described herein can open new opportunities for future research studies to apply the framework in full given the availability of such data. Finally, it is important to reiterate that complex dynamic network theoretic- and other analytics-based approaches are not proposed to replace but rather to complement the expertise and sensible judgment of project managers and the capabilities of available analysis tools. Specifically, the enriched visual and analytical insights together with the proactive and rapid adaptation capabilities facilitated by the developed framework can empower the new paradigm of *resilience-driven management* of complex spatiotemporally dynamic projects.

5.9 ACKNOWLEDGMENTS

The authors are grateful to the financial support of the Natural Science and Engineering Research Council of Canada (NSERC) CaNRisk-CREATE program, and the INTERFACE Institute and the INViSiONLab, both of McMaster University.

5.10 REFERENCES

Apostolato, I. A. (2013). An overview of software applications for social network analysis. International Review of Social Research, 3(3), 71-77.

Aritua, B., Smith, N. J., & Bower, D. (2009). Construction client multi-projects – A complex adaptive systems perspective. International Journal of Project Management, 27(1), 72-79.

Avlijaš, G. (2019). Examining the value of Monte Carlo Simulation for project time management. Management: Journal of Sustainable Business and Management Solutions in Emerging Economies, 24(1), 11-23.

Barker, K., & Haimes, Y. Y. (2009). Uncertainty analysis of interdependencies in dynamic infrastructure recovery: Applications in risk-based decision making. Journal of Infrastructure Systems, 15(4), 394-405.

Barker, K., Ramirez-Marquez, J. E., & Rocco, C. M. (2013). Resilience-based network component importance measures. Reliability Engineering & System Safety, 117, 89-97.

Cantarelli, C. C., van Wee, B., Molin, E. J., & Flyvbjerg, B. (2012). Different cost performance: different determinants?: The case of cost overruns in Dutch transport infrastructure projects. Transport Policy, 22, 88-95.

Castillo, T., Alarcon, L. F., & Pellicer, E. (2018). Influence of organizational characteristics on construction project performance using corporate social networks. Journal of Management in Engineering, 34(4), 1-9.

Chassiakos, A. P., & Sakellaropoulos, S. P. (2005). Time-cost optimization of construction projects with generalized activity constraints. Journal of Construction Engineering and Management, 131(10), 1115-1124.

Chester, M., & Hendrickson, C. (2005). Cost impacts, scheduling impacts, and the claims process during construction. Journal of construction engineering and management, 131(1), 102-107.

Di Maddaloni, F., & Davis, K. (2018). Project manager's perception of the local communities' stakeholder in megaprojects. An empirical investigation in the UK. International Journal of Project Management, 36(3), 542-565.

Duan, S., & Ayyub, B. M. (2020). Assessment Methods of Network Resilience for Cyber-Human-Physical Systems. ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, 6(1), 03119001.

Dueñas-Osorio, L., Craig, J. I., Goodno, B. J., & Bostrom, A. (2007). Interdependent response of networked systems. Journal of Infrastructure Systems, 13(3), 185-194.

Eriksson, P. E., Larsson, J., & Pesämaa, O. (2017). Managing complex projects in the infrastructure sector—A structural equation model for flexibility-focused project management. International journal of project management, 35(8), 1512-1523.

Estrada, E., & Knight, P. A. (2015). A first course in network theory. Oxford University Press, USA. ISBN 978-0198726463.

Ezzeldin, M., & El-Dakhakhni, W. E. (2019). Robustness of Ontario power network under systemic risks. Sustainable and Resilient Infrastructure, 1-20.

Faysal, M. A. M., & Arifuzzaman, S. (2018, December). A comparative analysis of large-scale network visualization tools. In 2018 IEEE International Conference on Big Data (Big Data) (4837-4843). IEEE.

Flyvbjerg, B. (2014). What you should know about megaprojects and why: An overview. Project management journal, 45(2), 6-19.

Fu, T., Lyu, Y., Liu, H., Peng, R., Zhang, X., Ye, M., & Tan, W. (2018). DNA-based dynamic reaction networks. Trends in biochemical sciences, 43(7), 547-560.

Goldstein, M. (2006). Subjective Bayesian analysis: principles and practice. Bayesian analysis, 1(3), 403-420.

Gondia, A., Siam, A., El-Dakhakhni, W., & Nassar, A. H. (2020). Machine Learning Algorithms for Construction Projects Delay Risk Prediction. Journal of Construction Engineering and Management, 146(1), 04019085.

Gong, M., Cai, Q., Ma, L., Wang, S., & Lei, Y. (2017). Computational Intelligence for Network Structure Analytics. Springer Singapore. ISBN 978-9811045578

Han, S. H., Yun, S., Kim, H., Kwak, Y. H., Park, H. K., & Lee, S. H. (2009). Analyzing schedule delay of mega project: Lessons learned from Korea train express. IEEE Transactions on Engineering Management, 56(2), 243-256.

Hariri-Ardebili, M. A. (2018). Risk, Reliability, Resilience (R3) and beyond in dam engineering: A state-of-the-art review. International journal of disaster risk reduction, 31, 806-831.

Hashem M. Mehany, M. S., Bashettiyavar, G., Esmaeili, B., & Gad, G. (2018). Claims and project performance between traditional and alternative project delivery methods. Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 10(3), 04518017.

Hazir, Ö. (2015). A review of analytical models, approaches and decision support tools in project monitoring and control. International Journal of Project Management, 33(4), 808-815.

Hernandez-Fajardo, I., & Dueñas-Osorio, L. (2013). Probabilistic study of cascading failures in complex interdependent lifeline systems. Reliability Engineering & System Safety, 111, 260-272.

Ipsilandis, P. G. (2007). Multiobjective linear programming model for scheduling linear repetitive projects. Journal of construction engineering and management, 133(6), 417-424.

Jarkas, A. M. (2017). Contractors' perspective of construction project complexity: definitions, principles, and relevant contributors. Journal of Professional Issues in Engineering Education and Practice, 143(4), 04017007.

Kao, T. W. D., Simpson, N. C., Shao, B. B., & Lin, W. T. (2017). Relating supply network structure to productive efficiency: A multi-stage empirical investigation. European Journal of Operational Research, 259(2), 469-485.

Kereri, J. O., & Harper, C. M. (2019). Social networks and construction teams: Literature Review. Journal of Construction Engineering and Management, 145(4), 03119001.

Larsen, J. K., Shen, G. Q., Lindhard, S. M., & Brunoe, T. D. (2015). Factors affecting schedule delay, cost overrun, and quality level in public construction projects. Journal of Management in Engineering, 32(1), 04015032.

Leon, H., Osman, H., Georgy, M., & Elsaid, M. (2017). System dynamics approach for forecasting performance of construction projects. Journal of Management in Engineering, 34(1), 04017049.

Loizou, P. & French, N. (2012). Risk and uncertainty in development: A critical evaluation of using the Monte Carlo simulation method as a decision tool in real estate development projects. Journal of Property Investment & Finance, 30(2), 198-210.

Love, P. E., Teo, P., Morrison, J., & Grove, M. (2016). Quality and safety in construction: Creating a no-harm environment. Journal of Construction Engineering and Management, 142(8), 05016006.

Lu, M., & AbouRizk, S. M. (2000). Simplified CPM/PERT simulation model. Journal of Construction Engineering and Management, 126(3), 219-226.

Lu, W., Xu, J., & Söderlund, J. (2020). Exploring the Effects of Building Information Modeling on Projects: Longitudinal Social Network Analysis. Journal of Construction Engineering and Management, 146(5), 04020037.

Luo, L., He, Q., Xie, J., Yang, D., & Wu, G. (2016). Investigating the relationship between project complexity and success in complex construction projects. Journal of Management in Engineering, 33(2), 04016036.

McGee, F., Ghoniem, M., Melançon, G., Otjacques, B., & Pinaud, B. (2019, September). The state of the art in multilayer network visualization. In Computer Graphics Forum (Vol. 38, No. 6, 125-149).

McKinsey Global Institute (2017). Bridging infrastructure gaps: Has the world made progress?. McKinsey & Company. Retrieved April 19, 2020, from https://www.mckinsey.com/business-functions/operations/our-insights/bridging-infrastructure-gaps-has-the-world-made-progress#.

Mok, K. Y., Shen, G. Q., & Yang, J. (2015). Stakeholder management studies in mega construction projects: A review and future directions. International Journal of Project Management, 33(2), 446-457.

Nasir, D., McCabe, B., & Hartono, L. (2003). Evaluating risk in construction–schedule model (ERIC–S): construction schedule risk model. Journal of construction engineering and management, 129(5), 518-527.

Ndekugri, I., Braimah, N., & Gameson, R. (2008). Delay analysis within construction contracting organizations. Journal of construction engineering and management, 134(9), 692-700.

Nepal, M. P., Park, M., & Son, B. (2006). Effects of schedule pressure on construction performance. Journal of Construction Engineering and Management, 132(2), 182-188.

Newman, M. E., Barabási, A. L. E., & Watts, D. J. (2006). The structure and dynamics of networks. Princeton university press. ISBN 978-0691113579.

Park, H., Han, S. H., Rojas, E. M., Son, J., & Jung, W. (2010). Social network analysis of collaborative ventures for overseas construction projects. Journal of construction engineering and management, 137(5), 344-355.

Project Management Institute. (2017). A guide to the project management body of knowledge. 6th ed. Newtown Square, PA: PMI. ISBN 978-1628251845.

Pryke, S. D. (2004). Analysing construction project coalitions: exploring the application of social network analysis. Construction management and economics, 22(8), 787-797.

Revelle, W. (2020) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.1.3.

Sadeghi, N., Fayek, A. R., & Pedrycz, W. (2010). Fuzzy Monte Carlo simulation and risk assessment in construction. Computer‐Aided Civil and Infrastructure Engineering, 25(4), 238-252.

Sarkar, P., & Moore, A. W. (2006). Dynamic social network analysis using latent space models. In Advances in Neural Information Processing Systems (1145-1152).

Schloerke, B., Crowley, J., Cook, D., Briatte, F., Marbach, M., Thoen, E., ... & Larmarange, J. GGally: Extension to'ggplot2', 2018. URL https://CRAN.R-project.org/package=GGally. R package version, 1(0), 361.

Schröpfer, V. L. M., Tah, J., & Kurul, E. (2017). Mapping the knowledge flow in sustainable construction project teams using social network analysis. Engineering, Construction and Architectural Management, 24(2), 229-259.

Serrador, P., & Turner, R. (2015). The relationship between project success and project efficiency. Project management journal, 46(1), 30-39.

Sun, J., & Zhang, P. (2011). Owner organization design for mega industrial construction projects. International Journal of Project Management, 29(7), 828-833.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tomczak, M., & Jaśkowski, P. (2020). New Approach to Improve General Contractor Crew's Work Continuity in Repetitive Construction Projects. Journal of Construction Engineering and Management, 146(5), 04020043.

Wehbe, F., Al Hattab, M., & Hamzeh, F. (2016). Exploring associations between resilience and construction safety performance in safety networks. Safety science, 82, 338-351.

Wilkinson, S., Chang-Richards, A. Y., Sapeciay, Z., & Costello, S. B. (2016). Improving construction sector resilience. International Journal of Disaster Resilience in the Built Environment, 7(2), 173-185.

Xue, X., Zhang, R., Wang, L., Fan, H., Yang, R. J., & Dai, J. (2018). Collaborative innovation in construction project: A social network perspective. KSCE Journal of Civil Engineering, 22(2), 417-427.

Yates, J. K., & Epstein, A. (2006). Avoiding and minimizing construction delay claim disputes in relational contracting. Journal of Professional Issues in Engineering Education and Practice, 132(2), 168-179.

Yeo, K. T. (1995). Planning and learning in major infrastructure development: systems perspectives. International Journal of Project Management, 13(5), 287-293.

Yu, T., Shen, G. Q., Shi, Q., Lai, X., Li, C. Z., & Xu, K. (2017). Managing social risks at the housing demolition stage of urban redevelopment projects: A stakeholder-oriented study using social network analysis. International journal of project management, 35(6), 925-941.

Zhu, J., & Mostafavi, A. (2015). An integrated framework for assessment of the impacts of uncertainty in construction projects using dynamic network simulation. In ASCE International Workshop on Computing in Civil Engineering (355-362).

Chapter 6:
SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

6.1 SUMMARY

The research presented in this dissertation aims at developing data-driven strategies geared towards resilience management of infrastructure projects that can support managers and stakeholders with actionable decisions to mitigate the impacts of systemic risks at both the intra-KPI and inter-KPI levels of their projects. To achieve this aim, three main objectives were followed.

The *first* objective involved conducting literature surveys to identify key intra-KPI factors such as schedule delay factors (Chapter 2) and injury precursors (Chapters 3 and 4), and further compiling objective data from previous projects that describe qualitative or quantitative assessments of such factors and their influence on schedule delays (Chapter 2) or workplace injuries (Chapters 3 and 4) KPIs. This objective also included data pre-processing to constitute structured datasets suited for applying predictive analytics (in the second objective), as well as better understanding such datasets through exploratory data analysis and visualization.

The *second* objective involved developing ML-based predictive tools to predict potential KPI disruptions, based on input values from sets of intra-KPI factors, such that proactive response/mitigation strategies and/or contingencies can be deployed in a timely manner. This objective also included the hyper-parameter tuning and validation of the ML algorithms that drive such tools. In Chapter 2, project delay risk predictive tools were developed to facilitate predictions of project schedule delay extents described as percentages of original schedule duration. In Chapter 3, injury severity risk level predictive tools were developed to enable identifying high-risk worksites in projects. In Chapter 4, predictive spatiotemporal site safety risk models were developed to generate forecasts of risk leading indicators (e.g., injuries' financial implications and body parts most likely affected) across different zones and over project lifecycles.

The *third* objective involved developing CDNT-based tools, in Chapter 5, to model the interdependences between project on-site contractor crews through dynamic networks, and further correlating such network behaviors with sets of KPIs. This objective then included facilitating predictive and adaptive solutions against potential interdependence-induced cascade disruptions at the inter-KPI level in the form of alternative project schedules and thus task re-arrangements/contractor re-coordination, such that, in the case of disruptive events, the most important set of KPIs can be rapidly restored.

6.2 CONCLUSIONS

In lieu of re-listing conclusions of the individual chapters, which can be found in Chapters 2 through 5 of the dissertation, this section is reserved for a discussion on the over-arching conclusions that can be drawn from the current work:

- The results of the exploratory analysis of intra-KPI factors, such as that presented and discussed in Chapter 2 for schedule delay factors and in Chapter 3 for injury precursors, revealed a deepened understanding of the complex nature of construction projects evident from the interdependence that exists, not only between such factors among themselves but also between the factors and their respective KPIs. Although such interdependence was explored only for schedule delay and workplace injury KPIs within the current work, similar analyses can be extended to unlock a better understanding of the interdependencies existing within other KPIs in construction projects.
- The aforementioned complexity and interdependence guided the selection of the analytics methods used in this work. ML was used in Chapters 2 through 4 because it has gained significant recognition for its capability to model and predict outcomes of complex interdependent systems while avoiding prespecified modelling assumptions, unlike in statistical-based methods for instance, which was an important consideration when dealing with systems

whose behaviors are largely complex/interdependent and thus unknown a priori. The excellent predictive performance of the developed ML-based predictive tools within this dissertation should, in theory, endorse ML as a viable option for tackling systems related to other KPIs in construction projects.

- The developed KPI predictive tools can be utilized to support both *proactive* and *dynamic* project management strategies through: 1) the ability to *proactively* assess project delays or worksite injuries from earlier project stages based on delay factors or injury precursors, respectively, identified from these stages; and 2) the continuous ability to adjust input values of such factors across different sites, as the project progresses, as scopes change or as more information becomes available, which can enable a better capture of the *dynamic* nature of construction work and reflect the current/real-time state of a project.
- The tools developed for schedule delay and workplace injury KPI prediction included both black-box and glass-box models (DT-based herein). Although the black-box models outperformed their counterparts across both considered KPIs, the differences in performance evaluation measures were closely comparable. In fact, leveraging glass-box models can be particularly opportune when: a) the outcome variable (i.e., KPI classes) distribution is relatively balanced; b) the input variables (i.e., KPI factors) do not contain

significantly strong predictors; and c) the dataset is not high-dimensional. Where possible, leveraging glass-box models may be useful since such models can not only support quantitative predictions but also enable qualitative interpretations of the input-output interdependencies, which are typically hidden in conventional ML methods resulting in a reluctance on the part of project managers to adopt them. Based on our discussions with project and safety managers, hesitancy to embrace change, such as a new predictive tool, is typically related to questions such as: a) why does a model predict the way it does? b) what to make out of the model's predictions? and c) how to trust a model's predictions? They informed us that they would favor and consider adopting white-box-based tools, such as the DT-based models developed in this dissertation, since such models are rule-based and display how the decisions are taken at each step, which ultimately enables an understanding of what the model did or might have done.

- The CDNT-based resilience-driven tools developed in Chapter 5 are aimed at empowering project managers with the ability to proactively mitigate the inter-KPI systemic risks of their projects and rapidly recover from cascade KPI disruptions. The tools enable a) visually modelling the dynamic interdependencies of on-site contractor crews across various site locations and over different project lifecycle stages through dynamic network layouts; b) analyzing such network layouts to identify critical contractors with the highest potentials for disrupting performance in different site locations and lifecycle stages; c) correlating such network layouts with sets of KPIs to proactively forecast the impacts of the underlying interdependence-induced vulnerabilities associated with the current contractor arrangement on overall project performance measured through multiple KPIs; and d) generating adaptive (self-organized) network layouts and subsequently alternative schedules/coordination strategies between contractors that quickly restore the project to a set of desired KPIs. It should be noted that the tools developed and described in this chapter can also be applied to other complex and diverse project types that are not limited to only construction or overhaul projects.

- In retrospect, the tools developed in this work support the two dimensions of a resilience-driven management project approach. The first dimension involves predicting potential KPI disruptions based on real-time and dynamic project conditions, which is thus supported by the schedule delay risk predictive models and the site safety risk predictive models developed in Chapters 2 through 4. The second dimension involves deploying adaptive solutions against potential inter-KPI cascade disruptions such that rapid restoration of the most important set of performance objectives can be restored, which is thus supported by the tools developed in Chapter 5. As such, organizations may use the suite of tools developed in this thesis based on their current management needs. For example, organizational groups that prioritize a specific performance objective, such as those that are time- or safety-focused, may utilize the tools from the first resilience dimension. Higher levels of the organization which are interested in multi-objective management may utilize the tools from the second resilience dimension, where one or more KPIs can be triggered to guide the adaptive solution generation.

Ultimately, this dissertation presents a cohesive body of work that is expected to: a) deepen construction organizations' understanding of the interrelated and dynamic nature surrounding large-scale and complex projects; b) address the need for transforming historical data of completed projects into useful business value that enables organizations to make evidence-based data-supported decisions to mitigate project KPI disruptions; and c) influence a real change in the way organizations improve overall project resilience under different systemic risk levels and interdependence-induced vulnerability extents.

6.3 RECOMMENDATIONS FOR FURTHER RESEARCH

The research presented in this dissertation contributes to the state-of-the-industry with data-driven resilience-guided strategies to support complex infrastructure project management through enabling the prediction of and adaptation to KPI disruptions in project environments with systemic risks, such as those between factors influencing project KPIs and between such KPIs. This research has also highlighted several avenues for possible extensions to expand the state-of-the-industry regarding the development of other data-driven resilience-based approaches to complex project management:

- The schedule delay prediction models in Chapter 2 were developed using a relatively small dataset of previous project information as similar data is currently not available in open literature to the best of the authors' knowledge. Larger datasets under a wider range of delay factor quantitative/qualitative assessments are therefore needed to increase the models' accuracy and extend their applicability. Larger datasets can also open opportunities for employing other ML algorithms, such as those used in other chapters of this thesis, which can achieve higher accuracy (e.g., ensemble algorithms) or facilitate numerical/quantitative predictions of delays (e.g., artificial neural networks, genetic algorithms or genetic programming).

- While the models developed in Chapter 2 are useful for predicting schedule delay at the project level, the ML methodologies described in the chapter can be employed to yield models that are useful for task-level delay estimation. Such models would need to be trained on datasets that are task-granular and include assessments of delay factors/sources on singular task delays along with other general task descriptors (e.g., nature of work, task scheduled duration, no. involved workers, whether a rework incident).

- The safety risk prediction models developed in Chapters 3 and 4 were demonstrated/trained on datasets describing injury precursors such as worksite-, work method- or worker-related hazards/conditions influencing worksite safety and injury outcomes. Such precursors were collected as part of post-incident inspections under legislations requiring employers to report the conditions surrounding worksite incidents and near hits. In future studies, it might be useful to employ datasets that are collected as part of pre-job inspections and safety audits where collections/identifications of precursors/hazards are free from any form of accompanied influence/newly developed awareness in the aftermath of an incident when its nature and narrative are known. Such training data can also endorse the reliability of using such models during pre-job inspections and attach multiple usages to the information collected.

- The full extent of the framework developed in Chapter 5 was not applied/demonstrated on the considered case study due to restrictions on key data imposed by the contacted corporation. As such, the framework tools and procedures described can open new opportunities for future research to study adaptive contractor network behaviors as driven by key KPI correlations, given the availability of sufficient data.

- The key challenge facing the management of infrastructure projects addressed in Chapter 5 pertained to the effects of contractor network interdependence over time on overall project performance. As such, relationships and disruptions across other stakeholder networks were not fully examined but are still needed especially in complex projects that involve many stakeholder groups. To develop a more holistic perspective on project dynamics, future research efforts can be directed toward employing multiplex network approaches that can extend the investigation to one which combines cross-sectional and longitudinal network designs of stakeholder cross-network interactions over time and the multilateral effects on overall project outcomes.