A data-driven approach for deploying safety policies for schedule planning in industrial construction projects: a case study

Maedeh Taghaddos [1]; Estacio Pereira[2]; Carlos Osorio-Sandoval[3]; Ulrich Hermann[4]; Simaan AbouRizk[5*]

[1] PhD Graduate, Department of Civil and Environmental Engineering, University of Alberta, Donadeo Innovation Centre for Engineering, 9211 116 Street NW, Edmonton, Alberta, Canada, T6G 1H9; Email: taghaddo@ualberta.ca

[2] Assistant Professor (Teaching), Department of Civil Engineering, University of Calgary, ENF 208, 2500 University Dr NW, Calgary, Alberta, Canada, T2N 1N4 Email: estacio.pereira@ucalgary.ca

[3] Assistant Professor, Department of Civil Engineering, University of Nottingham, B74 Coates Building, University Park, NG7 2RD, Nottingham, United Kingdom Email: carlos.osorio@nottingham.ac.uk

[4] Manager, Construction Engineering, PCL Industrial Management Inc., 9925 56 Ave., Edmonton, AB, Canada T6E 3P4; Email: rhhermann@pcl.com

[5] Professor, Department of Civil and Environmental Engineering, University of Alberta, 5-080 NREF, 9105 116 St NW, Edmonton, AB T6G 2W2, Email: abourizk@ualberta.ca

*Corresponding Author; 5-080 NREF, University of Alberta, Edmonton, Alberta, Canada T6G 2W2

ABSTRACT

Construction, by nature of its work, is more accident-prone than other industries despite advancements in improving safety performance. Proactive mitigation and assessment of safety performance on construction projects remain challenging due to the difficulty of acquiring, storing, and using data to produce accurate predictive models. This research is focused on devising methods that allow decision-makers to leverage existing data in the planning phase to streamline the development of predictive models. A data-driven approach to predict the probability of a safety incident occurring in a given construction project and within a novel discipline-level schedule is presented. By implementing the proposed model, decision-makers can evaluate and mitigate the risk of a given project incident occurring by deploying discipline-level safety policies in the planning phase and modifying the schedule accordingly. A predictive model is developed based on selected safety-related metrics extracted from a dataset comprising daily payroll data and incident reports, which represent 28 million working hours within eight different industrial construction projects in

Canada. The model was implemented in a case study based on an industrial project to demonstrate the framework's functionality and practical utility during the project planning phase. The results show that the revised safe plan can be achieved by incorporating safety considerations in the planning phase.

PRACTICAL APPLICATIONS

This research provides a practical solution for enhancing safety in the planning phase of construction projects through a data-driven model. By leveraging existing historical data, decision-makers can predict potential safety incidents within specific disciplines without the need for detailed quantitative planning information. This approach also enables effective adjustments to be made to the schedule in order to mitigate risks. Furthermore, the discipline-level approach facilitates proactive safety planning by implementing discipline-specific safety policies that align with the unique characteristics of each discipline.

Using a case study based on an industrial project, the proposed framework demonstrates its functionality and practical utility by identifying suitable safety-related metrics that construction enterprises typically record. These sources can include safety-related data, such as incident reports, as well as data recorded for other purposes, such as payroll data. The results highlight that incorporating safety considerations in the planning phase enables the development of a revised safety plan. In conclusion, the key takeaway is that by considering safety-related metrics and utilizing HR data available in all companies, organizations can proactively assess and improve safety performance.

INTRODUCTION

According to the Association of Workers' Compensation Boards of Canada (2019), the number of accepted lost-time claims per year in the construction industry has increased by 9% from 2016 (25,514) to 2018 (27,952). Additionally, the number of fatalities in the Canadian construction industry is higher than in any other industry, with 202 deaths recorded in 2016 and 199 in 2018 (AWCBC 2019). According to estimates from the International Labor Organization (ILO), an average of 4% of annual global Gross Domestic Product (GDP)—adding up to trillions of dollars—is lost due to direct and indirect costs incurred as a result of occupational accidents and diseases, which includes lost working time, workers' compensation, interruption of production, and medical expenses (Takala et al., 2014). Although AFPM (2017) demonstrated that the incident rates had decreased greatly in the heavy industrial sector when measured over the past 30 years (Recordable Incident Rate reduced by more than 1000%), there is still great concern about events with a serious injury or fatal consequences as they have remained consistent. Due to the construction industry's impact on Canada's economy, the improved safety performance of construction projects is critically needed.

Based on the literature, there are many causes for accidents at construction sites, and a holistic approach has to be considered to mitigate hazards (Ahn et al. 2020; Mohammadi et al. 2018; Pereira et al. 2020). Decision-makers should be aware of how they can impact project safety performance. For instance, poor project scheduling can impact safety performance by 1) generating delays and increasing the pressure on workers (Han et al. 2014; Mitropoulos et al. 2005); 2) developing site congestion — increasing the probability of workers being struck-by heavy equipment (Ahn et al. 2020; Zhang et al. 2018); 3) increasing the number of inexperienced people on, and 4) increasing the crew size and consequently reducing the number of inspections and safety observations (Jiang et al. 2015).

Although there is no doubt about how different aspects of a project can impact the occurrence of incidents, it is still challenging to integrate decisions from various project factors to mitigate incidents and assess safety performance proactively. Issues may arise because construction practitioners have difficulty identifying safety measures, collecting data, and proactively assessing the impact of factors on safety performance (Pereira et al. 2018a).

From a theoretical perspective, three levels can be considered for safety assessment: the first is project level, the second is discipline level, and the third is crew level. The project level has the least details (broadest) to plan for safety, and the crew level has the most details (finest). The most common approach is assessing safety performance at a project level, which may not be feasible in practice as there is insufficient detailed information available for safety planning (Goh and Chua 2013; Karakhan et al. 2018; Lingard et al. 2017; Salas and Hallowell 2016; W Guo and Wing Yiu 2016). Therefore, practitioners find making decisions based on project-level models challenging since their results may neither be economical nor produce the required output. For example, if a project has a large crew size, reducing the crew size in a specific discipline may not necessarily improve the project's safety performance as much as another measure (e.g., workers' age). In other words, evaluating the crew size for each discipline is essential rather than the overall evaluation in the entire project context. Hence, discipline-based evaluations are preferred rather than project-based evaluations.

Another practical challenge is related to data acquisition and utilization. Small companies have difficulty collecting safety-related data (e.g., available resources (Bavafa et al. 2018)). Some companies may not have the knowledge or skills required to produce safety predictive assessment models (Boon et al. 2019) using techniques like Machine Learning (ML) algorithms. Large organizations may have issues with data silos — the data repository — as it is not easily accessible to the entire organization. Although Pereira et al. (2020) demonstrated that integrating different department databases may lead to the development of better models to predict safety, concerns about privileges and sensitive information hinder database integration in practice. Moreover, departments may collect information from different levels (e.g., project, discipline, and task), which may make it impossible to integrate all into one single model. This study proposes to use Human Resources (HR) data (e.g., payroll), a common database among all companies, to address the data acquisition problem.

The HR dataset is one of the most reliable datasets, with considerable attention given to it because of its monetary aspect. Moreover, it is very informative and is often made available due to regulations. Several safety-related measures can be identified from this database, such as workers' age and experience, job type, number of workers, new workers' rate (i.e., the rate at which new workers are deployed to work on the project), the number of supervisors, etc. This manuscript presents a case study that uses the payroll database in a data-driven approach for deploying safety policies for discipline-level schedule planning. The approach described herein was able to identify several safety-related measures that companies can use to proactively measure safety performance at the discipline level without relying on subjective decisions.

Based on these measures, a predictive safety (or incident) assessment model is suggested, and a case study on using the findings in practice is presented. The approach presented in this manuscript may help organizations use their payroll database to derive safety models to proactively test scenarios and control the safety performance on construction projects.

The novelty of this paper is to **proactively** improve safety in a **discipline-level schedule** (in the planning phase). This prediction is through the **novel data acquisition** from a reliable, unbiased, and objective payroll dataset, which is commonly available in all companies. Moreover, this research can facilitate the safety forecast by planning based on qualitative data to evaluate the safety of various possible what-if scenarios. Therefore, the practitioners can consider the safety of the schedule in the planning phase.

LITERATURE REVIEW

Deficient project scheduling is considered a root cause of construction accidents as it leads to time pressure on workers, with subsequent problems including trade overlap, crowded workspaces, and reduced attention to detail (Haslam et al. 2005; Neale and Gurmu 2021). It is believed that safety can be improved by considering activity information regarding the number of workers, including their occupation types, in the schedule (Choe and Leite 2017), and by minimizing the number of workers on congested sites (Anvari et al. 2016). However, empirical data — needed to estimate risks accurately based on these constraints — is usually not captured, leading to a decision-making process that relies on subjective opinions. In addition, attributes such as the fundamental characteristics of the work site and environment (e.g.; weather, uneven surfaces, specific tools, and equipment (Hallowell et al. 2020)), as well as features of the workforce (e.g.; age, experience, crew size, and specific trade) can also contribute to safety incidents. The combination of these attributes that define the overall work environment can be used to predict safety outcomes (Hallowell et al. 2020; Tixier et al. 2016a).

Construction companies typically record some information on several safety indicators at the project level (e.g., site inspection logs, hazard reports, injury reports, etc.) to meet the regulator's requirements (Versteeg et al. 2019). Usually, this information is only provided to the safety department and is not used for predictive purposes (Pereira et al 2020). However, companies still need to be proactive and incorporate measurements of valid and reliable metrics that may be causally related to the occurrence of incidents or injuries (Lingard et al. 2017; Versteeg et al. 2019). New technologies, like wearable devices, automated data collection, bar codes, etc., enable companies to capture more data related to safety issues (Ahn et al. 2019). Further research in the area is still needed to access quality data. Nevertheless, there are still practical challenges on how to use the data to improve safety management in practice.

To analyze the collected data stored in different data sources, statistical and ML models have been applied widely to predict safety outcomes in construction projects at the early design stages. As stated above, one practical challenge is the lack of consistency regarding the safety outcomes and predictors measured across projects and organizations. This lack of consistency may result in the need for different prediction approaches on a case-by-case basis. For example, Esmaeli et al. (2015) tested the validity of generalized linear models to predict safety outcomes based on a large volume of data on attributes that cause "struck-by" accidents. Poh et al. (2018) used an ML approach to develop a model that forecasts accident occurrence and severity of construction worksites based on project-related and safety-related input features. Kang and Ryu (2019) proposed a Random Forest model that can predict occupational accidents based on accident and weather data and suggest which management features significantly contribute to the forecast. Sarkar et al. (2019) developed optimized ML-based models to predict incident outcomes at the workplace using Support Vector Machine (SVM) and Artificial Neural Network (ANN) algorithms on incident reports data. Baker et al. (2020) used a large dataset of over 90,000 incident reports and various ML algorithms to develop a model that predicts injuries and their severity and shows which attributes have high predictive power when the safety outcomes are external and independent.

Although several ML algorithms have been proposed, their use in proactively assessing safety performance in companies is not common. Several challenges exist since it is expected that a single set of metrics may not be suitable for all construction industry sectors (Nasir et al. 2012). These issues present a challenge regarding the selection of data that should be collected to predict safety performance. Therefore, we propose a data-driven method that leverages HR data, commonly recorded for reasons beyond performance metrics, and allows for an adequate model selection based on the available data. This approach should be flexible enough to incorporate new variables as their impact on safety is better understood.

METHODOLOGY

This paper reports on the experience obtained in a case study that implemented a data-driven approach to predict the probability of a safety incident occurring within a month based on preliminary schedule planning information. The predictive model used in this approach was developed based on daily payroll information from previous projects and dated safety incident reports. The model was implemented in a hypothetical case based on an industrial project to demonstrate the framework's functionality during the project planning phase. Figure 1 depicts an overview of the methodology followed in this study.

Project Background

The functionality of the proposed framework is illustrated through a hypothetical case study inspired by a previous study (Taghaddos et al. 2021). This case study demonstrates that an industrial construction enterprise can use the data available from various departments within the enterprise to devise methods for predicting safety incidents based on discipline-specific information. The case study also exemplifies the framework's functionality during the project planning phase. The case study focuses on the planning phase of a small (6-month) in-situ drainage oil sands project in Alberta, Canada, and mainly involves five disciplines: civil, operators, pipefitters, electrical workers, and ironworkers.

The case study analyzes the risk of accident occurrence over the 6 months of the project at the planning phase based on qualitative data. Data categorization was implemented based on 28 million working hours of historical data. Utilizing the discipline-based incident predictions generated by the proposed model, the project plan was adjusted accordingly to mitigate safety risks. This case study provides a concrete example of how a construction enterprise can put to use existing data within the proposed framework and apply it to a real-world scenario. Moreover, it demonstrates the manner in which the framework can be tailored to meet the unique needs of various disciplines within the construction industry.

Data Collection

A large dataset of approximately 28 million working hours was collected across eight industrial construction projects in Canada. The dataset reports different features obtained from daily payroll information: working hours, age, work experience, time working on the project of workers and foremen, crew sizes, number of operators, changes in the number of workers in the project, and project progress. The data were categorized into five trade disciplines (electrical, ironworkers, pipefitters, civil, and operators).

Safety reports were collected to determine the monthly occurrences of incidents related to each discipline in each project throughout the data collection period. Incident reports, lost-time injuries, medical aid injuries, and modified work injuries were considered safety incidents. The payroll information dataset and the monthly safety incident occurrence information were collated and integrated into a single dataset. Table 1 shows the features contained in the final dataset. A monthly data record was collected for each discipline in each project.

The data is split into four qualitative categories: very high, high, moderate, and low. The value categories range from 4 (very high) to 1(low). This qualitative categorization can ease the use of the model in the planning phase. For example, the practitioners have some ideas about time-dependent crew size during the project (i.e., how it varies during construction) but do not yet have the exact numbers. For example, let's assume July is the peak month for the structural-steel discipline's work in a particular project; hence, one can predict a large crew size will be required in this case. Such high-level prediction can easily be combined with planning to improve safety measurements.

Table 2 shows the number of data points collected for each discipline in this study. Each data point represents information (reported monthly) on the features listed in Table 1, including the occurrence of each reported safety incident (i.e., the target variable). The incident rates per discipline were found to be as follows: ironworkers (34.93%), pipefitters (52.72%), civil (47.45%), operators (3.78%), and electrical workers (20.74%).

Model Development

The final dataset was retrieved from eight industrial construction projects. Ten input features were selected by applying the Boruta feature selection algorithm, which iteratively removes features that prove to be less relevant than random probes (Kursa and Rudnicki 2010). These ten features are categorized into five different trade disciplines. Five different ML models (Taghaddos and Mohamed 2021): SVM, Decision Tree, Naïve Bayes, Naïve Bayes (Kernel), and Fast Large-Margin, were developed from the resulting dataset containing the ten selected predictor features and the target variable (whether or not a safety incident occurred). A cross-validation method was implemented to select the model that worked best for the collected data by comparing the prediction performance of the five developed models (Zhang and Yang 2015) using two measures:

accuracy and incident recall. Accuracy measures the percentage of correctly predicted records, whereas incident recall measures the true positives recognition rate (Al-Turaiki et al. 2016). In the context of construction safety, a false negative (i.e., predicting an incident as a "No incident") is more expensive than a false positive (i.e., predicting a safe situation as an "incident") because false positives only impose precautions to the system. For this study, an incident recall with a 75% threshold was considered an important criterion for model selection. The framework was developed using the educational version of the well-known data-mining tool, RapidMiner Studio (Mierswa and Klinkenberg 2018). A screenshot of the developed model, as it appears in the data-mining software, is provided in Figure 2.

Model Implementation

The proposed model was implemented in the hypothetical case study described above to demonstrate its successful utilization in adjusting the project plan based on discipline-based incident predictions. The model was applied to predict monthly occurrences of incidents on the original program with planned information. A sensitivity analysis was performed on different planning strategies considering modifications to the predictor features to reduce the number of predicted incidents.

RESULTS

Table 3 shows the ten selected predictor features. These features are selected by applying the Boruta feature selection algorithm to the dataset and using experts' opinions.

Five ML models were then developed and applied to the dataset to predict the target variable (i.e., whether or not an incident occurred) based on the selected predictor features. This process is illustrated in Figure 2. In this figure, "CV" represents the cross-validation techniques in each model. The accuracy and incident recall of the prediction performance of the models were measured and compared after applying a cross-validation method. Table 4 shows the prediction performance measures of the five models. Based on the results in Table 4, the Naïve Bayes (Kernel) model was selected to demonstrate the framework's functionality in a hypothetical case study.

Similar to the eight-project data stored in a database of industrial projects, the case study project involves discipline-specific information. As such, the safety prediction incident results are also discipline based. Moreover, the predictions are time-dependent since the prediction depends on the features for the selected time period. Table 5 presents the original project schedule (Scenario 1). As demonstrated in the framework description section, each discipline is considered an individual data point. The framework analyzed 28 entries for 6 months (civil, pipefitters, and ironworkers) or 4 months (electrical/operators).

In Scenario 1 (Table 5), the framework assessed the entries based on the trained model and with the original planning strategy, it predicted ten incidents (which is equivalent to 35% of the entries) within the 6 months — Pipefitters (4), Civil (2), Ironworkers (3), and Electrical (1). In Scenario 2 (Table 6), a tentative planning strategy is used to reduce the number of accidents. Due to the framework's ability to consider a variety of strategies capable of reducing the likelihood of an accident, the following discipline-based mitigation strategies were adopted (individually or grouped): decreasing the percentage of foremen older than 50 years, decreasing the rate at which new workers are deployed to work on the project, reducing the crew size, increasing workers with more than 3 years of experience, and reducing discipline working hours. The output results shown are accident-free, reinforcing the framework's ability to test different planning strategies to improve the project safety performance.

It is important to mention that just one discipline (Pipefitter – month 3) is required to reduce the working hours. This strategy should be considered carefully since reducing working hours may lead to further project completion delays and/or increase production pressure on workers in the months ahead. The case study demonstrated that discipline-level data can produce better-tailored strategies than those developed for the project level. As Pereira et al. (2020) suggested, crew size is a factor that may impact project safety performance, and increasing the number of foremen may reduce the likelihood of accidents. The framework presented in this research demonstrated that reducing the crew size, not considering the discipline, may lead to increased project cost and would not improve safety performance. Moreover, the framework also demonstrated that schedule planning strategies such as new workers rate (i.e., the rate at which new workers are deployed to work on the project), discipline working hours, discipline-based progress, and accumulative working hours should be considered concomitantly with their impact on the project safety performance.

While some features can be considered in the planning phase, others may assist the organization in identifying flaws in the Safety Management System. Workers' and foremen's age features show that a reinforcement of the organization's policies should account for this specific group to ensure they are better prepared to identify hazards or follow safety procedures. The impact of workers' experiences also demonstrated that safety induction and safety training should assess the workers' abilities to retain the course knowledge and apply it in practice.

DISCUSSION

This case study validates the premise that the proposed predictor features in Table 3 can predict the occurrence of safety incidents, enabling an informed decision-making process regarding discipline-level schedule planning. The approach can simulate risks for any context, including new work if the fundamental attributes remain stable (Hallowell et al. 2020). The proposed framework predicts the likelihood of a safety incident occurring, overcoming a common limitation of traditional attribute-based safety risk assessment, which predicts the outcome of an incident should one occur (Choi et al. 2020; Esmaeili et al. 2015; Hallowell et al. 2020; Koc et al. 2021; Tixier et al. 2016a; b).

The proposed framework uses historical data, both at the business and project management levels, to discover data-driven knowledge and use it to support project management decisions, as advocated by You and Wu (2019). The study emphasizes the importance of effective data management in the context of construction safety. The method incorporates metrics extracted from daily payroll data, which are typically not used in traditional safety planning. By leveraging existing data, the method streamlines the development of predictive models for safety incidents, equipping decision-makers with a useful tool for the proactive assessment and mitigation of risk. Another novel feature of the developed framework is its ability to adjust the construction plan based on safety incident predictions in order to mitigate risk. These adjustments could include changing the crew composition, crew size, or time-dependent discipline working hours, all of which can affect the project duration.

The case study presented in this paper demonstrates how a data-driven model can be incorporated into the scheduling process contributing to safety planning. This finding agrees with Yi and Langford (2006), who advocate for scheduling construction to reduce accident risks. The suggested discipline-based scheduling is aligned with Hallowell and Gambatese (2009), who recommend considering risks based on activities to target high-risk activities in safety programs.

The feature selection process results align with other studies predicting safety outcomes. For example, Rivas et al. (2011) identified "task duration in hours", "length of time doing the job",

"job type" and "worker age" within the five most relevant features predicting accidents. Poh et al. (2018) used the Boruta algorithm, and the features "percentage of project completion" and "average monthly project manpower" were within the selected features to predict accident occurrence and severity. Choi et al. (2020) also found that "age" and "service length" were among the most important factors in predicting the likelihood of fatal accidents. These factors are commonly associated with accident precursors (Pereira et al. 2018b). While the proposed model does not attempt to uncover underlying causality, these alignments with knowledge of the construction safety domain were essential during the development of the model and the subsequent interpretation of its predictions (Mannering et al. 2020).

CONCLUSION

This manuscript proposes a novel data-driven approach for deploying safety policies for discipline-level schedule planning. This novel approach enables practitioners to account for safety considerations in the planning phase and proactively make appropriate decisions without needing detailed quantitative information. Five ML models were developed from payroll data collected in eight large industrial construction projects. The accuracy and incident recall of the prediction performance of the models were measured and compared to select the model that worked best on the collected data. Subsequently, the practical utility of the model was demonstrated through a case study.

The findings reveal that the predicted occurrence of safety incidents can be reduced by modifying the predictor features during the project's planning phase, achieving a safer planning strategy. In the case study project, the original plan and schedule were revised based on discipline-specific tentative planning strategies—e.g., decreasing the rate of new workers and crew sizes (to varying degrees depending on the discipline). Accordingly, the incident rate was reduced from 35% to 0%, resulting in an incident-free plan. The discipline-based safety plan for the early planning stage as proposed herein is beneficial to practitioners in that it provides the basis for expanding the plan to the work-package level later in the project.

This study makes three important contributions to knowledge and practice in this domain. First, it provides a framework for proactive safety improvement in the planning phase and for deploying discipline-level safety policies by identifying suitable safety-related metrics that construction enterprises typically record for other purposes. In this manner, it helps with tackling the large volumes of project-level data to identify, capture, and analyze the features that affect safety performance by leveraging existing data for model development. Moreover, the discipline-based approach used in the case study demonstrates the adaptability of the proposed framework to meet the discipline-specific needs, and align with the unique features, of the construction industry.

Second, it provides a novel data acquisition method. Specifically, this study demonstrates that payroll data and incident reports, which tend to be more reliable and unbiased than other data sources due to their monetary/regulatory nature, can be used to develop a model for safety-related decision support.

Third, this study integrates discipline-level scheduling with safety prediction. A key consideration in this regard is that the discipline-specific level of scheduling is not so high-level as to miss the vital discipline-specific features that are important in decision-making (such as in the case of project-level scheduling), and not so detailed as to make planning and decision-making cumbersome (such as in the case of crew-level scheduling).

The authors believe that this research can help project practitioners identify which data should be collected in their projects and define strategies to improve their construction plans in terms of safety based on insights emerging from the data. By leveraging the data-driven discipline-level safety prediction model, project teams can make informed decisions and implement proactive measures to enhance safety performance. Furthermore, the model's flexibility allows for the inclusion of additional factors specific to each project, ensuring a comprehensive and tailored approach to safety management. For example, in a heavy construction project involving excavation work, factors such as the type of equipment used (e.g., excavators, bulldozers, etc.), as well as the competency of equipment operators, can significantly influence safety outcomes. Ultimately, this research aims to contribute to the advancement of construction safety practices by promoting evidence-based decision-making and proactive risk mitigation strategies.

LIMITATIONS and FUTURE WORK

While the developed framework has been found to be capable of predicting the probability of a safety incident occurring in a construction project and within a novel discipline-level schedule, this study is subject to certain limitations.

One notable limitation is the lack of worker-level data that could lead to more accurate predictor features. For example, research has shown that the psychological status of workers on the site directly influences unsafe behaviors (Guo et al. 2017); however, this feature is difficult to measure or predict. Particularly as some types of worker-level data are typically subject to data protection laws. In future work, site-related data at the project level should be collected (in the form of incident days and incident-free days) in order to develop a more comprehensive model (Choi et al. 2020).

The example presented in this paper can be developed further by enhancing the dataset. The dataset could be expanded to include near-miss incidents, which would add an additional metric on safety performance (Shen and Marks 2015), provided that the incident reports specify the trade(s) involved in the incident and therefore are aligned with the data structure followed in the present study. In this manner, the prediction of near-miss incidents could be incorporated to build upon the present work.

DATA AVAILABILITY STATEMENT

Data used in this study were provided by a third party. Direct requests for these materials may be made to the provider indicated in the Acknowledgments.

ACKNOWLEDGMENTS

This project was supported by a Collaborative Research and Development Grant (CRDPJ 492657) from the Natural Sciences and Engineering Council of Canada. The authors would also like to thank PCL Industrial Management Inc. for their continued support, collaboration, in-depth knowledge, and provision of historical project data.

REFERENCES

American Fuel and Petrochemical Manufacturers – AFPM (2017). 2017 AFPM Occupational Injury and Illness Report. <https://www.afpm.org/sites/default/files/issue_resources/AFPM_Charts.pdf> (accessed 5 December 2022).Ahn, C. R., S. Lee, C. Sun, H. Jebelli, K. Yang, and B. Choi. 2019. "Wearable Sensing Technology Applications in Construction Safety and Health." *J. Constr. Eng. Manag.*, 145 (11): 03119007. American Society of Civil Engineers. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001708.

Ahn, S., L. Crouch, T. W. Kim, and R. Rameezdeen. 2020. "Comparison of Worker Safety Risks between Onsite and Offsite Construction Methods: A Site Management Perspective." *J. Comput. Civ. Eng.*, 146 (9): 05020010. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001890.

Al-Turaiki, I., M. Alshahrani, and T. Almutairi. 2016. “Building predictive models for MERS-CoV infections using data mining techniques.” *J. Infect. Public Health*, 9 (6): 744–748. https://doi.org/10.1016/j.jiph.2016.09.007.

Anvari, B., P. Angeloudis, and W. Y. Ochieng. 2016. “A multi-objective GA-based optimisation for holistic Manufacturing, transportation and Assembly of precast construction.” *Autom. Constr.*, 71 (Part 2): 226–241. Elsevier. https://doi.org/10.1016/J.AUTCON.2016.08.007.

AWCBC. 2019. *Overview The Association of Workers’ Compensation Boards of Canada produces this annual report, which provides national statistics on work-related Fatalities and Lost Time Claim compensation for injuries and diseases*.

Baker, H., M. R. Hallowell, and A. J. P. Tixier. 2020. “AI-based prediction of independent construction safety outcomes from universal attributes.” *Autom. Constr.*, 118: 103146. Elsevier B.V. https://doi.org/10.1016/j.autcon.2020.103146.

Bavafa, A., A. Mahdiyar, and A. K. Marsono. 2018. “Identifying and assessing the critical factors for effective implementation of safety programs in construction projects.” *Saf. Sci.*, 106 (March): 47–56. Elsevier. https://doi.org/10.1016/j.ssci.2018.02.025.

Boon, J., H. Yap, P. Eng, I. N. Chow, and K. Shavarebi. 2019. “Criticality of Construction Industry Problems in Developing Countries: Analyzing Malaysian Projects.” *J. Manag. Eng.*, 35 (5): 04019020–12. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000709.

Choe, S., and F. Leite. 2017. “Construction safety planning: Site-specific temporal and spatial information integration.” *Autom. Constr.*, 84: 335–344. Elsevier B.V. https://doi.org/10.1016/j.autcon.2017.09.007.

Choi, J., B. Gu, S. Chin, and J. S. Lee. 2020. "Machine learning predictive model based on national data for fatal accidents of construction workers." *Autom. Constr.*, 110: 102974. Elsevier. https://doi.org/10.1016/J.AUTCON.2019.102974.

Esmaeili, B., M. R. Hallowell, and B. Rajagopalan. 2015. "Attribute-Based Safety Risk Assessment. II: Predicting Safety Outcomes Using Generalized Linear Models." *J. Constr. Eng. Manag.*, 141 (8): 04015022. American Society of Civil Engineers (ASCE). https://doi.org/10.1061/(ASCE)CO.1943-7862.0000981.

Goh, Y. M., and D. Chua. 2013. "Neural network analysis of construction safety management systems: a case study in Singapore." *Constr. Manag. Econ.*, 31 (5): 460–470. https://doi.org/10.1080/01446193.2013.797095.

Guo, H., Y. Yu, T. Xiang, H. Li, and D. Zhang. 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." *Autom. Constr.*, 82: 207–217. Elsevier. https://doi.org/10.1016/J.AUTCON.2017.06.001.

Hallowell, M. R., S. Bhandari, and W. Alruqi. 2020. "Methods of safety prediction: analysis and integration of risk assessment, leading indicators, precursor analysis, and safety climate." *Constr. Manag. Econ.*, 38 (4): 308–321. Routledge. https://doi.org/10.1080/01446193.2019.1598566.

Hallowell, M. R., and J. A. Gambatese. 2009. "Activity-Based Safety Risk Quantification for Concrete Formwork Construction." *J. Constr. Eng. Manag.*, 135 (10): 990–998. American Society of Civil Engineers. https://doi.org/10.1061/(ASCE)CO.1943-7862.0000071.

Han, S., F. Saba, S. Lee, Y. Mohamed, and F. Pena-Mora. 2014. "Toward an understanding of the impact of production pressure on safety performance in construction operations.” *Accid. Anal. Prev.*, 68 (7): 106–116. https://doi.org/10.1016/j.aap.2013.10.007.

Haslam, R. A., S. A. Hide, A. G. F. Gibb, D. E. Gyi, T. Pavitt, S. Atkinson, and A. R. Duff. 2005. “Contributing factors in construction accidents.” *Appl. Ergon.*, 401–415. Elsevier Ltd.

Jiang, Z., D. Fang, and M. Zhang. 2015. “Understanding the causation of construction workers ’ unsafe behaviors based on system dynamics modeling.” *J. Manag. Eng.*, 31 (6): 10.1061/(ASCE)ME.1943-5479.0000350, 04014099. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000350.

Kang, K., and H. Ryu. 2019. “Predicting types of occupational accidents at construction sites in Korea using random forest model.” *Saf. Sci.*, 120: 226–236. Elsevier B.V. https://doi.org/10.1016/j.ssci.2019.06.034.

Karakhan, A. A., S. Rajendran, J. Gambatese, and C. Nnaji. 2018. “Measuring and Evaluating Safety Maturity of Construction Contractors: Multicriteria Decision-Making Approach.” *J. Constr. Eng. Manag.*, 144 (7): 1–13. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001503.

Koc, K., Ö. Ekmekcioğlu, and A. P. Gurgun. 2021. “Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers.” *Autom. Constr.*, 131: 103896. Elsevier. https://doi.org/10.1016/J.AUTCON.2021.103896.

Kursa, M. B., and W. R. Rudnicki. 2010. “Feature selection with the boruta package.” *J. Stat. Softw.*, 36 (11): 1–13. https://doi.org/10.18637/jss.v036.i11.

Lingard, H., M. Hallowell, R. Salas, and P. Pirzadeh. 2017. “Leading or lagging? Temporal analysis of safety indicators on a large infrastructure construction project.” *Saf. Sci.*, 91: 206–220. Elsevier. https://doi.org/10.1016/j.ssci.2016.08.020.

Mannering, F., C. R. Bhat, V. Shankar, and M. Abdel-Aty. 2020. “Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis.” *Anal. Methods Accid. Res.*, 25: 100113. Elsevier. https://doi.org/10.1016/J.AMAR.2020.100113.

Mierswa, I., and R. Klinkenberg. 2018. “RapidMiner Studio (9.1).”

Mitropoulos, P., T. S. Abdelhamid, and G. A. Howell. 2005. “Systems model of construction accident causation.” *J. Constr. Eng. Manag.*, 131 (7): 816–825.

Mohammadi, A., M. Tavakolan, and Y. Khosravi. 2018. “Factors influencing safety performance on construction projects: A review.” *Saf. Sci.*, 109 (2018): 382–397. https://doi.org/10.1016/j.ssci.2018.06.017.

Nasir, H., C. T. Haas, J. H. Rankin, A. R. Fayek, D. Forgues, and J. Ruwanpura. 2012. “Development and implementation of a benchmarking and metrics program for construction performance and productivity improvement 1 This paper is one of a selection of papers in this Special Issue on Construction Engineering and Management.” *Can. J. Civ. Eng.*, 39 (9): 957–967. https://doi.org/10.1139/l2012-030.

Neale, J., and A. Gurmu. 2021. “Production pressures in the building sector of the construction industry: a systematic review of literature.” *J. Eng. Des. Technol.* Emerald Publishing Limited. https://doi.org/10.1108/JEDT-12-2020-0529.

Pereira, E., M. Ali, L. Wu, and S. Abourizk. 2020. “Distributed Simulation-Based Analytics Approach for Enhancing Safety Management Systems in Industrial Construction.” *J. Constr. Eng. Manag.*, 146 (1): 1–12. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001732.

Pereira, E., S. Han, and S. AbouRizk. 2018a. “Integrating Case-Based Reasoning and Simulation Modeling for Testing Strategies to Control Safety Performance.” *J. Comput. Civ. Eng.*, 32 (6): 04018047. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000792.

Pereira, E., U. Hermann, S. Han, and S. AbouRizk. 2018b. “Case-Based Reasoning Approach for Assessing Safety Performance Using Safety-Related Measures.” *J. Constr. Eng. Manag.*, 144 (9): 04018088. American Society of Civil Engineers. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001546.

Poh, C. Q. X., C. U. Ubeynarayana, and Y. M. Goh. 2018. “Safety leading indicators for construction sites: A machine learning approach.” *Autom. Constr.*, 93: 375–386. Elsevier B.V. https://doi.org/10.1016/j.autcon.2018.03.022.

Rivas, T., M. Paz, J. E. Martín, J. M. Matías, J. F. García, and J. Taboada. 2011. “Explaining and predicting workplace accidents using data-mining techniques.” *Reliab. Eng. Syst. Saf.*, 96 (7): 739–747. Elsevier. https://doi.org/10.1016/J.RESS.2011.03.006.

Salas, R., and M. Hallowell. 2016. “Predictive validity of safety leading indicators: empirical assessment in the oil and gas sector.” *J. Constr. Eng. Manag.*, 142 (10): 10.1061/(ASCE)CO.1943-7862.0001167, 04016052. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001167.

Sarkar, S., S. Vinay, R. Raj, J. Maiti, and P. Mitra. 2019. “Application of optimized machine learning techniques for prediction of occupational accidents.” *Comput. Oper. Res.*, 106: 210–224. Elsevier Ltd. https://doi.org/10.1016/j.cor.2018.02.021.

Shen, X., and E. Marks. 2015. “Near-Miss Information Visualization Tool in BIM for Construction Safety.” *J. Constr. Eng. Manag.*, 142 (4): 04015100. American Society of Civil Engineers. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001100.

Taghaddos, M., H. Taghaddos, U. Hermann, Y. Mohamed, and S. AbouRizk. 2021. “Hybrid multi-mode simulation and optimization for subarea scheduling in heavy industrial construction.” *Autom. Constr.*, 125 (2021): 1–18. https://doi.org/10.1016/j.autcon.2021.103616.

Tixier, A. J. P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016a. “Application of machine learning to construction injury prediction.” *Autom. Constr.*, 69: 102–114. Elsevier B.V. https://doi.org/10.1016/j.autcon.2016.05.016.

Tixier, A. J. P., M. R. Hallowell, B. Rajagopalan, and D. Bowman. 2016b. “Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports.” *Autom. Constr.*, 62: 45–56. Elsevier. https://doi.org/10.1016/J.AUTCON.2015.11.001.

Versteeg, K., P. Bigelow, A. M. Dale, and A. Chaurasia. 2019. “Utilizing construction safety leading and lagging indicators to measure project safety performance: A case study.” *Saf. Sci.*, 120: 411–421. Elsevier. https://doi.org/10.1016/j.ssci.2019.06.035.

W Guo, B. H., and T. Wing Yiu. 2016. “Developing Leading Indicators to Monitor the Safety Conditions of Construction Projects.” *J. Manag. Eng.*, 32 (1): 04015016–14. https://doi.org/10.1061/(ASCE)ME.

Yi, K.-J., and D. Langford. 2006. “Scheduling-Based Risk Estimation and Safety Planning for Construction Projects.” *J. Constr. Eng. Manag.*, 132 (6): 626–635. https://doi.org/10.1061/(ASCE)0733-9364(2006)132:6(626).

You, Z., and C. Wu. 2019. “A framework for data-driven informatization of the construction company.” *Adv. Eng. Informatics*, 39: 269–277. Elsevier. https://doi.org/10.1016/J.AEI.2019.02.002.

Zhang, P., N. Li, Z. Jiang, D. Fang, and C. J. Anumba. 2018. “An agent-based modeling approach for understanding the effect of worker-management interactions on construction workers’ safety-related behaviors." https://doi.org/10.1016/j.autcon.2018.10.015.

Zhang, Y., and Y. Yang. 2015. "Cross-validation for selecting a model selection procedure." *J. Econom.*, 187 (1): 95–112. https://doi.org/10.1016/j.jeconom.2015.02.006.

List of Figure Captions

TABLE 1. Features collected from daily payroll information

feature name	Description
Proj_id*	Project identifier
WH_month	Monthly working hours
WH_cml	Monthly cumulative working hours
WH_cml-pct	Percentage of cumulative working hours
WH_diff	Increase/decrease in working hours compared to the previous month
Age_wrks-30-less	Percentage of workers aged 30 or less
Age_frmn-30-less	Percentage of foremen aged 30 or less
Age_wrks-50-more	Percentage of workers aged 50 or more
Age_frmn-50-more	Percentage of foremen aged 50 or more
WEx_wrks-new	Percentage of new workers compared to the previous month
WEx_wrks-3-less	Percentage of workers with up to 3 years of experience
DS_wrks	Workers' average number of days on the site
DS_frmn	Foremen's average number of days on the site
Crew-size	Crew size
Proj_s-curve-inc	Monthly S-Curve increase
Proj_pct-cplt	Percentage of project completion
Proj_ramp	Increase/decrease of workers on the project
S_incident	Monthly safety incidents occurring

***All features, except Proj_id, were collected by discipline**

TABLE 2. Number of collected data points by discipline

Trade discipline	Total data points	Incident data points
Ironworkers	146	51
Pipefitters	165	87
Civil	177	84
Operators	132	5
Electrical	135	28

TABLE 3. Selected predictor features

feature name	Description
WH_month	Monthly working hours
WEx_wrks-3-less	Percentage of workers with up to 3 years of experience
Age_frmn-30-less	Percentage of foremen aged 30 or less
Age_wrks-30-less	Percentage of workers aged 30 or less
WEx_wrks-new	Percentage of new workers compared to the previous month
Proj_s-curve-inc	Monthly S-Curve increase

Crew-size	Crew size
DS_frmn	Foremen average number of days on the site
Age_frmn-50-more	Percentage of foremen aged 50 or more
WH_cml-pct	Percentage of cumulative working hours

TABLE 4. Prediction performance measures

Model	Accuracy	Standard deviation	Incident recall
SVM	72.2	3.83	61.39
Naïve Bayes (Kernel)	70.47	3.65	75.25
Decision Tree	67.97	3.31	57.43
Naïve Bayes	62.90	3.89	74.26
Fast Large-Margin	62.90	3.89	74.26

Table 5. Original project planning - Scenario 1

Month #	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**2**	**3**	**3**	**3**	**3**	**3**	**4**	**4**	**4**	**4**	**4**	**5**	**5**	**5**	**5**	**5**	**6**	**6**	**6**	**6**	**6**
Discipline-Civil (Yes-1/No-0)	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
Discipline-Ironworkers (Yes-1/No-0)	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
Discipline-Pipefitters (Yes-1/No-0)	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
Discipline-Electrical (Yes-1/No-0)	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
Discipline-Operators (Yes-1/No-0)	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0
Age_frmn-30-less	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
DS_frmn	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Age_wrks-50-more	1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
Age_wrks-30-less	4	4	4	4	4	4	4	4	4	3	3	3	3	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2
NewWorkersRate	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3	3	3	1	1	1	1	1	1	1	1	1	1
Crew_Size	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	3	3	3	3	3	3	3	3	3
WEx_wrks-3-less	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	2	1	1	2	1	2	2	1
WH_month	2	1	1	2	3	1	1	1	3	3	2	3	3	3	3	3	4	3	3	3	3	4	3	3	3	3	4	3
Proj_s-curve_inc	1	1	1	1	2	1	1	1	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
WH_cml-pct	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1	2	2	2	2	2	2	2	2	2	2
Incident (Yes/No)	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**Yes**	**No**	**Yes**	**Yes**	**No**	**Yes**	**Yes**	**Yes**	**No**	**No**	**No**	**Yes**	**Yes**	**No**	**No**	**No**	**Yes**	**Yes**	**No**

Table 6. An adjusted project planning to reduce incidents - Scenario 2 (the colored cells are changed between scenarios)

Month #	**1**	**1**	**1**	**2**	**2**	**2**	**2**	**2**	**3**	**3**	**3**	**3**	**3**	**4**	**4**	**4**	**4**	**4**	**5**	**5**	**5**	**5**	**5**	**6**	**6**	**6**	**6**	**6**
Discipline-Civil (Yes-1/No-0)	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0
Discipline-Ironworkers (Yes-1/No-0)	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
Discipline-Pipefitters (Yes-1/No-0)	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
Discipline-Electrical (Yes-1/No-0)	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1
Discipline-Operators (Yes-1/No-0)	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0
Age_frmn-30-less	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3
DS_frmn	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Age_wrks-50-more	1	1	1	1	1	1	1	1	2	2	2	2	1	3	1	3	1	3	3	3	1	1	3	3	3	1	1	3
Age_wrks-30-less	4	4	4	4	4	4	4	4	4	3	3	3	3	3	3	2	2	2	2	2	2	2	2	2	2	2	2	2
NewWorkersRate	4	4	4	4	4	4	4	4	3	1	3	2	1	3	3	1	1	3	1	1	1	1	1	1	1	1	1	1
Crew_Size	4	4	4	4	4	4	4	4	4	4	4	1	3	4	4	1	1	4	3	3	1	1	3	3	3	1	1	3
WEx_wrks-3-less	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	1	2	1	1	1	1	2	1	1	1	1
WH_month	2	1	1	2	3	1	1	1	3	3	2	2	3	3	3	3	4	3	3	3	3	4	3	3	3	3	4	3
Proj_s-curve_inc	1	1	1	1	2	1	1	1	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
WH_cml-pct	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1	2	1	2	2	2	2	2	2	2	2	2	2
Incident (Yes/No)	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**	**No**

Figure 1

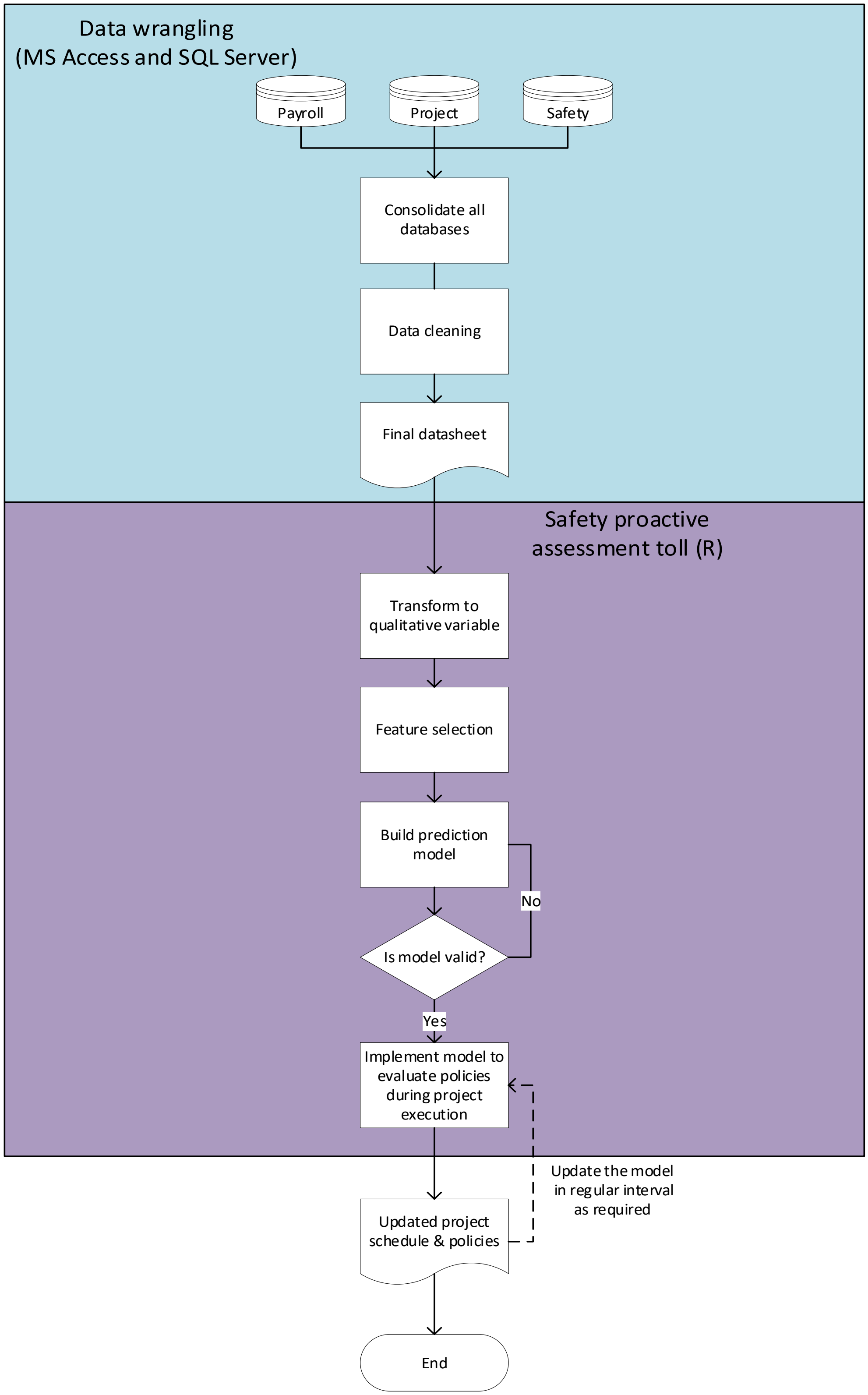
Data wrangling
(MS Access and SQL Server)
Payroll
Project
Safety
Consolidate all databases
Data cleaning
Final datasheet
Safety proactive assessment toll (R)
Transform to qualitative variable
Feature selection
Build prediction model
No
Is model valid?
Yes
Implement model to evaluate policies during project execution
Update the model in regular interval as required
Updated project schedule & policies
End

Figure 2

Click here to access/download;Figure;Figure 2.pdf