

REPUBLIC OF TURKEY
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

**DEVELOPMENT OF A MACHINE LEARNING PREDICTION
MODEL FOR CONSTRUCTION SAFETY MANAGEMENT**

Kerim KOÇ

DOCTOR OF PHILOSOPHY THESIS

Department of Civil Engineering

Program of Civil Engineering

Supervisor

Assoc. Prof. Dr. Aslı Pelin GÜRGÜN

October, 2022

REPUBLIC OF TURKEY

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

**DEVELOPMENT OF A MACHINE LEARNING PREDICTION
MODEL FOR CONSTRUCTION SAFETY MANAGEMENT**

A thesis submitted by Kerim KOÇ in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY is approved by the committee on 26.10.2022 in Department of Civil Engineering, Civil Engineering Program.

Assoc. Prof. Dr. Aslı Pelin GÜRGÜN
Yildiz Technical University
Supervisor

Approved By the Examining Committee

Assoc. Prof. Dr. Aslı Pelin GÜRGÜN, Supervisor
Yildiz Technical University

Assoc. Prof. Dr. Zeynep IŞIK, Member
Yildiz Technical University

Prof. Dr. Gül POLAT TATAR, Member
Istanbul Technical University

Assoc. Prof. Dr. Cenk BUDAYAN, Member
Middle East Technical University Northern Cyprus Campus

Assoc. Prof. Dr. Atilla DAMCI, Member
Istanbul Technical University

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of Development of a Machine Learning Prediction Model for Construction Safety Management supervised by my supervisor, Assoc. Prof. Dr. Aslı Pelin GÜRGÜN. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Kerim KOÇ

Dedicated to my family



ACKNOWLEDGEMENTS

First and foremost, I offer my sincerest gratitude to my supervisor. Assoc. Prof. Dr. Aslı Pelin GÜRGÜN. This dissertation would not have been a real fulfillment without the backing and cooperation of her who has supported me throughout my thesis with her patience and knowledge. I would also like to thank my colleague, companion, and friend; Dr. Ömer EKMEKCİOĞLU, who was always with me in dealing with the hardship of academia.

Didem KOÇ, my wife, there are no words to express my gratitude to her. Being happy with you was probably the most effective means to complete this thesis. Further, I would like to show my gratitude to my parents, who were with me throughout this process. Finally, I would also like to thank the Social Security Institution for supporting me and providing the dataset used throughout the thesis.

Kerim KOÇ

TABLE OF CONTENTS

LIST OF SYMBOLS	vii
LIST OF ABBREVIATIONS	ix
LIST OF FIGURES	xii
LIST OF TABLES	xiv
ABSTRACT	xv
ÖZET	xviii
1 INTRODUCTION	1
1.1 Literature Review.....	1
1.2 Objective of the Thesis	4
1.3 Hypothesis.....	5
2 LITERATURE REVIEW	7
2.1 Construction Safety Management in Turkey	7
2.2 Machine Learning Applications in Construction Safety	8
3 INTEGRATING FEATURE ENGINEERING, GENETIC ALGORITHM AND TREE-BASED MACHINE LEARNING METHODS TO PREDICT THE POST- ACCIDENT DISABILITY STATUS OF CONSTRUCTION WORKERS	10
3.1 Introduction.....	10
3.2 Background	12
3.3 Research Methodology	22
3.4 Results and Discussion.....	32
3.5 Conclusions and Future Work.....	44
4 ACCIDENT PREDICTION IN CONSTRUCTION USING HYBRID WAVELET- MACHINE LEARNING	46
4.1 Introduction.....	46
4.2 Research Objective and Novelty	47
4.3 Background of Used Methods and Related Implementations in Construction Safety Literature.....	48
4.4 Research Framework.....	56
4.5 Materials and Methods.....	58
4.6 Results and Discussion.....	63
4.7 Conclusions.....	76
5 SCENARIO-BASED AUTOMATED DATA PREPROCESSING TO PREDICT SEVERITY OF CONSTRUCTION ACCIDENTS	79
5.1 Introduction.....	79

5.2 Literature Review	82
5.3 Data Description	86
5.4 Model Development.....	92
5.5 Results.....	101
5.6 Discussion	109
5.7 Conclusion	115
6 PREDICTION OF CONSTRUCTION ACCIDENT OUTCOMES BASED ON IMBALANCED DATASET THROUGH INTEGRATED RESAMPLING TECHNIQUES AND MACHINE LEARNING METHODS	118
6.1 Introduction.....	118
6.2 Background	120
6.3 Research Methodology	127
6.4 Results and Discussion.....	135
6.5 Implications for Safety Authorities and Practical Use of the Study	145
6.6 Conclusion and Future work	151
7 RESULTS AND DISCUSSION	154
REFERENCES	162
PUBLICATIONS FROM THE THESIS	184

LIST OF SYMBOLS

a	Missing Data Handling Step Score
A	Dataset after Missing Data Handling
A_r	Recorded Number of Accidents
A_p	Predicted Number of Accidents
$\overline{A_r}$	Average of Observed Number of Accidents
$\overline{A_p}$	Average of Predicted Number of Accidents
b	Data Binning Step Score
B	Dataset After Data Binning
c	Outlier Handling Step Score
C	Dataset After Outlier Handling
$C_{i,j}$	The Number of Observations Observed in Class i Predicted to Be Class j in Confusion Matrix C
d	Data Scaling Step Score
D	Dataset After Data Scaling
e	Data Encoding Step Score
E	Dataset After Data Encoding
f	Data Resampling Step Score
F	Dataset After Data Resampling
J_0	Maximum Number of Levels
k	Number of Hours in the day shift
K	Number of Workers
LL	Lower Limit of Numerical Features
m	Mean
n	Number of Days
N	Number of Data Samples
N_i	The number of Cause Material
N_{Month}	Number of Instance in the Testing Set to Run the Model Monthly
N_{Trial}	Number of Instance in the Testing Set to Run the Model Daily
$Q1$	25 th Percentile
$Q3$	75 th Percentile
R^2	Determination Coefficient
s	Standard Deviation

T_i	The Rork-hour Time for Worker i
UL	Upper Limit of Numerical Features
X	Dataset After Preprocessing
z	Number of Principal Component



LIST OF ABBREVIATIONS

ACF	AutoCorrelation Functions
AdaBoost	Adaptive Boosting
ANN	Artificial Neural Network
ARIMA	Autoregressive Integrated Moving Average
AUC	Area Under Curve
AUROC	Area Under the Receiver Operating Characteristic
BFGS	Broyden- Fletcher-Goldfarb-Shanno
BL	Border-Line
CART	Classification And Regression Trees
C-BiLSTM	Convolutional Bidirectional Long Short-Term Memory
CCUS	Case Control Under-sampling
CHAID	Chi-squared Automatic Interaction Detection
COS	Cluster-based Over-sampling
CUS	Cluster-based Under-sampling
CV	Cross-validation
CWT	Continuous Wavelet Transform
DBIN	Data Binning
DENC	Data Encoding
DNN	Deep Neural Network
DRSM	Data Resampling
DSCL	Data Scaling
DT	Decision Tree
DWT	Discrete Wavelet Transform
EEG	Electroencephalography
ENN	Edited Nearest Neighbors
ET	Extra Trees
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FT	Fourier Transform
GA	Genetic Algorithm
GBDT	Gradient Boosting Decision Trees

GBM	Gradient Boosting Model
HAN	Hierarchical Attention Network
IQR	Interquartile Range
KM	K-means
KMA	Korea Meteorological Agency
KNN	K Nearest Neighbor
KOSHA	Korea Occupational Safety and Health Agency
LR	Logistic Regression
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MARS	Multi-variate Adaptive Regression Splines
MDH	Missing Data Handling
ML	Machine Learning
MLP	Multi-layer Perceptron
MOEL	Ministry of Employment and Labor
MRA	Multiple Regression Analysis
MW-MOTE	Majority Weighted Minority Over-Sampling Technique
NB	Naive Bayes
N-PD	Not permanently Disabled
NSE	Nash-Sutcliffe Efficiency
OHS	Occupational Health and Safety
OHSA	Occupational Health and Safety Administration
OUTH	Outlier Handling
OVR	One-vs-Rest
PACF	Partial AutoCorrelation Functions
PC	Principal Component
PCA	Principal Component Analysis
PD	Permanently Disabled
PPE	Personal Protective Equipment
RAW	Imbalance data
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root Mean Squared Error
ROC	Receiver Operating Characteristic
ROS	Random Over-sampling

RPSS	Rank Probability Skill Score
RUS	Random Under-sampling
SGTB	Stochastic Gradient Tree Boosting
SME	Small and Medium-sized Enterprises
SMOTE	Synthetic Minority Over-sampling Technique
SQP	Sequential Quadratic Programming
SSI	Social Security Institution
SVC	Support Vector Classifier
SVM	Support Vector Machine
SVR	Support Vector Regression
TF-IDF	Term Frequency – Inverse Document Frequency
TN	True Negative
TP	True Positive
TPR	True Positive Rate
TSI	Turkish Statistical Institute
W-ANN	Wavelet - Artificial Neural Network
W-MARS	Wavelet - Multi-variate Adaptive Regression Splines
W-SVR	Wavelet - Support Vector Regression
WT	Wavelet Transform
XGBoost	EXtreme Gradient Boosting

LIST OF FIGURES

Figure 3.1 Working principles of tree-based ensemble machine learning methods	18
Figure 3.2 Research flow	24
Figure 3.3 Data encoding procedure	28
Figure 3.4 Principal component selection procedure.....	33
Figure 3.5 Trials of principal components for each classifier	34
Figure 3.6 Heat map of PCA.....	34
Figure 3.7 Convergence graph of genetic algorithm for (a) Random Forest, (b) XGBoost, (c) AdaBoost, (d) Extra Trees.	35
Figure 3.8 Confusion matrix for ML models	38
Figure 3.9 ROC curve of ensemble models	38
Figure 3.10 Guide for practitioners on model implementation.....	39
Figure 4.1 Analysis procedures of the proposed study	57
Figure 4.2 The time series of the daily occupational accidents	60
Figure 4.3 Illustration of lag and lead times	60
Figure 4.4 The partial autocorrelation functions.....	63
Figure 4.5 The scatter plots of the stand-alone models obtained for the test set (grey continuous lines represent 1:1 perfect lines and red dashed lines represent the trend lines).	65
Figure 4.6 Spectral bands decomposed by discrete wavelet transform	67
Figure 4.7 The scatter plots of the hybrid wavelet models obtained for the test set (grey continuous lines represent 1:1 perfect lines and red dashed lines represent the trend lines)	68
Figure 4.8 Recorded and best modelled time series with respect to the testing set. (a) T+1, (b) T+7, (c) T+30	70
Figure 4.9 Taylor Diagrams. (a) T+1, (b) T+7, (c) T+30	70
Figure 4.10 Dynamic utilization of the proposed framework.....	71
Figure 4.11 The role of inspection boards and sector-basis investigation.....	72
Figure 4.12 Proactive risk management and dynamic utilization of the study.....	73
Figure 4.13 Example for continuous monitoring.....	75
Figure 4.14 Answers to 5W1H	76
Figure 5.1 Flow of the research	82
Figure 5.2 Features used in the prediction model	87
Figure 5.3 Cause material with accident severity level	91
Figure 5.4 Construction type with accident severity level.....	91

Figure 5.5 Education level with accident severity level	92
Figure 5.6 Gender with accident severity level	92
Figure 5.7 Research steps	93
Figure 5.8 Scenario development process.	95
Figure 5.9 Illustration of preprocessing process	99
Figure 5.10 Confusion matrices of best scenarios	104
Figure 5.11 Performances of preprocessing scenarios in terms of (a) f1-score, (b) precision, (c) recall	105
Figure 5.12 Confusion matrices of the best recall scenarios based on output labels...	107
Figure 5.13 Boxplot of labels based on recall values	108
Figure 5.14 Permutation importance of features.....	109
Figure 5.15 Example of a practical use of the research	114
Figure 6.1 Flowchart of the study	128
Figure 6.2 ROC curve of a. Raw models b. RUS-based models, c. ROS-based models, d. SMOTE-based models.	137
Figure 6.3 Distribution of number of past accidents in the company versus number of workers in the company based on a. raw data (RAW), b. resampled (RUS), c. resampled (ROS), d. resampled (SMOTE).....	145
Figure 6.4 Practical use of the research	148
Figure 6.5 Example of a utilization plan.....	149

LIST OF TABLES

Table 3.1 Selected studies of occupational accident prediction in the construction industry.	13
Table 3.2 Description and distribution of attributes.....	26
Table 3.3 Possible cases between actual and predicted conditions.....	31
Table 3.4 Comparative results of machine learning methods	36
Table 3.5 Summary of evaluated test scores	37
Table 4.1 Overview of the used machine learning algorithms.....	52
Table 4.2 The statistical features of the time-series data	59
Table 4.3 The equations of performance criteria	62
Table 4.4 Performance comparison of stand-alone models	65
Table 4.5 Determination of the optimum decomposition level (for 1-day lead time) ...	66
Table 4.6 Performance comparison of hybrid wavelet models.....	68
Table 5.1 Some preprocessing elements and their descriptions.....	85
Table 5.2 Number of accidents by severity level with respect to numerical features....	88
Table 5.3 Number of accidents by severity level with respect to time-related features	90
Table 5.4 Initial preprocessing steps before comprehensive preprocessing	94
Table 5.5 Number of output labels.....	94
Table 5.6 Performances of first, best and last scenarios	102
Table 6.1 Studies addressing class imbalance issue in construction safety literature..	126
Table 6.2 Imbalanced dataset of construction accidents.....	129
Table 6.3 Description of attributes.....	130
Table 6.4 Summary of evaluated scores	138
Table 6.5 Alternative test/train splitting scenarios.....	139
Table 6.6 Attribute ranks based on permutation importance	141
Table 6.7 Respondents' opinions about the proposed model.....	150

Development of a Machine Learning Prediction Model for Construction Safety Management

Kerim KOÇ

Department of Civil Engineering

Doctor of Philosophy Thesis

Supervisor: Assoc. Prof. Dr. Aslı Pelin GÜRGÜN

Central to the entire discipline of construction safety management is the concept of construction accidents. Occupational accident rates in construction projects are usually higher than other industries in most countries, even though safety management systems are continuously improving. Hence, the preliminary studies exploring the consequences of occupational accidents have received considerable attention in research society. With the advance in data collection and analysis techniques, data-driven models have been proposed in the construction safety management research domain to predict potential outcomes of occupational accidents. In this context, machine learning (ML) methods can be used to predict various output variables to improve safety management applications in construction projects as part of data-driven model. Thus, the essential purpose of this study is to develop a data-driven ML model to predict several output variables with the aim of preventing occupational accidents, minimizing their severe consequences, and improving the conditions of workers/companies regarding occupational health and safety (OHS) management. To achieve the study objective, a dataset that includes nearly 400,000 occupational accidents experienced in the construction industry in Turkey was taken from the Social Security Institution, which enabled to develop a robust prediction model and analyze occupational accidents and their precursors statistically. Based on the comprehensive literature review, several research gaps have been identified and

addressed in this study. Majority of the accident outcome prediction studies developed their models based on dataset captured in developed countries, while this study, for the first time, used a huge national dataset of Turkey. Based on the careful examination of the dataset and identified research gaps, four prediction models were developed.

First, a study that integrates metaheuristics optimization algorithms and ML methods has not been found in the literature. Besides, despite the superiority of tree-based ML methods was well documented in the construction safety management research field, no previous study compared the performance of different tree-based ML methods. Moreover, even though preprocessing is the foundation of ML models, previous studies mainly ignored the role of comprehensive feature engineering. Finally, although several accident outcome prediction studies were performed in the literature, post-accident permanent disability status of construction workers has not been acknowledged in the literature. Therefore, feature engineering, tree-based ML methods and genetic algorithm were integrated in the first prediction model to predict post-accident permanent disability status of construction workers.

The second prediction model was developed to contribute to the body of construction safety management by coupling wavelet transform and different ML methods to predict the number of occupational accidents using time series data. The motivation of developing this model was that time series analysis of occupational accidents has not been employed previously. Therefore, based on trend of past accidents, the number of future accidents was predicted with high prediction accuracies for short-term, mid-term and long-term time periods, i.e., 1-day, 7-day, and 30-day ahead, respectively. Accordingly, a dynamic utilization plan was proposed to the field of safety management by introducing a new theoretical and practical framework.

Literature survey also showed that existing studies that uses ML methods in the construction safety management research field are limited and usually focus on selecting the most appropriate ML method rather than identifying the most effective preprocessing pipeline before the prediction. However, majority of time required for developing ML models is devoted to pre-processing. Therefore, a scenario-basis automated preprocessing model that identifies the best pre-processing pipeline is developed to predict the severity of construction accidents in the third prediction model of the study.

Finally, the literature on accident outcome prediction is limited in terms of dealing with imbalanced dataset through integrated resampling techniques and ML methods in the construction safety domain. Besides, previous studies on ML-based safety management in Turkey did not include accident outcome prediction regarding fatal and non-fatal accidents. Therefore, the last model includes prediction of accident outcomes (fatal vs non-fatal) by incorporating ML methods and resampling techniques.

Overall, this study proposes a data-driven safety management platform and examines its applicability with four prediction rationale. With the findings emerged according to the analysis results, it is highlighted that data-driven models, which have become easily adaptable to the existing project management applications in the era of data, can save the lives of construction workers. Several utilization plans are proposed to help industry professionals adopt the proposed models in construction sites to predict future incidents. It is suggested to develop a national data-driven safety management model, as the utilized dataset was acquired from national institution, who mandatorily require these data from companies after the occurrence of accidents. The proposed framework can be used in construction sites to detect several output variables along with the four models examined in this study, which can be a valuable decision-making input for safety professionals to reduce the unexpected consequences of accidents. With the incorporation of this thesis into the construction industry, some of the occupational accidents can be prevented and the lives of some of the construction workers can be saved.

Keywords: Machine learning, Artificial intelligence, Construction Safety, Occupational health and safety (OHS).

İnşaat İş Güvenliği Yönetimi için Makine Öğrenmesine Dayalı Tahmin Modeli Geliştirilmesi

Kerim KOÇ

İnşaat Mühendisliği Anabilim Dalı

Doktora Tezi

Danışman: Doç. Dr. Aslı Pelin GÜRGÜN

İnşaat güvenliği yönetiminin merkezinde inşaat kazaları kavramı yer almaktadır. İnşaat projelerinde iş kazası oranları, güvenlik yönetim sistemlerindeki sürekli iyileştirmelere rağmen çoğu ülkede diğer sektörlerden daha yüksektir. Bu nedenle, iş kazalarının sonuçlarını araştıran çalışmalar son yıllarda araştırmacılar tarafından büyük ilgi görmektedir. Veri toplama ve analiz tekniklerindeki ilerlemeyle birlikte, iş kazalarının olası sonuçlarını tahmin etmek için inşaat güvenliği yönetimi araştırma alanında veriye dayalı modeller önerilebilmektedir. Bu bağlamda, veriye dayalı güvenlik yönetimi uygulamalarını iyileştirmek için çeşitli çıktı değişkenlerini tahmin etmek amacıyla makine öğrenmesi (MÖ) yöntemleri kullanılabilir. Bu çalışmanın temel amacı, iş kazalarını önlemek, ciddi sonuçlarını en aza indirmek ve çalışanların/şirketlerin iş sağlığı ve güvenliği (İSG) ile ilgili koşullarını iyileştirmek amacıyla çeşitli çıktı değişkenlerini tahmin etmek için veriye dayalı bir MÖ platformu geliştirmektir. Bu çalışmadaki amaçlara ulaşabilmek için Sosyal Güvenlik Kurumu'ndan Türkiye'de inşaat sektöründe gerçekleşen yaklaşık 400.000 iş kazasını içeren bir veri seti elde edilmiş ve buna bağlı olarak tahmin modeli geliştirilerek kazalar ve bunlara etki eden faktörler istatistiksel olarak analiz edilmiştir. Kapsamlı literatür taraması yapılarak çeşitli araştırma boşlukları tespit edilmiş ve bunlar titizlikle ele alınmıştır. Kaza sonucu tahmin çalışmalarının çoğu

modellerini gelişmiş ülkelerde toplanan veri setlerine dayalı olarak geliştirirken, bu çalışmada ilk kez Türkiye’de toplanan oldukça büyük bir veri seti kullanılmıştır. Veri seti dikkatli bir biçimde incelenmiş ve literatür taraması ile belirlenen araştırma boşluklarına da dayanarak dört adet tahmin modeli geliştirilmiştir.

İlk olarak, inşaatta İSG literatüründe sezgisel optimizasyon algoritmaları ile MÖ yöntemlerini entegre eden bir çalışmaya rastlanmamıştır. Ayrıca, ağaç tabanlı makine öğrenmesi yöntemlerinin inşaat güvenliği yönetimi alanında oldukça iyi performans gösterdiği diğer birçok araştırmacı tarafından belirlenmiş olmasına rağmen ağaç tabanlı makine öğrenmesi yöntemlerinin performansını karşılaştıran hiçbir çalışmaya da rastlanmamıştır. Bununla birlikte, makine öğrenmesi modellerinin temelinde veri ön işleme yatmasına rağmen, önceki çalışmalar kapsamlı bir özellik mühendisliğinin rolünü göz ardı etmiştir. Son olarak, literatürde birçok kaza sonucu tahmin çalışması olmasına rağmen, inşaat çalışanlarının kaza sonrası kalıcı iş görmezlik durumları literatürde yer almamaktadır. Bu nedenle, ilk tahmin modelinde özellik mühendisliği, ağaç tabanlı MÖ yöntemleri ve genetik algoritma entegre edilerek inşaat çalışanlarının iş kazası sonrası olası iş görmezlik durumları araştırılmıştır.

İkinci tahmin modeli, zaman serisi verilerini kullanarak iş kazalarının sayısını tahmin etmek için dalgacık dönüşümü ve çeşitli MÖ yöntemlerini birleştirerek inşaat güvenliği yönetimine katkıda bulunmak için geliştirilmiştir. Bu modelin geliştirilmesinin sebebi, iş kazalarının zaman serisi analizlerinin daha önce yapılmamış olmasıdır. Bu nedenle, geçmiş kaza sayılarına dayalı olarak, ulusal ölçekte gelecek iş kazası sayıları kısa vadeli, orta vadeli ve uzun vadeli zaman dilimleri için, yani sırasıyla 1 gün, 7 gün ve 30 gün ilerisi için yüksek doğruluk oranlarıyla tahmin edilmiştir. Buna göre, daha önce literatürde hiç odaklanılmamış özgün bir teorik ve pratik çerçeve geliştirilerek İSG yönetimi için dinamik bir kullanım planı önerilmiştir.

Literatür araştırması ayrıca, inşaat güvenliği yönetimi araştırma alanında MÖ yöntemini kullanan mevcut çalışmaların sınırlı olduğunu ve bu çalışmaların genellikle tahmin yapılmadan önce en etkili veri ön işleme adımlarını belirlemek yerine daha çok en uygun MÖ yöntemini seçmeye odaklandığını göstermiştir. Ancak, MÖ modellerini geliştirmek için gereken zamanın çoğu genellikle veri ön işleme ayrılmaktadır. Bu nedenle, bu çalışmanın üçüncü modelinde inşaat kazalarının şiddetini tahmin etmek için en iyi senaryoyu belirleyen senaryo bazlı otomatik bir veri ön işleme modeli geliştirilmiştir.

Son olarak, kaza sonucu tahminleri ile ilgili inşaat güvenliği yönetimi alanındaki literatürde, yeniden örnekleme teknikleri ve MÖ yöntemlerinin entegre edilerek dengesiz veri kümesiyle başa çıkmaya işaret eden herhangi bir çalışmaya rastlanmamıştır. Ayrıca, Türkiye'de MÖ tabanlı güvenlik yönetimi ile ilgili daha önce gerçekleştirilen çalışmalar, ölümlü ve yaralanmalı kazalara ilişkin kaza sonucu tahminini içermemektedir. Bu nedenle, bu çalışmada geliştirilen son tahmin modelinde, MÖ yöntemlerini ve yeniden örnekleme tekniklerini entegre ederek kaza sonuçlarının (ölümlü ve yaralanmalı) tahminine odaklanılmıştır.

Genel olarak, bu çalışma veriye dayalı bir İSG yönetim platformu önermekte ve bu platformun uygulanabilirliğini dört tahmin modeliyle test etmektedir. Bu çalışmada ayrıca, analiz sonuçlarına göre ortaya çıkan bulgulara bağlı ve günümüz veri çağında mevcut proje yönetimi uygulamalarına kolayca uyarlanabilir bir model geliştirerek, inşaat sektörü çalışanlarının hayatlarının kurtarılabileceği vurgulanmaktadır. Sektör profesyonellerinin gelecekteki potansiyel iş kazalarını ve/veya bunların etkilerini tahmin etmek için ve bu modelleri uygulamalarına yardımcı olmak için çeşitli yol haritaları da bu çalışma kapsamında önerilmiştir. Kullanılan veriler, kazaların meydana gelmesinden sonra ilgili kurumlarca zorunlu olarak talep edildiği için, ulusal bir veriye dayalı güvenlik yönetimi modelinin geliştirilmesi ve buna ön ayak olması bu çalışmanın en büyük katkısıdır. Önerilen platform, kazaların beklenmedik ve istenmeyen sonuçlarını azaltma noktasında İSG uzmanları için değerli bir karar verme girdisi olabilecek ve önerilen dört modelle birlikte çeşitli çıktı değişkenlerini tespit etmek için şantiyelerde kullanılabilecektir. Bu çalışmanın sektöre entegrasyonu ile birlikte iş kazalarının bazıları engellenebilecek ve bazı çalışanların hayatları kurtarılabilecektir.

Anahtar Kelimeler: Makine öğrenmesi, Yapay zekâ, İnşaat güvenliği, İş sağlığı ve güvenliği (İSG).

1.1 Literature Review

Construction industry is envisaged to be among the most demanding and dangerous industries [1]. Therefore, it is well documented that construction workers have higher probabilities to experience severe accidents compared to workers in other sectors [2]. Construction sector accounts for nearly 35% of occupational fatalities globally compared to other sectors [3]. Even though several improvements have been performed in safety management practices in the industry, still, there is a need for growing interest given by both practitioners and researchers. Thus, central to the entire discipline of construction safety management becomes the concept of construction accidents.

Goh and Ubeynarayana [4] states that past construction accidents can provide a significant information to estimate future accidents and inevitably reduce their severe consequences. Based on the collected data from various institutions, machine learning (ML) methods can be used effectively to improve construction safety management applications within the context of smart construction and safety management. A comprehensive literature survey revealed that there are a variety of output variables that can be predicted to prevent their future re-occurrences. Some of the pivotal areas are classification of construction accident type [5], accident outcome [1], severity of accidents [6], body part injured [7], type of work [8], and energy type [7].

From the methodological point of view, several ML methods were implemented in the construction safety management such as support vector machine (SVM) [9], adaptive boosting (AdaBoost) [2], naïve bayes (NB) [10], and random forest (RF) [11], artificial neural network (ANN) [12], and deep neural network (DNN) [13]. In order to calibrate the optimum parameter configurations of the adopted ML methods, many researchers used trial-error [6] or grid search [14] algorithms. Similarly, some of the preprocessing elements such as data resampling (i.e., RUS [11]), data binning (i.e., age of the worker [15]), or outlier handling [4] were employed to prepare the data before the predictions. These algorithms were subjected to several data sources such as national data [2], data collected from construction site [16], or accident narratives subjected to natural language

processing [5]. These datasets were collected from a variety of countries such as the U.S. [5], Korea [2] and Turkey [16].

However, comprehensive literature review shows that there are significant research gaps in the relevant literature, which are addressed in this thesis. First of all, majority of research on accident outcome prediction relied on datasets taken from construction companies or accident reports. Only a limited number of studies used past accident records taken from national institutions. Similarly, majority of the used data were from developed countries, while this study statistically investigated national occupational accident dataset of Turkey.

On the other hand, the literature on accident outcome classification disregarded metaheuristic optimization algorithms despite their effectiveness in finding optimum hyperparameters. In one of the models developed in this study, GA was incorporated in the permanent disability prediction, and accordingly parameters of four tree-based ensemble ML models were configured. Likewise, despite addressing some of the preprocessing steps, none of the previous studies in the pertinent literature adopted a systematic feature engineering (or preprocessing) approach, while this was also considered in the permanent disability prediction model. In this vein, one-hot encoding was subjected to nominal features while label encoding was subjected to numerical features and all the features were scaled with the standard scaler method. Then, since the dimension of the dataset was increased significantly, there exist a significant computational burden, especially considering the GA method application. Therefore, principal component analysis (PCA) was employed to reduce the dimension of the dataset, while ensuring the statistical significance in the remaining information.

During the development of permanent disability model, it was observed that selecting the most appropriate preprocessing pipeline is probably as significant as selecting the most appropriate ML method. In fact, majority of time to develop prediction model is generally devoted to preprocessing, not the ML model setup. Existing literature, neither on construction safety management nor any other research domain, however, did not indicate any certain model that selects the most powerful preprocessing pipeline. Therefore, to fill the gap in the literature, we further developed a scenario-based automated preprocessing model to predict the severity of construction accidents.

The dataset used in this study was severely imbalanced regarding fatal accidents and non-fatal accidents. The literature highlights the role of resampling techniques in imbalanced ML applications. Therefore, in one of the models developed within the scope of this thesis, we examined the impact of several data resampling methods in the accuracy of accident outcome prediction. Class imbalance is particularly important in construction safety management research domain as minority class usually indicates more important outcomes such as fatality compared to non-fatal accidents or accidents compared to non-accident cases. Since detecting minority class is of importance in practice, this model is expected to provide practitioners with significant basis for developing proactive safety management platform.

It was also realized that time series analysis has not been performed previously despite huge national datasets. Since, the collected data comprises accident instances throughout nearly a decade, the number of accidents in each day can be acquired and statistically investigated. Then, with the concept of time-series analysis, the number of future accidents can be predicted based on past accident trends. In this vein, for the first time in the construction safety literature, one of the developed models within the scope of this thesis integrated ML methods with wavelet transform to predict the number of accidents for one-day, one-week, and one-month ahead.

Overall, several research gaps were identified based on a comprehensive literature survey and various prediction schemes were developed in this thesis. The common objective of all the proposed models is to develop a prediction model based on national occupational accident dataset in order to present a proactive and data-driven construction safety management platform. In the era of industry 4.0, the importance of data-driven platforms cannot be ignored, especially in the construction safety management research field. Because the application of data-driven models can save hundreds of lives if effectively implemented in construction projects. Here, the proposed models can be used in construction projects as the included features are already required by the Social Security Institution (SSI). The findings of this thesis are also expected to help researchers develop more robust ML models not only in the construction safety management field, but also other sub-fields of construction management.

1.2 Objective of the Thesis

The aim of this thesis is to develop machine learning prediction models using national dataset acquired from the Social Security Institution. The main motivation to perform this study was the Occupational Health and Safety Law (Law no: 6331) [17] enacted in 2012. The law states that employers must not only record occupational accidents but also report these accidents to the SSI within three days following the accident. Therefore, despite there were only 425 recorded occupational accidents in the Turkish construction industry between 1987 and 2011, there were nearly 400,000 recorded accidents between 2012 and 2020 (as of October, with NACE codes of 41, 42, and 43). Accordingly, several researchers used this dataset to observe the precursors of occupational accidents experienced in Turkey. Comprehensive literature survey showed that one of the promising research directions to improve construction safety management applications is to develop data-driven prediction models. Therefore, national dataset collected from the SSI was diligently investigated to develop effective machine learning models, address identified gaps, and show the applicability of using nation-wide dataset to prevent severe accidents and/or their consequences. Overall aim of the thesis was to develop a national prediction framework that detects several undesired outcomes prior to their occurrence to put them in use for effective data-driven safety management applications. Specific objectives of the proposed models can be summarized as follows:

- Model 1 aims to develop a comprehensive framework to predict the post-accident disability status of construction workers. Comprehensive feature engineering methodology, genetic algorithm (GA), and comparison of tree-based ML methods are initially employed, which constitute the methodological novelty of the study.
- Model 2 aims to contribute to the body of construction safety management by coupling discrete wavelet transform (DWT) and different ML methods to predict the number of occupational accidents using time series data. Methodological contribution of the study is that DWT and ML methods were initially integrated in the construction safety management literature.
- Model 3 aims to develop a scenario-basis automated preprocessing model that identifies the best scenario to predict the severity of construction accidents. Main motivation to conduct this study is that existing studies usually focus on selecting

the most appropriate ML method rather than identifying the most effective preprocessing pipeline before the prediction. Hence, a novel pre-processing model is proposed in this study.

- Model 4 aims to predict occupational accident outcomes using ML methods coupled with several resampling strategies. The main motivation to develop this prediction scheme was that there is a significant imbalance between number of fatal and non-fatal accidents, and selection of data resampling technique impacts the prediction rationale significantly.

Overall, this thesis is expected to contribute to society through several models to prevent high severe accidents in construction projects. The proposed solutions can also contribute to theory by filling several gaps in the literature. With this thesis, the importance and effectiveness of data-driven safety management applications are acknowledged.

1.3 Hypothesis

Main hypothesis is that data-driven safety management framework can be generated based on national occupational accident dataset, aiding to prevent some of the severe accidents. Correspondingly, sub-hypotheses of this study can be summarized as follows:

- Hypothesis 1: National occupational accident datasets can be used in prediction models to detect and prevent high severe accidents with high accuracy.
- Hypothesis 2: Occupational accident outcomes are influenced by all worker-based, time-based, and company-based attributes. Therefore, necessary mitigation measures should be developed based on the characteristics of workers and dynamics of companies, and particular constraints of time periods.
- Hypothesis 3: Integrating feature engineering, genetic algorithm, and tree-based ML methods can present a useful means to detect post-accident permanent disability status of construction workers.
- Hypothesis 4: By using historical time series data, number of future accidents can be predicted with high accuracies for short-, medium-, and long-term time horizons.
- Hypothesis 5: Hybridizing discrete wavelet transform (DWT) with ML methods can improve prediction performance of stand-alone ML models in detecting number of future nation-wide occupational accidents in the construction sector.

- Hypothesis 6: Comprehensive scenario-based preprocessing model may help improve prediction performance of ML methods significantly.
- Hypothesis 7: Preprocessing pipeline of: (i) Not removing missing data, (ii) Not applying data binning, (iii) Considering outliers, (iv) Applying Min-Max-Scaler (v) Applying one-hot encoding, and (vi) data resampling with random oversampling (ROS) yields the highest prediction performance compared to other preprocessing elements.
- Hypothesis 8: Proactive data-driven construction safety management framework can be used throughout specific periods, i.e., daily, weekly, or monthly-basis.
- Hypothesis 9: Data resampling methods have significant impacts on the accident outcome prediction and different resampling methods can yield the highest prediction power if integrated with different ML methods.

2.1 Construction Safety Management in Turkey

Safety is one of the most prevalent issues in construction projects worldwide (Choi et al. 2020). The lack of enforcement, particularly in developing countries, makes regulations non-functional and falls short of achieving their objectives regarding the safety performance of construction projects [18]. In Turkey, “Occupational Health and Safety Regulation in Construction Works” [19] defines the procedures and necessary measures, the responsibilities of parties, and the way safety plans should be prepared in the industry. Statistics show that Turkey is the 9th and 18th most unsafe country (among 73) regarding the rate of occupational fatalities and non-fatal occupational injuries to insured workers, respectively [20]. According to the Social Security Institution of Turkey [21], there were 368 fatal accidents in the construction industry in 2019. In the same year, non-fatal accidents caused over 2.5 million lost workdays.

Due to the above-mentioned reasons, construction safety management is one of the most investigated topics in the country to address diverse range of challenges. Gurcanli et al. [22] stated that the majority of the fatal occupational accidents occurred on small or middle scaled residential projects. Yılmaz and Uğur [23] compared the estimated OHS cost with actual OHS cost in the maintenance and repair projects in Turkey and found that they were 3.58% and 3.98% of the approximate total costs, respectively. Uzun et al. [24] proposed a framework for safety practices in architectural restoration and conservation projects by concentrating particular risk factors common in these projects including physical, chemical, biological, and ergonomic. Olcay et al. [25] examined fatal construction accidents in Turkey between 2012 and 2019, and found that there were no strong interrelationships among several attributes such as geographical zone, type of accident, profession, injured body part, construction site, and accident environment. Başağa et al. [26] addressed the need to improve the efficiency of safety training in the construction sector and investigated how the OHS training would be provided to the employees. They also underlined that taking divergent employee characteristics, such as education level, age, and work experience, into the consideration is essential during OHS

training. Since the fatality rate is still high in the country, developing prediction models and data-driven safety management plan aiding to detect accident outcomes proactively is required to prevent accidents and improve OHS management applications.

2.2 Machine Learning Applications in Construction Safety

An evolving and nascent AI body of knowledge has shown an expansive improvement in the field of construction safety management, which is recently supported by more accessible accident datasets. In this context, past datasets are valuable sources of information required for constituting virtual environments to make smart decisions regarding the accidents that are likely to occur in future (Goh and Ubeynarayana 2017). Based on the scope of the datasets and varying research goals, many scholars developed a great number of prediction models, providing a new practical lens to the body of construction safety management.

In this vein, Mistikoglu et al. (2015) used the decision tree (DT) method based on national data from the US Occupational Safety and Health Administration for a binary classification problem of accident outcome (fatal vs. non-fatal). A similar approach was also adopted by Ayhan and Tokdemir (2020) for a binary classification of accident severity (low severe vs high severe) by using an artificial neural network (ANN) with the data collected from construction companies. With a comparative perspective, Choi et al. (2020) evaluated the performances of several ML methods such as DT, logistic regression (LR), AdaBoost and RF to provide the most effective method in accident outcome prediction based on the national accident dataset of the Republic of Korea. The researchers suggested the RF method according to the performance scores obtained. Along with accident severity, Tixier et al. (2016) approached other aspects of accident outcomes such as energy type, injured body part, and type of accident as multi-label classification problems. They used RF and stochastic gradient tree boosting (SGTB) to model construction injury reports subjected to the natural language processing (NLP).

Accident type classification was another frequently focalized direction and considered by Kang and Ryu (2019), Zhang (2019), and Cheng et al. (2020). In this context, Cheng et al. (2020) utilized the NLP before applying the gated recurrent unit (GRU), in which the best parameters were identified via symbiotic organisms search (SOS). Zhang (2019) developed a deep neural network (DNN) structure and compared its results with five baseline ML methods such as DT, LR, support vector machine (SVM), k-nearest neighbor

(KNN), and naïve bayes (NB). While these two studies adopted the NLP approach, Kang and Ryu (2019) used the national dataset of the Republic of Korea, and focalized the feature importance to reduce and/or prevent occupational injuries based on the RF method. Ajayi et al. (2020) adopted a text mining approach and developed four prediction models, such as injured body part, lost time, binary outcome classification and, damage to fleet, utility, and equipment, using the DNN method. By adopting a similar approach but several ML methods such as RF, SVM, and XGBoost, Baker et al. (2020) focused on three aspects of accidents such as injured body part, accident severity, and type of accident. Different from all these studies, Ghodrati et al. (2018) approached safety outcomes with reference to the number of claims and number of entitlement claims using multiple regression (MR) analysis. A detailed literature survey showed the following research gaps: (1) A study that predicts the post-accident permanent disability status of construction workers has not been performed previously, (2) None of the past research integrated ML methods and metaheuristic optimization algorithms, (3) Despite the importance of data resampling methods, the pertinent literature overlooked the comparative analysis of data resampling methods and ML methods in the construction safety management, (4) A comprehensive feature engineering approach has not been well documented in the construction safety management research domain, (5) Despite the robustness of different tree-based methods has been acknowledged in the literature, no prior study compared the performances of several tree-based methods, (6) Time series analysis that integrates ML methods with wavelet transform has been neglected so far, (7) Although preprocessing is the fundamental of ML model development, there is no study in the literature that provides the most effective preprocessing pipeline. These research gaps formed the motivation to develop prediction models within the context of this thesis.

INTEGRATING FEATURE ENGINEERING, GENETIC ALGORITHM AND TREE-BASED MACHINE LEARNING METHODS TO PREDICT THE POST-ACCIDENT DISABILITY STATUS OF CONSTRUCTION WORKERS

This chapter was published in *Automation in Construction*, vol. 131, p. 103896 as a research article.

3.1 Introduction

The uncertain nature of construction projects entails high risk with a diversity of consequences such as project delay, cost overrun, occupational accidents resulting in permanent disability, and fatalities globally [7], [27], [28]. Particularly, safety issues in the construction industry cause not only permanent disability, and heavy loss of casualties for workers, but also significant social and economic costs [29], [30]. The risk of workers experiencing occupational accidents in the construction industry is higher than in other sectors in most countries [2]. Therefore, there is an urgent need to reduce the number of construction accidents and their severe consequences to provide workers with safer work-life conditions in construction projects.

Permanent disability or permanent incapacity for work is the condition of an injured worker who loses a part of or the entire earning capacity due to an occupational accident. Accidents leading to fatality, permanent disability, and irreversible health effects are considered to be at the top of injury severity levels, referred to as “catastrophic events” [31]. Over 11% of all disabilities occur as a result of occupational injuries globally [32]. As a result, construction accidents that result in permanent disability have a variety of adverse effects, including productivity loss, recruiting cost, compensation cost, compensation claim, and reputation loss [33]. Besides, construction workers who carry out their job to meet the housing needs of future generations are at high risk of not being able to meet their own needs.

According to the Republic of Turkey, Social Security Institution [21], 1,586 workers received permanent disability income after severe construction accidents in 2019, accounting for nearly 37% of all industries. In Turkey, the rate of permanent disability is even higher than the fatality rate in the construction industry (32%) compared to other industries. For every 10-million-dollar GDP contribution to the economy, a worker becomes disabled in the construction industry in the country [34]. Hence, proposing a remedy to minimize work disability among construction workers could reduce extensive social and economic burdens among workers, employers, industry, and even society [35]. Additionally, predicting the post-accident disability status of construction workers could aid in productivity planning, input availability for enhanced decision-making, and early payment planning for practitioners and policymakers.

Despite all these facts, the condition of permanently disabled construction workers after the accident is seldom examined in the literature regarding the predictive models. Therefore, this study aims to predict the permanent disability status of construction workers post-accident. Based on the diligent investigation of models and approaches proposed in the literature, it is attempted to fill research gaps in several ways and contribute to the well-being of construction workers. Major research gaps and corresponding contributions of the study are:

- It aims to predict construction injury outcomes for the worker regarding permanent disability in a binary classification by using several machine learning (ML) algorithms. In other words, whether the workers would be incapable of performing their works permanently or not as a result of construction accidents is the primary focus of the investigation. The proposed model can be run by construction firms and/or related institutions after an occupational accident as a practical implication. Literature related to permanent disability heavily relies on qualitative studies. On the other hand, limited quantitative studies use several methods and approaches rather than prediction models based on machine learning algorithms.
- A prediction model is proposed based on tree-based ensemble ML algorithms: Random Forest (RF), Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting (XGBoost), and Extremely Randomized Trees or shortly Extra Trees (ET). Tree-based ensemble ML methods have not been comparatively investigated in the construction safety management literature. Past researches frequently highlighted

the high performance of tree-based algorithms in the field [2], [11], [36]. The approach proposed in this study aims to provide a potential guideline for researchers who investigate means to increase the accuracy of prediction models in this area. It also provides insights for safety professionals and policymakers to improve safety management processes and applications in construction projects.

- This study uses the state-of-the-art optimization algorithm, namely genetic algorithm (GA), for parameter tuning in all the adopted ML models. GA has been successfully applied in predictive models in other domains [37], [38]. However, it has limited implementations for parameter tuning of ML methods specific to the construction safety management area. Integrating optimization algorithms for the hyperparameter tuning of ML methods provides a powerful means to enhance the accuracy of predictions. Thus, coupling the tree-based ensemble models with GA could aid in developing more robust safety management practices by increasing the performance of applied models.
- Detailed multi-step feature engineering is presented for pre-processing, which includes data encoding, data scaling, dimension reduction, and data resampling. Such a comprehensive pre-processing approach has rarely been applied in this field. Data pre-processing is a fundamental and preliminary step for developing prediction models, which also forms a basis for predicting a target variable. Therefore, an integrated feature engineering procedure through multi-stage pre-processing is aimed to improve the prediction accuracy of enhanced safety of workers in the construction industry.

3.2 Background

3.2.1 Machine Learning in Construction Safety

There has been a promising awareness among scholars about the contributions of historical databases to predict future accidents, which would aid the prevention of severe accidents [4]. Studies have been published regarding the prediction of several measurable units to enhance safety practices in construction projects recently. Some of the pivotal outcomes predicted are construction accident outcome, accident type, energy type, and body part injured. Researchers used either dataset collected from various national institutions and construction companies, or narratives subjected to natural language processing. Some studies in occupational accident prediction in the construction industry are shown in Table 3.1.

Table 3.1 Selected studies of occupational accident prediction in the construction industry.

Reference	Year	Data source	Adopted ML models	Predicted output	Number of output elements	Major results
[1]	2015	National data (US OSHA)	DT (C5.0 and CHAID)	Accident outcome (i.e., injury and death)	2	Fatality risk is positively correlated with fall distance and negatively correlated with the provision of safety training. The prediction accuracy was slightly higher with the C5.0 algorithm.
[7]	2016	Construction injury reports (captured from 470 contractors)	RF, SGTB	Injury type (fall, struck by, etc.), energy type (biological, mechanical, etc.), body part (head, neck, etc.), injury severity (first aid, medical case, etc.)	5,7,5,3	Based on RPSS values, SGBT showed better performance in all classifications.
[9]	2016	Data captured from wearable inertial measurement units	One-class SVM	Near miss fall detection (i.e., near miss, normal walk)	2	Among experiments performed in the laboratory and outdoors, total accuracy was slightly higher in the laboratory (86.8% compared to 85.2%).
[4]	2017	National construction narratives (US OSHA)	SVM, LR, RF, KNN, DT, NB	Accident type (i.e., electrocution, exposure to chemical substances, explosion, falls, etc.)	11	Linear SVM and RBF-SVM with unigram tokenization were the best classifiers based on the F1 score.
[39]	2018	Foot plantar pressure distribution data collected by a wearable insole pressure system	ANN, DT, KNN, SVM	Awkward working postures (i.e., overhead working, squatting, stopping, semi-squatting, one-legged kneeling)	5	Based on cross-validation results, SVM showed the highest prediction performance according to accuracy score.

Table 3.1 Selected studies of occupational accident prediction in the construction industry (continued)

Reference	Year	Data source	Adopted ML models	Predicted output	Number of output elements	Major results
[11]	2019	National accident (KOSHA) and meteorological (KMA) data	RF	Accident type (i.e., falls from height, collision by object, rollover, falling object)	4	The accuracy of RF was 71.3%. Assailing materials and original-cause materials were the most influential attributes.
[5]	2019	National construction narratives (US OSHA)	SVM, LR, KNN, DT, NB	Accident type (i.e., electrocution, exposure to chemical substances, explosion, falls, etc.)	11	SQP algorithm to optimize the weight of each classifier to develop ensemble model outperformed rest of the models based on weighted average F1 score.
[16]	2019	Construction companies comprising 87 construction sites	ML, ANN	The severity of accidents (at-risk behavior/near miss, incidents with material damage, first aid, medical intervention, lost workday cases, fatalities)	6	ANN showed higher performance while integrating fuzzy set theory improved the prediction of at-risk behavior and lost workday cases.
[2]	2020	National accident data (MOEL of the Republic of Korea)	LR, DT, RF, AdaBoost	Accident outcome (i.e., injury and death)	2	RF showed the highest prediction accuracy according to AUROC. The month and employment size were the most influential factors.
[8]	2020	From a construction company	SVM, DT	Year, type of work, type of accident, injured part, assailing materials, severity, and cause of the accident	Several combinations	Seven trials by taking one attribute as output and others as input showed that accuracy was the highest when the cause of the accident was set as an output variable.

Accident classification is another under-researched area in the literature. Goh and Ubeynarayana [4] compared the performances of SVM, LR, RF, KNN, and NB algorithms to classify construction accidents into 11 labels, such as electrocution, exposure to chemical substances, and explosion. The researchers used word stemming for pre-processing and grid search algorithms for parameter optimization of machine learning models. Based on the data collected from the US OSHA, it was found that SVM with unigram tokenization outperformed other models with respect to the F1 score, in which electrocution was the most correctly labeled case. By using the dataset from the same data source, Zhang et al. [5] proposed an ensemble prediction model based on SVM, LR, KNN, DT, and NB algorithms. The researchers used a sequential quadratic programming algorithm to optimize the weights of stand-alone ML models, which improved the prediction performance with respect to the average weighted F1 score. Kang and Ryu [11] focused on the RF algorithm for construction accident type prediction. The researchers removed the classes of less than 10% and used random under-sampling (RUS) to solve the class imbalance issue. The RF algorithm showed a prediction accuracy of 71.3%, and assailing materials and original-cause materials were the most influential attributes.

For accident severity classification, Ayhan and Tokdemir [16] conducted the latent class clustering analysis for dimension reduction before performing the ANN algorithm. The researchers showed that implementation of fuzzy set theory improved the prediction accuracy with respect to at-risk behavior and lost workday cases severity levels. Antwi-Afari et al. [39] compared the performances of ANN, DT, KNN, and SVM to predict awkward working postures based on data captured through a wearable insole pressure system. The cross-validation results showed that SVM achieved the highest accuracy score. Tixier et al. [7] developed a prediction model for injury type, energy type, body part, and injury severity classifications using random forest and SGBT algorithms. They found SGBT was showing better performance compared to random forest algorithm in all classifications based on the construction injury report collected from 470 contractors. Another study to develop different classification schemes was performed by Lee et al. [8], based on a dataset collected from a company. The researchers predicted several attributes such as the year, cause of the accident, and type of the accident with SVM and DT models. They found that the accuracy was the highest when the cause of the accident was set as the output variable. A comprehensive literature review shows that the accident outcome with respect to permanent disability has not been investigated on a prediction

level. In addition, the adoption of GA for tuning the hyperparameters of ML models has been overlooked in the construction safety literature. Furthermore, past research on construction accident outcome prediction lacks a detailed feature engineering approach, which has been considered as a multi-step perspective in this study.

3.2.2 Methodological Background

In recent years, quantitative and qualitative approaches have gained significant interest in the construction safety literature. Between the two, quantitative research has become more dominant, particularly with the advancements in artificial intelligence techniques. Recent evidence suggests that machine learning methods are among the most prominent instruments. Therefore, prediction researches related to construction accidents have been carried out using various machine learning methods, such as RF [11], DT [1], SVM [9], KNN [39], NB [5], ANN [16], and Adaboost [2].

Data pre-processing has a significant impact on prediction models by improving the practicality and model efficiency, as the data sets associated with construction accidents contain plenty of variables. There are several categorical attributes affecting construction accident outcomes, and the dimension of a problem increases with the implementation of data encoding (specifically one-hot encoding). Therefore, performing the dimensionality reduction strategies is effective in reducing the complexity of both the model and computations.

Hyperparameter tuning is one of the challenges inherent in ML methods. Depending upon the selected ML methods, several internal parameters are needed to specify the optimum model configuration. Therefore, several techniques have been presented in the pertinent literature to optimize the parameters of ML methods. Choosing an appropriate hyperparameter tuning strategy is critical for controlling the overall behavior of the employed methods. The methodological background of this study comprises the following sub-sections: 3.2.2.1. Tree-based machine learning methods, 3.2.2.2. Dimensionality reduction, and 3.2.2.3. Hyperparameter optimization.

3.2.2.1 Tree-based ML Methods

The classical learning algorithms generate a single model from a training sample, while tree-based ensemble methods ground on the training of several algorithms to deal with a single problem. Hence, ensemble methods compose multiple models by training samples and combining the results depending on classification or regression purposes. These

algorithms have been the subject of great interest in prediction studies, as they provide information on causal effects and processes [40]. Additionally, they can render the importance levels of the attributes and allow improved evaluations of the obtained results. Thus, the ability to explain the cause and effects between the predictors and targets makes the tree-based models more interpretable compared to other ML methods [41]. They can provide considerably high performance in handling different predictor variables [42].

In this study, Random Forest, XGBoost, Adaboost, and Extra Trees were used to predicting the permanent disability of construction workers. The working principles of these methods are illustrated in Figure 3.1. In general, tree-based machine learning methods are systems that comprise a set of branches and nodes, providing the best splitting strategy by starting from the top or root of the tree to estimate the class label or numerical values [40]. They are mainly based on the recursive splitting strategy of training samples. This process stops at the last branches, where the splitting can no longer be achieved, and the uniformity (or homogeneity) is provided in the sub-nodes.

The fundamental of tree-based methods is the splitting of decision trees (DT), which relies on a set of “if-then” rules that can be visualized by a tree [43]. Even though decision trees have some pros, such as enabling interpretability with their easy-to-understand schemes [44], they are also associated with some cons that make them less useful. For instance, the predictions are likely to be time-consuming and less accurate, especially when the problem is complex. Besides, they are prone to overfitting due to the inherent model instability, in which a slight change in a training set may cause significant variations in the predicted outcomes [45]. To overcome these challenges, more advanced models, such as random forest and boosted trees, are introduced.

RF is a non-parametric method that combines DT with the ideas of bagging and bootstrap, giving considerably accurate results in both regression and classification problems [46]. The RF method provides high predictive performance in large datasets containing numerous independent variables and missing values. In addition, it is described as a resolute algorithm against commonly encountered problems in ML algorithms, such as overfitting. The RF adopts the principle of combining predictive trees based on the values of independently sampled random vectors [47]. The classical Classification And Regression Trees (CART) algorithm is used for the generation of trees in the RF modeling system [48]. However, different from the CART, the best branching methodology is grounded in using the entire database. RF determines the best-branched variable in each

leaf node using the random subspace method. Despite these advantages, RF has a downside of predisposing the features, which affect the dataset with categorical attributes at different levels significantly.

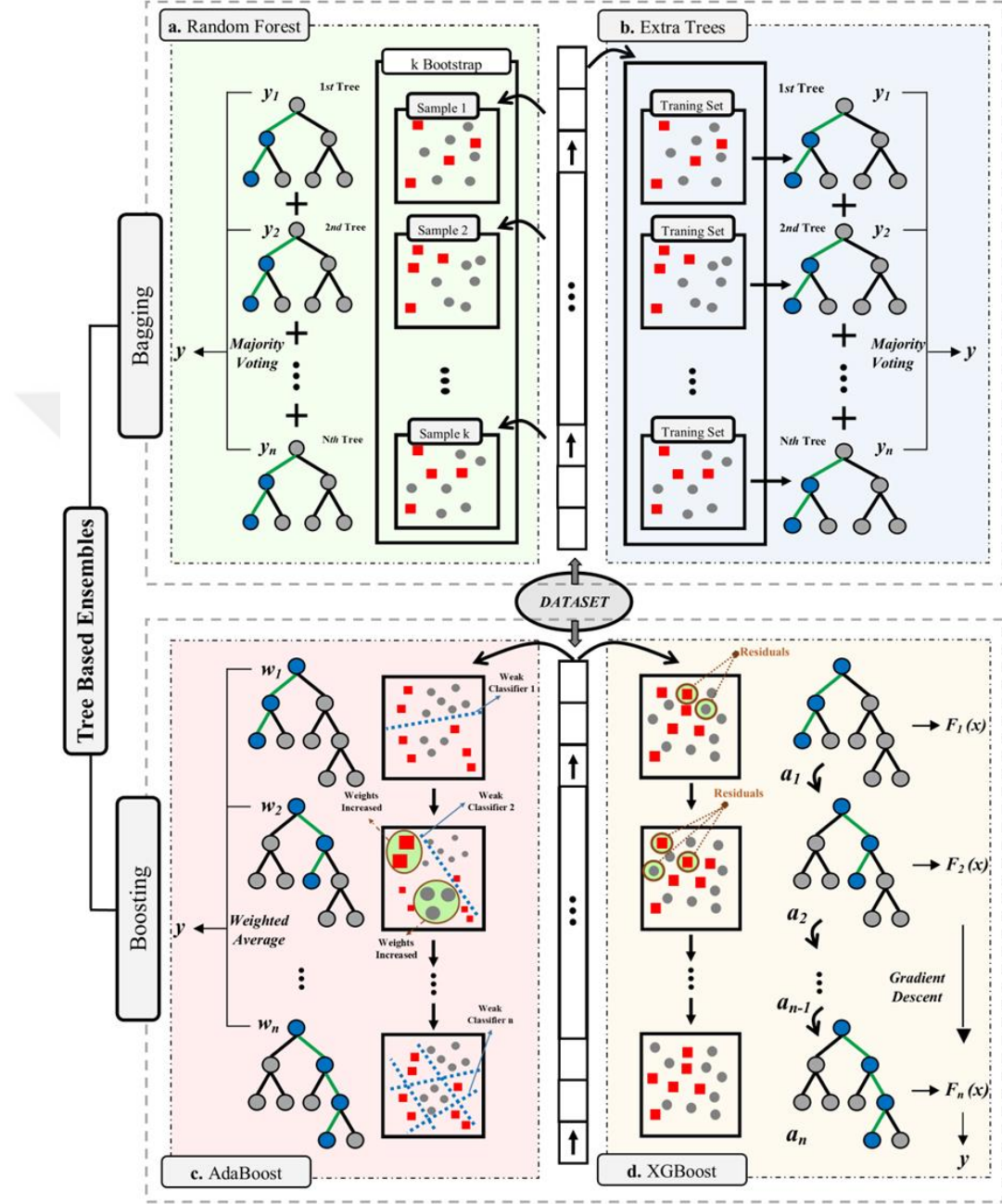


Figure 3.1 Working principles of tree-based ensemble machine learning methods

Extremely Randomized Trees or shortly Extra Trees (ET) is used for both classification and regression problems. Essentially, ET has a similar philosophy to the classical RF, yet it differs according to two fundamental points [49]. The first difference comes from the fact that ET separates the nodes by choosing completely random cut-points. Similar to RF, a random subset of input features is used. However, instead of searching for the best

threshold value when splitting the nodes, a random cut-point value is assigned for each aspect in the input data [50]. The aim is to reduce the high computation load, which occurs as a result of finding the optimum split for each node, as well as to achieve more tree diversity. In addition, the learning process is furthered in the ET method in comparison with other tree-based algorithms. The second difference comes from using the entire learning instance to grow trees rather than bootstrap copying [51]. Even though using the whole dataset for training results with a decrease in bias with the application of random splitting for each node, this process also leads to a decrease in the variance.

As a general ensemble methodology, boosting is based on the principle of employing many weak learners, such as classifiers that have relatively high bias error, one after the other, to increase the classification accuracy [2]. The AdaBoost inherently benefits from the DT in the generation of the weak classifiers. Each weak classifier focuses on the training of instances that are incorrectly classified by the previous weak classifier to re-classify them properly. In other words, the misclassified samples are identified at each iteration to increase their weights as the next classifier elaborates to provide the correct classification of the previously misclassified data points. It is important to note that the output of the system is obtained by the weighted average of the outputs of weak learners. AdaBoost mainly minimizes the exponential loss function, which makes the algorithm sensitive to outliers.

Gradient boosting was introduced for the utilization of any differentiable loss functions [52]. Even though the Gradient Boosting Decision Trees (GBDT) are more robust than Adaboost in the detection of outliers, its proper regularization of additive iterative process limits the generalization of the GBDT algorithm, such that the regularization is controlled by the complexity of the model. Therefore, the XGBoost algorithm was proposed to improve the performance of the GBDT. Different from the GBDT, XGBoost may yield higher accuracy by using various regularization techniques and controlling the complexity of trees [53]. With its advanced system comprising parallel tree boosting, both regression and classification problems can be solved by employing the constructed boosted trees [48]. In addition, the primary reason underlying the superiority of the XGBoost is the objective function used during the learning process. The objective function comprises a loss function and a regularization term. The loss function calculates the difference between each estimated value and the actual value, as the regularization term controls the complexity of the model, all while avoiding overfitting in the model. In

addition, XGBoost adopts the second-order Taylor expansion to the loss function to achieve the minimization of objective function [54]. Thus, the algorithm not only provides fast computation processing but also enables achieving more reliable results based on its robust structure.

3.2.2.2 Dimension Reduction

In several disciplines, large datasets could be challenging to use, particularly with the application of machine learning methods along with parameter optimization. Therefore, dimensionality reduction strategies have been proposed to represent high-dimensional data with a smaller number of components. With these methods, preserving the statistical information and increasing the interpretability is satisfied [55]. There are several dimension reduction techniques employed in the literature, such as discriminant analysis, singular value decomposition, and Principal Component Analysis (PCA). Among them, PCA is the most widely used method in construction safety literature as a prominent tool for dimensionality reduction [3], [56]–[58].

The basic principle of the PCA relies on representing the high dimensional sample space using the lower dimensional sample sets without losing vital information. Each principal component contains information from all variables. Through PCA, both visualization and interpretability of the original data set are provided [59]. Additional features, which are transformed into the new orthogonal coordinate system, are extracted and arrayed in descending order, in which the first component explains an essential part of the original data by capturing the most significant variance. In addition, accomplishing the uncorrelated structure among the principal components, one of the preeminent problems frequently encountered in modeling studies [60], is vital to avoid linearity [61].

3.2.2.3 Hyperparameter Optimization

Determining the hyperparameters of machine learning methods is crucial for the identification of optimum model configurations. Hyperparameter optimization controls a model's learning process. There are numerous methods for hyperparameter tuning, such as trial-error, grid search, random search, Bayesian optimization, and meta-heuristic algorithms. Some of them include an exhaustive process for users, while others provide automatic calibration utilities.

The trial-error can be referred to as manual calibration and requires adjustments in the parameter values by a user within the lower and upper boundaries. Despite being one of

the most widely used methods of parameter tuning [62], [63], its contributions to machine learning models are limited [64], [65]. Several researchers adopted the grid search algorithm to minimize tedious iterative searching tasks commonly encountered during trial-error [4], [9], [66]. However, the grid search algorithm is also based on a time-consuming procedure, searching parameters for each combination within a particular grid space.

Due to the abovementioned disadvantages of traditional methods (trial-and-error and grid search), optimization algorithms have been presented for the hyperparameter tuning of ML methods. As a rule of thumb, optimization algorithms perform the hyperparameter mapping by using the present parameter value for the next iteration, unlike performing blind searching, as in traditional methods. Also, model efficiency is increased regarding both computational time and predictive power by employing these techniques [67]. There are many optimization algorithms proposed for hyperparameter tunings, such as Bayesian optimization, gradient-based optimization, and heuristics, such as particle swarm algorithm and genetic algorithm (GA) [65]. Even though the heuristics do not guarantee finding the optimal solution, they are widely used as fast and efficient tools to find a near-optimal solution even with a small memory space [68].

The heuristic optimization algorithms are based on concepts, which have already been explored in nature. They can be identified as the “experience-based” methods and have become considerably feasible for solving complex optimization problems [69]. These methods perform effective searching by scanning a particular part of the solution space instead of a whole. GA, of the most widely used meta-heuristics [65], is a stochastic approach providing parallelized search for population points instead of a single point. This approach differentiates GA from the other algorithms as it avoids being trapped in local optima and discovering the global solution [70]. GA does not depend on alternative techniques, while other metaheuristics such as differential evolution and firefly algorithms require their parameters to be tuned by different algorithms [70]. In addition, there are several features of GA which make it beneficial compared to other meta-heuristics. These features include being a universal optimizer for solving a great deal of nonlinear and non-smooth types of optimization problems, being easy to implement, having the ability to conjunct with the other meta-heuristics, and solving problems that have high computational complexity [71].

Owing to its benefits mentioned above, GA was used in this study to define the optimum hyperparameters of four tree-based machine learning methods separately. As an adapted version of the natural selection mechanism of the computer environment [72], all available data are subjected to elimination using genetic operators to obtain the best solution in the GA operation process. The best solution comprised individuals that mostly adapted to the environment, in other words, survivals as a result of iterations. In GA, there is no strict rule for determining the number of individuals (i.e., chromosomes) in the population. While many chromosomes extend the computational time, a few individuals in the population may lead to a failure in reaching the optimum solution as well as causing the system to get stuck in a particular solution space. Therefore, choosing the optimum number of individuals depending on the characteristics of the problem is critical in GA applications [73]. Several replicas of the genes belonging to successful individuals are transferred to the next generations, which leads the general population to the best solution. Furthermore, crossover and mutation operators play a significant role in creating a new generation and substituting with the old generation. It is also important to note that recent advances in the usage of GA have displayed that GA could be successfully applied, especially in the optimization of complex systems [74]–[77].

3.3 Research Methodology

3.3.1 Research Design

This study aims to predict the outcomes of construction injuries that construction workers might have due to occupational accidents. In other words, the constructed framework envisions whether the victim will be permanently disabled (PD) or not (N-PD). To achieve this holistic approach, a five-step comprehensive methodology was designed, as illustrated in Figure 3.2. First, a dataset comprising construction accidents in Turkey between 2012 and 2020 was obtained from the Republic of Turkey, Social Security Institution. The attributes that are directly influential in predicting the outcomes of occupational injuries with respect to any permanent disability of workers were diligently investigated. In the second step, the dataset was thoroughly examined regarding the missing values, duplicates, inconsistencies, etc., for data cleaning. As a result, 47,938 construction accident data were attained, including 1,074 PD and 46,864 N-PD instances. Since the dataset contained some categorical variables, data encoding was required for ML implementations. Thus, all the categorical data were grouped as either nominal or

ordinal, and subjected to one-hot encoding and label encoding, respectively. As a result, attained attributes were subjected to the standard scaling as a rule of thumb.

It should be noted that dimension reduction is required after applying a one-hot encoding scheme since it results in extremely high dimensionality. Therefore, the pre-processed dataset was subjected to principal component analysis for dimension reduction. In order to test different principal component (PC) scenarios, transformed data was divided into two major categories as training set and testing set, by taking 70% and 30% of all data in the third step. The accuracies of the training set with different principal components were tested through k-fold stratified cross-validation (CV). As a result, the dimension of the dataset was reduced.

The original dataset was highly imbalanced, such that the N-PD worker class was 44 times more prevalent than the PD worker class. Therefore, a random under-sampling algorithm was applied to the training set only to address the class imbalance issue. After implementing several pre-processing methods, such as data cleaning, variable encoding, principal component analysis, and random under-sampling, four widely used tree-based ensemble models, RF, XGBoost, AdaBoost, and ET, were used to predict injury outcomes in the fourth step. A state-of-the-art optimization algorithm, Genetic Algorithm, was adopted to tune the hyperparameters of four classifiers to improve the prediction accuracy based on k-fold stratified cross-validation subjected to the training set. The best parameters in each classifier were then rigorously examined with blinded testing set by using several performance evaluation metrics, such as precision, recall, F1 score, accuracy, and AUROC. This approach enabled an increase in practical implication and applicability of adopted pre-processing steps and GA for the newly introduced testing set.

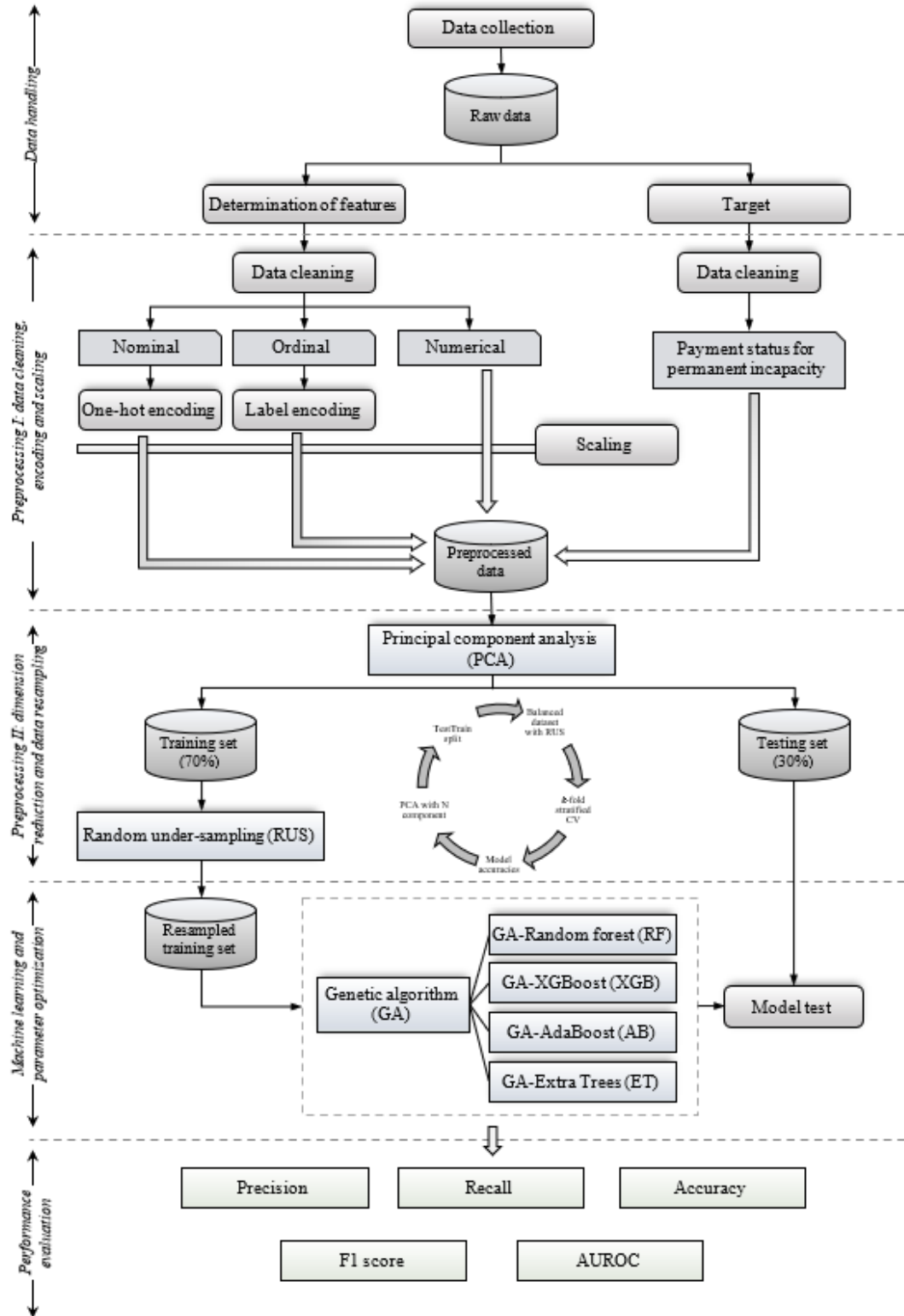


Figure 3.2 Research flow

3.3.2 Data description

This paper explores the classification of permanent disabilities resulting from construction accidents recorded in Turkey between 2012 and 2020. The used data was

obtained from the Republic of Turkey, Social Security Institution. In this context, eight attributes were determined as inputs to predict the target attribute, the permanent disability. The summary of the predictor and target attributes is presented in Table 2. The data contained three different variable types; categorical (education level, type of injury, injured body parts, materials used, and the type of construction) and numerical (working days lost, number of past accidents of the worker, and daily wage) for the inputs, and binary for the output.

It is important to note that some attributes, such as age, gender, and tenure, were not included in this study. These characteristic variables could have notable effects on the outcomes of construction incidents in different contexts, such as whether near miss or accident and fatal or non-fatal. However, their impact on disability status is relatively indirect. Disability occurs when a particular body part of a worker [31] is lost (at some degree) after exposure to certain injury types. Thus, attributes were selected in a way that defines the work environment and accident condition of workers rather than their demographics.

Attributes such as education level, materials used, type of construction, and daily wage were included since these attributes could be related to the work performed by a worker. Kamardeen and Rameezdeen [31] showed that the injured body part and injury type were strong determinants of severe accidents, such as permanent disability. As for the working days lost, the type of accident (such as falling from a height or being hit by an object) could be directly related to the number of workday losses [78]. For instance, Hess et al. [79] stated that a particular work of drywall handling that strains the back accounted for the highest percentages of permanent disabilities and working days lost. The workday loss could even be regarded as a disability for temporary incapacities. Similarly, Mohd Kamar et al. [80] found that accident costs, which were chiefly associated with disability rates, were mainly affected by injured body parts, type of injury, and working days lost. Hence, eight attributes (Table 3.2) contributed to predicting the permanent disability status of workers in this study.

It is important to note that some of the categorical input variables comprised a considerable number of classes, some of which included only a few instances. Considering the computational efficiency and practical applicability, three input variables such as education level, materials used, and type of construction were re-classified before the application of ML methods to reduce the dimension of the input space. Hence, the

number of elements was reduced from 11 to 6, 157 to 17, and 322 to 10 for education level, materials used, and type of construction, respectively.

Table 3.2 Description and distribution of attributes

Variable names	Data type	Number of elements	Definitions
Education level	Categorical	11(6)	The education level of a worker. It includes six categories: Illiterate or only literate, elementary school, secondary school, high school, graduate from two-year college, BSc or MSc or Ph.D.
Working days lost	Numerical	0 – 512	The number of days that worker cannot work temporarily. It includes numerical values between 0 and 512 days.
Type of injury	Categorical*	7	Type of injury exposed to the workers, which was expressed in seven categories: Acute infections, exposure to chemical substances, exposure to extreme temperatures, closed fracture, shock due to assault and threat, fluid effects, and more than one type.
Injured body parts	Categorical*	14	The body part affected by an accident, which was expressed in fourteen categories such as hand and wrist; cervical vertebrae; neck (others); arm; tooth; chest; neurotic; head; leg and femur; ear; shoulder; foot and fingers; thoracic vertebra; backbone (others).
Materials used	Categorical*	157(17)	Any material involved in an accident, which was seventeen categories, such as construction machinery; chemical, biological substances, waste, and contaminants; construction materials (concrete, brick, etc.); office materials; creatures; assembly kits (nail, bolt, etc.); storage systems; hand tools with an engine (cutting machines); gas and explosive materials; building parts (door, window, etc.); hand tools without engine; distribution and piping systems; dust, silt, grain; underground areas (tunnels, sewage, etc.); stairways, ladder, and surfaces; disasters; other.
Type of construction	Categorical*	322(10)	Type of construction has ten categories, such as residential building projects; non-residential building projects; road projects; railway and metro projects; infrastructure projects (tunnel, road, bridge projects); electricity and telecommunication projects; water projects (including harbors); demolition works; earthworks; other.
Number of past accidents of the worker	Numerical	0 – 6	The total number of past accidents of the worker attained from the occupational accident database and includes values between 0 and 6.
Daily wage	Numerical	0 – 639.6	The daily wage of a worker and includes payments between 0 (for intern) and 639.6 in Turkish Lira (TL).
Permanent disability	Binary (target)	2	Target variable indicating whether workers became incapable of performing their works permanently (PD) or not (N-PD).

Note: The values in the parenthesis indicate the number of elements after re-classification.

*Variables were subjected to one-hot encoding.

Also, since the entire input space includes both nominal and ordinal categorical data types, the nominal attributes, such as the type of injury, injured body parts, materials used, and type of construction, were subjected to one-hot encoding. The rest of the categorical variables were subjected to label-encoding. This pre-processing step was done since natural ordering among categorical variables was most likely to lead to fallacious classifications. Overall, 47,938 construction accidents, consisting of 1,074 permanently disabled and 46,864 not permanently disabled records, were considered as a result of diligent examinations of inconsistencies, duplicates, and missing values.

3.3.3 Data Pre-processing

3.3.3.1 Data Encoding

The dataset used in this study includes categorical features, such as education level, type of injury, injured body parts, the material used, and type of construction. However, most ML methods work well with numerical features [81]. Therefore, categorical features should be transformed into their numerical counterparts. In this context, there are two prevalent methods to deal with categorical data; label encoding and one-hot encoding [82]. In label encoding, the feature values are converted into specific numerical values so that the total number of features remains the same [83]. Differently, one-hot encoding generates a new binary feature for each element in a categorical variable. Figure 3.3 provides an example to illustrate and compare label and one-hot encoding procedures. Label encoding is easier compared to one-hot encoding. However, ordering the issue could cause misinterpretation of numeric values [84]. Thus, only education level was subjected to label encoding in this study since ordering existed from 1 to 6 in the education level feature. In contrast, there is not any ordering matter for other categorical features. Hence, features such as type of injury, injured body parts, the material used, and type of construction were subjected to one-hot encoding. This procedure converted 8-feature dataset into a 52-feature dataset, including, *i*) 4 features from education level, working days lost, the number of past accidents of the worker, and daily wage; *ii*) 7 features from the type of injury; *iii*) 14 features from injured body parts, *iv*) 17 features from materials used, and *v*) 10 features from the type of construction.

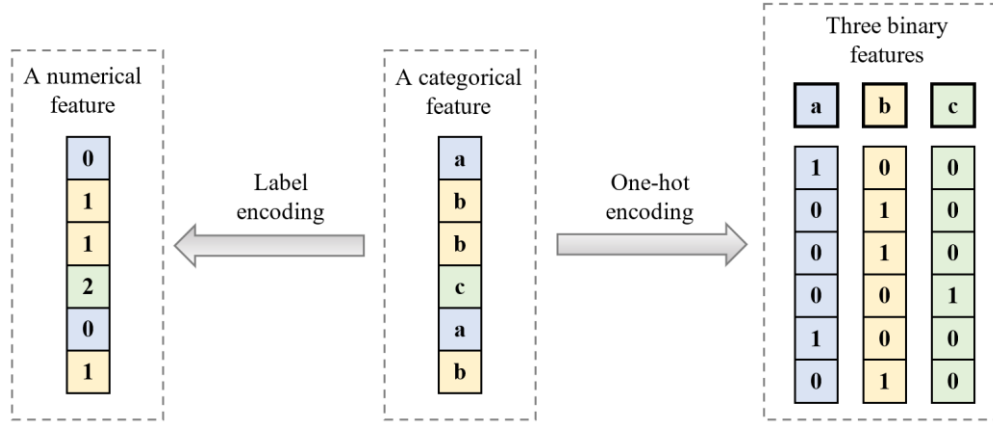


Figure 3.3 Data encoding procedure

3.3.3.2 Data Scaling

Most ML methods do not perform well with different scaling of numerical features [81]. There are two widely used solutions to address this issue: the *Min-Max scaler* and the *Standard-scaler* [85]. The *Min-Max scaler* creates a range (ordinarily 0-1) and scales numerical features into the predetermined range through zooming in or out [86]. On the other hand, the *Standard-scaler* converts numeric features in such a way that results in a distribution having a mean value and standard deviation of 0 and 1, respectively [87]. This study adopted the *Standard-scaler* technique for feature scaling before performing PCA analysis due to its wide applications [82]–[84], [88] and effectiveness [86]. Test x 's regular score was defined by subtracting the mean value from each test value and then divided by the standard deviation as:

$$z = \frac{x - m}{s} \quad (3.1)$$

where u is the mean value and s is the standard deviation.

3.3.3.3 Dimension Reduction

Since one-hot encoding increases the dimension of the dataset considerably, dimension reduction methods have widely been used in the literature after applying one-hot encoding [84], [85], [87]–[89]. Dimension reduction is a process that transforms a large number of variables into a small number of variables without losing crucial information [84]. Considering its benefits and application of one-hot encoding that results in significantly increased dimensions, PCA was used in this study. Two consecutive steps were adopted for principal component selection after PCA analysis. In the first step, Kaiser's criteria

[90] was adopted, in which eigenvectors whose eigenvalues were lower than 1 could be ignored. Therefore, the number of principal components with an eigenvalue higher than 1 was set as the z value. Since prediction accuracy is of utmost importance for ML models, it was considered in the second step of principal component selection [91]. Hence, the accuracies of four adopted ML algorithms were further calculated for different scenarios, by setting the number of principal components as 1, 2, ..., z . This process highlighted the effects of different lengths of the feature vector on the accuracy of the ML models. For each scenario, training (after resampling) and testing sets were captured to be used in the subsequent stages of the study to perform fair comparisons. The average of the ML algorithm's accuracies was then computed, and the principal component value giving the highest average accuracy was selected.

3.3.3.4 Data Resampling

The dataset used in this study was severely imbalanced, such that the ML models achieved a prediction rate of 98% erroneously in the case of classifying all the data as N-PD. Under these circumstances, the dataset needed to be resampled to produce the class distribution balance. In the relevant literature, there are three most widely used resampling strategies, which are *i*) random under-sampling (RUS), *ii*) random over-sampling (ROS), and *iii*) synthetic minority over-sampling technique (SMOTE) [2]. Class distribution is balanced by randomly eliminating the majority class in the RUS algorithm [11] while repeating the minority class in the ROS algorithm [92]. Meanwhile, SMOTE generates synthetic samples of the minority class through cluster-based interpolation [93]. In this study, the RUS strategy was used due to several reasons. First, ROS might lead to overfitting and reduce the prediction performance in invisible instances [94]. In addition, artificial positive instances are generated by interpolation in SMOTE, which could work well on numerical data [2]. RUS was chosen as the best strategy for this investigation after examining several categorical variables in the dataset and taking computational time and storage into account [94]. Although there was a risk of information loss in RUS, no artificial instances were generated. Besides, RUS is argued to be a powerful alternative when the dataset is large [95]. Its wide implementations showed its successful applications as a resampling strategy and were used [11], [96] or suggested by several researchers [5], [97].

3.3.4 Training the Machine Learning Methods

In this study, four tree-based machine learning methods were integrated into the genetic algorithm to predict the permanent disability of construction workers. The hyperparameter tunings of the RF, ET, AdaBoost, and XGBoost methods were performed using a genetic algorithm. In tree-based methods, the model parameters can vary according to the working principle and internal dynamics of the algorithms. However, a parameter called the *number of trees* is prevailing in each algorithm. Determining the number of trees is crucial as it plays a key role in providing an improvement in the models and enhancing the generalization [98].

It is important to note that the optimal number of trees is highly related to the adopted methodology. For instance, with bagging techniques like random forest and extra trees, the more trees there are, the better the model's performance is expected to be [99]. However, boosting algorithms, such as Adaboost and XGBoost, can overfit the model when the number of trees increases. It is essential to note the trade-off between the number of trees and computational cost [100]. Other parameters in tree-based ML methods include the learning rate, which is common in boosting algorithms, as well as the maximum depth and the minimum number of samples, which are prevalent in RF and ET.

Overall, since the parameters mentioned above could vary regarding the adopted methodology and used dataset in pre-defined ranges, the state-of-the-art genetic algorithm was adopted for hyperparameter tuning. 1,000 combinations were searched for each algorithm to establish a decent datum line. The range of the number of trees was fixed between 50 and 1,000, with a step of 50 for each algorithm. Similarly, the range of learning rate and the range of maximum depth were fixed between 0.01 and 0.50 (step = 0.01) and between 1 and 50 (step = 1), respectively. The internal parameters of GA were determined based on the trial error. The population size was determined as 25, and the tournament size was chosen as 5 [101]. In addition, the mutation rate and the crossover rate were identified as 0.25 and 0.5, respectively [102].

3.3.5 Performance and Evaluation Metrics

There are several performance evaluation metrics to measure the accuracy of the proposed models, including Precision, Recall, F1 score, and AUROC. These performance measures

depend on the classification outcomes, including four potential cases in binary classification, as listed in Table 3.3.

Table 3.3 Possible cases between actual and predicted conditions

Actual cases	Predicted cases (Yes)	Predicted cases (No)
Yes	True Positive (TP)	False Negative (FN)
No	False Positive (FP)	True Negative (TN)

Precision is an evaluation metric showing the accuracy of the positive predictions as the rate of true positive instances to the sum of positive instances, while recall is the rate of true positive instances to the sum of true positive and false negative instances [103].

$$Precision = \frac{TP}{TP + FP} \quad (3.2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3.3)$$

The F1 score provides an overall assessment scheme by combining precision and recall by taking their harmonic average [104].

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3.4)$$

With balanced class distribution in a binary classification, the model performance can be measured as the correctly classified instances divided by the sum of all instances, such that the accuracy is defined by [105]:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (3.5)$$

However, this evaluation metric is impractical if the class distribution is imbalanced. In such a case, the prediction could become biased towards the majority class and propose erroneously higher accuracies without even predicting a single minority class accurately [106]. Therefore, two performance evaluation metrics should be used simultaneously as true positive rate (TPR) and false positive rate (FPR) [2]. TPR and FPR are calculated as correctly predicted positive values to the true positive values and incorrectly predicted

negative values to true negative values, respectively. They are defined in Eqs. (3.6) and (3.7).

$$TPR = \frac{TP}{TP + FN} \quad (3.6)$$

$$FPR = \frac{FP}{FP + TN} \quad (3.7)$$

The receiver operating characteristic (ROC) curve is a robust graphical method to measure the classification performance [107]. It plots TPR as a function of FPR, which are placed on the vertical and horizontal axes, respectively. The closer the curve is to the upper left corner of the graph, the better is the classifier performance. In other words, the higher area under the ROC curve (AUROC), the better the prediction of a classifier. It is considered a reliable instrument to measure the performances of several ML algorithms, which has been widely adopted by scholars for its easy visualization and interpretation [108]. Along with other evaluation metrics, this study essentially focused on AUROC values for model comparisons.

3.4 Results and Discussion

3.4.1 Generation of features

52 features were obtained as a result of the one-hot encoding applied to the nominal categorical variables. Since the ML algorithms work relatively slower when the number of inputs increases [40], PCA was adopted to reduce the dimension of input space. The findings from the PCA analysis revealed that the eigenvalue of the corresponding eigenvector becomes lower than 1 after 31 principal components, which explained 71.99% of the total variance. This result showed that the number of inputs could be chosen as 31 to satisfy the statistical significance. In this research, the threshold value was set as 31 for the number of principal component selections. The dimension of the input space was reduced by using four standalone tree-based ensemble ML methods. In this sense, the accuracies of the XGBoost, AdaBoost, ET, and RF were obtained step-by-step, starting from 1 component up to the 31 components. The entire procedure adopted for principal component selection is illustrated in Figure 3.4.

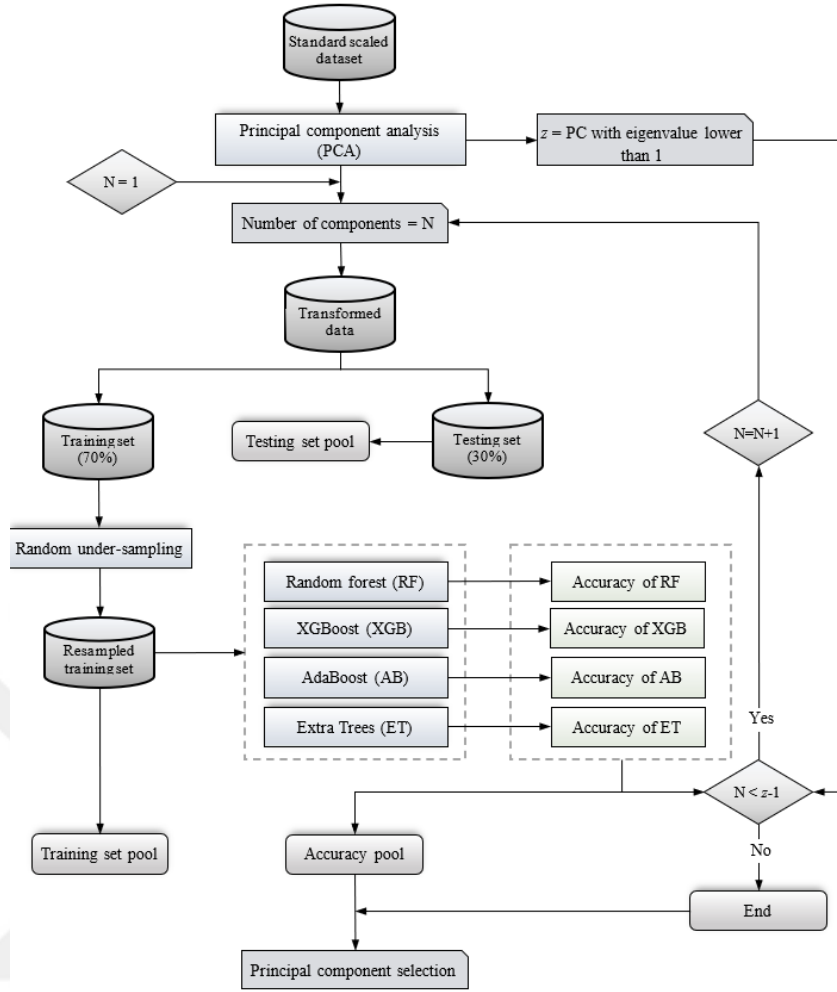


Figure 3.4 Principal component selection procedure

The results of estimations attained for each component are presented in Figure 3.5. The different colors of solid lines show the accuracy of four distinct stand-alone ML methods, and the black dashed line represents the average accuracy of the ML methods. Since the ML methods yielded varied prediction performances for different principal component scenarios, the average of four ML methods was chosen as the optimum number of principal components. Figure 3.5. depicts that the optimum number of PCs could be five since the maximum average accuracy was obtained as 77.49%. Computational efficiency was enabled for further predictions, particularly considering the integration of GA for hyperparameter optimization. The heat map showing the factor loadings of each attribute is presented in Figure 3.6. The “working days lost” was directly proportional with Component 1, while the “daily wage” had strong loading onto Component 2. Nearly all elements of “type of injury” were positively loaded onto Component 3 and Component 4. Even though some instances of “type of construction” were inversely loaded onto

Component 5, significant loadings could be observed onto Component 5 considering the overall instances.

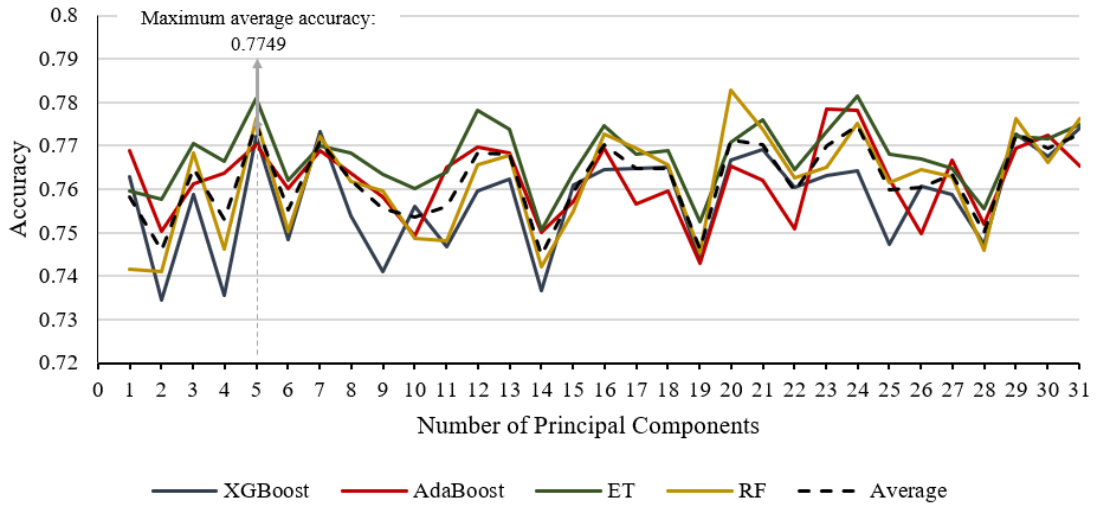


Figure 3.5 Trials of principal components for each classifier

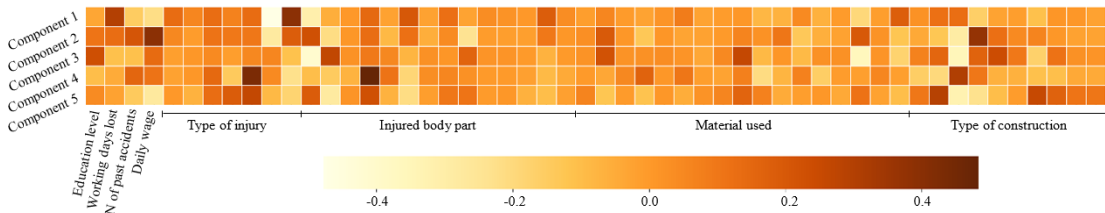


Figure 3.6 Heat map of PCA

3.4.2 Model Calibration

Optimization of the hyperparameters of ML algorithms is essential not only to increase predictive accuracies but also to save computational time. Therefore, in this study, the Genetic Algorithm was adopted to optimize the internal parameters of ML methods (RF, XGBoost, AdaBoost, and ET) and stabilize the error rates in the generated models.

Figure 3.7. provides the calibration of the ML algorithms with respect to the number of generations, in which the maximum, average, and minimum accuracies of each iteration were obtained in the pre-defined population size. As seen from the figure, the maximum number of generations was set to 500. The results obtained from the hybrid GA-ML methods for the training sets showed that the maximum accuracy was obtained at iteration 79 with 78.41% accuracy for Random Forest (Figure 3.7a) and iteration 17 with 78.31% accuracy for XGBoost (Figure 3.7b). A summary of the established search mechanisms, including the parameter ranges, best parameters, and the number of total combinations, is

illustrated in Table 3.4. The optimal values for the number of trees and maximum depth were obtained as 950 and 9, respectively for RF, while the learning rate and the number of trees were attained as 0.1 and 50, respectively for the XGBoost algorithm. In addition, the best parameters were found as 250 and 0.11 for the number of trees and the learning rate in the calibration of AdaBoost at iteration 72 with 78.27% accuracy (Figure 3.7c). Finally, the optimal scenario was achieved at iteration 73 with 78.34% accuracy (Figure 3.7d), and the corresponding parameters were found as 250 and 14 for the number of trees and the maximum depth, respectively.

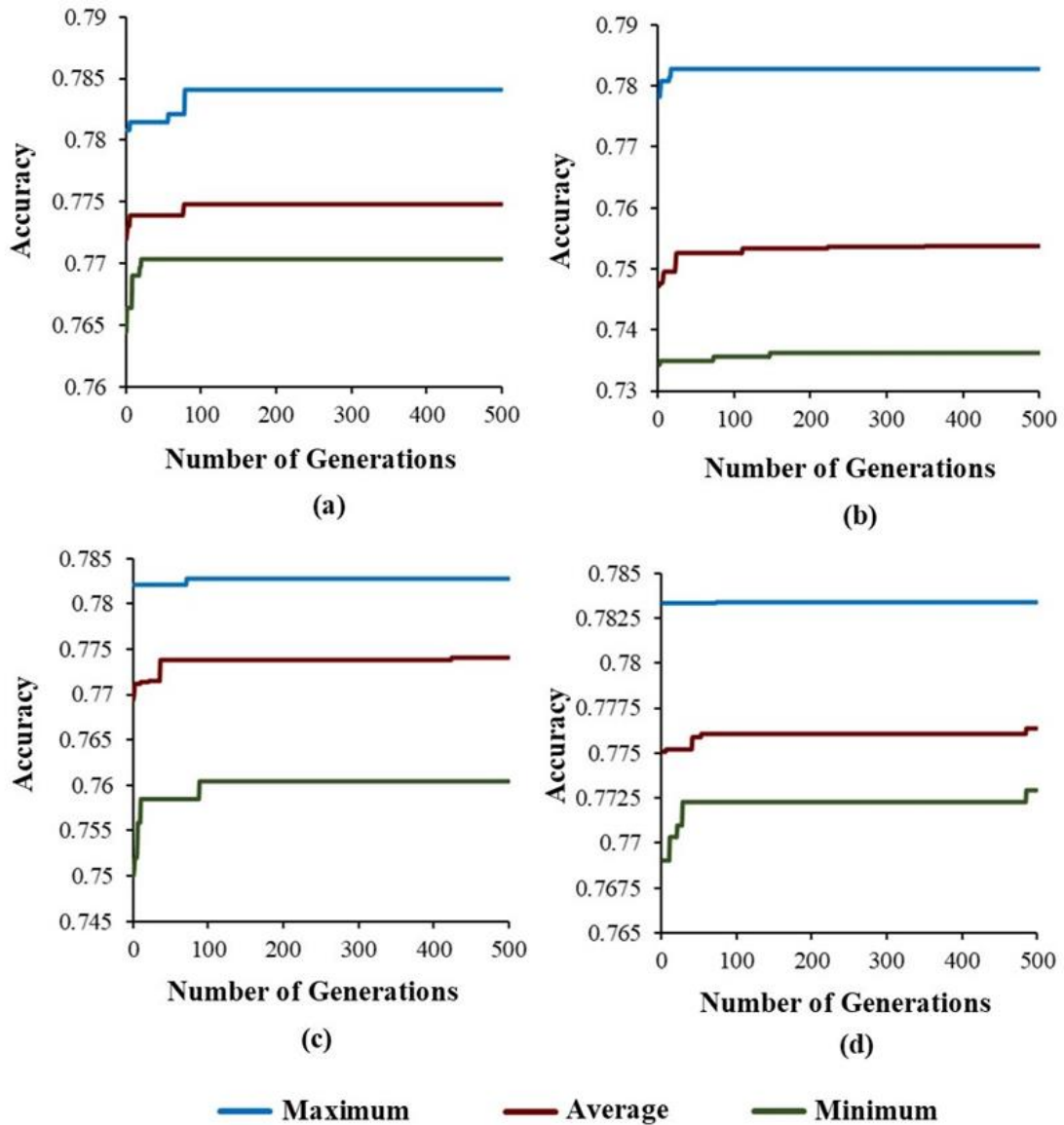


Figure 3.7 Convergence graph of genetic algorithm for (a) Random Forest, (b) XGBoost, (c) AdaBoost, (d) Extra Trees.

Table 3.5 Comparative results of machine learning methods

Model	Parameters	Parameter ranges	Best parameters	Total combination
RF	N of trees	50-1000; step = 50	950	1,000
	Maximum depth	1-50; step = 1	9	
	Minimum N of samples	fixed	10	
XGBoost	Learning rate	0.01-0.50; step = 0.01	0.1	1,000
	N of trees	50-1000; step = 50	50	
	Minimum child weight	fixed	10	
AdaBoost	N of trees	50-1000; step = 50	250	1,000
	Learning rate	0.01-0.50; step = 0.01	0.11	
ET	N of trees	50-1000; step = 50	250	1,000
	Maximum depth	1-50; step = 1	14	
	Minimum N of samples	fixed	4	

3.4.3 Classification Results

This study compares the performances of four well-known tree-based ensemble ML algorithms integrated with genetic algorithms for parameter optimization. Performance evaluation scores of each classifier were measured with the remaining 30% of the data (testing set). TP, FN, TN, FP, accuracy, precision, and F1 score of ensemble models are provided in Table 3.5. The confusion matrix developed for each tree-based classifier with respect to recall values is also shown in Figure 3.8. The findings highlighted that the ET algorithm showed the best performance in classifying PD workers, with a recall value of 69.03% (Figure 3.8), through 214 correctly classified instances out of 310 (Table 3.5). However, it showed the worst performance in classifying N-PD workers with 2,599 FP instances out of 14,072, accounting for a recall rate of 81.53%. In this respect, XGBoost presented the highest recall rate, 83.25%, by classifying 11,715 instances correctly, while it achieved the worst performance with respect to PD workers (67.741%). Still, XGBoost presented relatively higher prediction power in terms of accuracy (0.8292) and F1 score (0.8887) compared to other ML models (Table 3.5). The results implied that XGBoost could detect 67.74% of all PD workers (Figure 3.8) by focusing on only 17.85% of the

accidents $((210+2,357)/(310+14,072))$ (Table 3.5). Although ET could aid in preventing 69.03% of all accidents that resulted in PD workers, it required a careful examination of 19.56% of all accidents $((214+2,599)/(310+14,072))$. In terms of precision values, each classifier's performance was nearly the same, in which AdaBoost was the most powerful classifier (0.9721).

Table 3.5 Summary of evaluated test scores

Models	PD (Support: 310)		N-PD (Support: 14,072)		Performance metrics (average)		
	TP	FN	TN	FP	Accuracy	Precision	F1 score
GA-RF	211	99	11,526	2,546	0.8161	0.9718	0.8807
GA-XGBoost	210	100	11,715	2,357	0.8292	0.9719	0.8887
GA-AdaBoost	212	98	11,680	2,392	0.8269	0.9721	0.8873
GA-ET	214	96	11,473	2,599	0.8126	0.9720	0.8786

All these classification metrics could be considered as meaningful outcomes. However, AUROC is regarded to be a better alternative and has widely been used in case of imbalanced class distribution by many researchers [2], [11], [92], [94], [109], [110]. Hence, ROC curves of ensemble models based on the best parameters captured from GA were computed under an untrained testing set (Figure 3.9). The findings showed variations in the performance of four classifiers, ranging between 0.8012 and 0.8120. According to analysis results, XGBoost presented the highest prediction rate with an AUROC value of 0.8120, followed by ET and RF, by 0.8110 and 0.8094, respectively. On the other hand, AdaBoost showed the lowest prediction rate with an AUROC value of 0.8012. Since XGBoost was also the best performing model in terms of accuracy and F1 score, it was determined to be the most suitable tree-based ensemble model in the overall scheme.

		<i>Random Forest (RF)</i>		<i>XGBoost (XGB)</i>		<i>AdaBoost (AB)</i>		<i>Extra Trees (ET)</i>	
True class	RF	PD	N-PD	PD	N-PD	PD	N-PD	PD	N-PD
		68.06%	31.94%						
	N-PD	18.09%	81.91%						
	XGB	PD	N-PD	PD	N-PD	PD	N-PD	PD	N-PD
AB	ET	67.74%	32.26%	68.39%	31.61%	69.03%	30.97%	69.03%	30.97%
		16.75%	83.25%	17.00%	83.00%	18.47%	81.53%	18.47%	81.53%

Figure 3.8 Confusion matrix for ML models

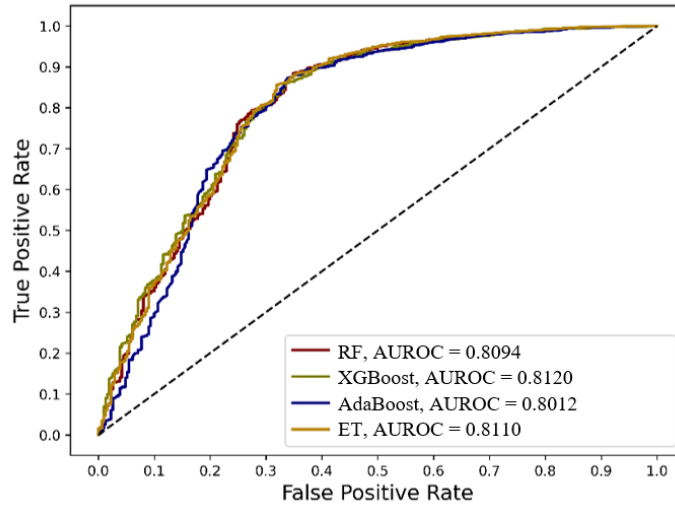


Figure 3.9 ROC curve of ensemble models

3.4.4 Feature Importance

The importance of each principal component used in the prediction framework was obtained through the Gain importance criteria, which shows the improvements in accuracy that an attribute provides to corresponding branches widely. The results of the feature importance analysis revealed that the most crucial attribute was Component 1 with a value of 0.5272 over 1, while the second most influential attribute was Component 4 with a value of 0.2578. Figure 3.6 shows that “working days lost” had one of the highest contributions to explain Component 1. In addition, the last and next to the last members of the “type of injury” variable had relatively higher direct and inverse proportional effects in explaining Component 1, respectively. It should be noted that the first denotes the injuries related to “shock due to assault and threat” and the latter represents the injuries related to “exposure to extreme temperatures”. Hence, it would be correct to state that

workers are likely to experience permanent disability after construction accidents if they are exposed to extreme temperatures. Contrarily, there is a relatively lower probability of workers experiencing permanent disability in case of occupational accidents related to shock due to assault and threat. In addition, Component 4 was mostly explained by the fifth instance of “type of injury” and the fourth instance of the “injured body part” variable, which were “exposure to chemical substances” and “arm”, respectively. Hence, workers practicing close to chemical substances have a higher probability of experiencing permanent disability. The findings on injured body parts also revealed that workers who had had an arm injury were more likely to become permanently disabled.

3.4.5 Contributions of the Study

This study aims to provide insights to improve occupational safety management, productivity planning, input availability for enhanced decision-making, early payment planning, a safer working environment, and policy development for practitioners in the construction industry and policymakers from governmental authorities. Suggested contributions and a guide for implementing the proposed model are illustrated in Figure 3.10 and explained in the following sub-sections.

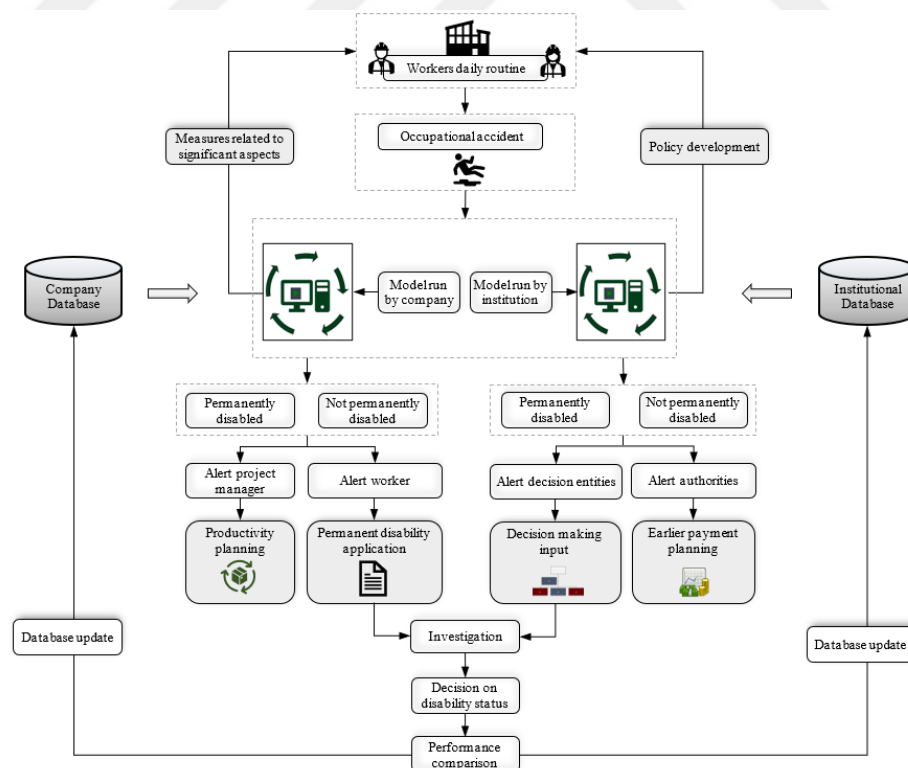


Figure 3.10 Guide for practitioners on model implementation

3.4.5.1 Contribution to Practice

This study approaches safety management based on the accidents leading to permanent disability for workers in construction projects. An integrated GA-tree-based ensemble ML method explicitly contributed to the examination of permanent disability. Evidence suggests that high accident rates in the construction industry lead to loss of productivity [111]. Here, two different productivity losses can be observed from the disabled worker: productivity loss of the injured worker and productivity loss due to worker replacement [112]. The proposed model can be used for the early planning of productivity loss in the projects by safety management departments of construction companies and project managers. The outcomes of the study could give insights on the potential workforce loss to control productivity in the projects. In addition, Mohd Kamar et al. [80] address that accident costs are strongly associated with disability rates. Hence, by predicting the disability status of construction workers, employers can also anticipate the effect of accidents on the cost of the projects.

Analysis results showed that the integrated use of feature engineering, tree-based machine learning methods, and state-of-the-art optimization techniques could provide efficient and accurate prediction models. The proposed model showed that the ET algorithm could help to detect cases with permanent disability since it predicted the highest number of true positive cases. In addition, the XGBoost algorithm can highly function in estimating the disability of construction workers after construction injuries since it achieved the highest prediction rate with respect to the performance evaluation indicators such as accuracy, F1 score, and AUROC. The findings suggest that using the XGBoost-based prediction model, safety professionals can detect 67.74% of permanent disability cases by focusing on only 17.85% of the recorded accidents.

This study also exhibited the influential variables and underlying causes leading to permanent disabilities in construction projects by attribute importance. This part of the proposed model could contribute to the improvement of safety management by taking necessary measures proactively for the prevention of accidents resulting in permanent disability. In this sense, safety management professionals could be advised to focus on some of the key variables to enhance occupational health and safety policies, contributing to the minimization of permanent disability incidents in the construction industry. One of the important findings of this study is that “*working days lost*” was found among the most influential attributes in detecting permanent disability. According to Social Security

Institution [21], “working days lost” was higher when workers faced incidents such as falls, struck by a falling object, arm (or hand or finger) being torn or cut off. Thus, workers experiencing these accidents are more likely to lose their earning capacity, resulting in permanent disability for the workers and productivity loss for the projects. In this context, “*lack of workers’ awareness*” and “*fall protection*” are the primary reasons for falls from height [113]. Therefore, safety management departments are suggested to organize periodic training sessions for workers working at height more frequently and be careful about the use of personal protective equipment to minimize fall-related accidents.

One of the other issues regarding “*working days lost*” was found as “*struck by a falling object*”. Since the probability to be struck by a falling object is very high while working with or near cranes, dropped object protection systems can be suggested to reduce workers’ permanent disability risk, even up to 50% [114]. Considering injured body parts, the “*arm*” was identified as the most vulnerable bodily organ. Yan et al. [115] address that physically demanding tasks, especially those that are performed outdoor, could cause permanent disability. Here, the arms of workers are often vulnerable due to demanding tasks and material handling with no protective equipment. It is important to note that one of the major reasons for not using personal protection equipment highlighted by construction workers in Turkey was the protective gear being uncomfortable [116]. Thus, the comfortability factor should be considered in developing protective gear regarding arms to minimize the possibility of accidents resulting in disabilities for construction workers. In addition, bit extension shafts can also be used in construction sites to protect the arms of the workers by reducing stresses to screw hard-to-reach components [117].

Findings also show that “*type of injury*” plays a determining role in predicting the permanent disability of construction workers. Different injury types could have divergent consequences in workers’ lives [7], and the most significant types of permanent disabilities were determined as “*exposure to extreme temperatures*” and “*exposure to chemical substances*”. According to the injury type classification of Robert et al. [118], the injuries originated from extreme temperatures are classified as severe and likely to cause permanent disabilities. Chi and Han [119] address that the prevailing human errors related to extreme temperature are the misjudgment and malfunction in securing or warning operation. Also, lack of respiratory protection and engineering control, besides the misjudgment, are common human errors reported for chemical exposure. It is important to note that exposure to extreme temperatures and chemical substances are not

frequently encountered injury types in construction sites [120]. Thus, safety measures should be taken at the maximum level, and engineering supervision is essential before and during these works for effective risk response. In this context, a proper place could be allocated for the works that might lead workers to expose chemical substances or extreme temperatures in safety professionals' agenda.

Workers are the key stakeholders of safety management. According to the facts mentioned above, the fundamental requirement for a higher level of safety in construction projects is the protective gear used by the workers. In this regard, periodic control and maintenance should be performed, particularly for protective gears and equipment. The study by Nath et al. [121] emphasizes the advantages of real-time personal protection equipment detection, which construction companies can adopt and use on building sites. Workers in the Turkish construction sector stated that they would wear personal protective equipment if adequate safety training and rewards were provided regularly [116]. These points were stated as crucial for the workers to embrace personal protective equipment voluntarily. These suggestions could be shared with workers (or their representatives) during safety meetings for more effective management [122]. By considering the impact of protective gear on accident outcomes, collecting related data besides other requirements could improve the investigations regarding the underlying causes of permanent disabilities.

3.4.5.2 Suggestions for Policymakers

Permanent disability of workers is determined by the “Working Power and Professional Gaining Power Loss Rate Determination Procedures Regulation” [123] according to the “Social Insurance and General Health Insurance” law (Law no: 5510) [124] in Turkey. A worker can be regarded as “permanently disabled” if the worker loses at least 10% of its earning capacity. After an occupational accident leading to a permanent disorder, the worker can apply “permanent disability income” with a set of documents, such as the disability application form, health report of employment, occupational accident report, and medical report after an accident. Then, Social Security Institution evaluates a worker's disability based on several aspects such as type of injury, injured body part, the amount of body part injured, type of work, and medical report. The procedure to determine the disability of workers usually takes some time. Thus, after the occupational accident or permanent disability application of a worker, related institutions can run the proposed model by incorporating the information of the worker into the testing phase of

the model, instead of the testing set used in this study. Hence, the proposed prediction model can be used to predict whether the workers would be paid permanent disability income or not by the related institutions. Considering that the proposed ML model could provide early detection of the disability status of workers, potential expenditures regarding compensation payments for workers can be pre-planned.

There are some limitations due to the data collection method regarding construction accidents in Turkey [78]. For instance, according to the dataset used in this study, there were only seven injury types selected by construction companies to define the type of injury as a result of occupational accidents. In addition, factors such as the type of accident could not be used due to the significant amount of missing information. Besides, some other aspects of the accidents, such as safety gear used by the workers, site alerts about hazardous conditions, and type of work, were not available in the dataset, which could have contributed to a better and more extensive representation of the working environment. Thus, registration of these attributes is recommended while recording occupational accidents. A safety management training program for safety staff and project managers could be improved by not only focusing on the most hazardous working conditions and personal protection equipment applications but also emphasizing the importance of proper data collection.

Measures taken in construction works are identified in Occupational Health and Safety Regulation in Construction Works [19], based on Occupational Health and Safety law (Law no: 6331) [17]. According to the regulation, the minimum working conditions for working at height must provide some measures, including the existence of fall protection platforms, scaffold protection, full-body harnesses. Horizontal and vertical lifelines that would be used during the performance of such works must be considered in the project planning phase as a part of safety management and included in the health and safety plan. Similarly, regarding the falling objects, the regulation addresses that entrance should be curtailed or closed passages should be created where there is a risk of falling objects. In addition, workers must wear hard hats, and this should be controlled by the employers. However, no precise statements regarding working with extreme temperatures or chemicals and adequate arm protection in construction are expressed, which should be considered by policymakers to minimize workers' disability.

Besides all, one of the primary reasons for poor workplace safety in many developing countries is the lack of enforcement despite its existence [122], [125], [126]. Thus,

governmental authorities could be suggested to increase monitoring and controlling of safety rules in construction sites. Another applicable approach would be combining safety promotion with safety enforcement, as this approach showed significant improvement in accident rates in Hong Kong [122]. This might also be extended to proper data recording since the recorded data is usually of low quality, such that missing data or “other” labeled expressions are observed for different attributes in many instances. By implementing the suggestions made according to the analysis results, the permanent disability of workers can be reduced in construction sites.

3.5 Conclusions and Future Work

In this study, a machine-learning-based permanent disability prediction model was developed based on 47,938 recorded construction accident data. A detailed multi-step feature engineering approach was carried out in the pre-processing stage, including data encoding, data scaling, dimension reduction, and data resampling to code categorical features as numerically, to transform numerical features into the same scale, improve the efficiency of the model, and highlight the imbalanced data problem often encountered in engineering practices, respectively. State-of-the-art optimization method, a genetic algorithm was used for parameter tuning of adopted tree-based ensemble machine learning models, i.e., RF, XGBoost, AdaBoost, and ET. The proposed genetic algorithm-based machine-learning model showed a prediction rate of 0.8292 with respect to accuracy and 0.8120 with respect to AUROC when the model was tested through the XGBoost algorithm. In addition, feature importance results highlighted that workers exposed to extreme temperatures and chemical substances have higher probabilities of experiencing permanent disabilities.

This study is one of the first studies conducted to predict the outcomes of construction accidents with respect to the permanent disability of construction workers with an imbalanced dataset. Further investigations, such as analyzing the disability levels of workers, can explore the severity of worker disability and underpinning the potential causes for high-level disabilities. This study used some attributes that are post-accident such as working days lost, type of injury, and injured body parts. On the other hand, some pre-accident attributes were not included in this study, such as safety gear used by the workers, site alerts about hazardous conditions, type of work, age, gender, and tenure. Therefore, based on the availability of the data, interested researchers can investigate the effects of additional pre-accident attributes to predict outcomes of construction accidents

for the development of proactive accident prevention strategies. In a future study, the authors plan to expand the scope of the current study to predict accident severity levels by using proactive data.

The methodological approach used in this study is expected to contribute to the body of knowledge in construction safety management such that interested researchers can enhance the methodology underneath. This study adopted one of the most advanced optimization techniques, genetic algorithm, for parameter tuning. Further investigations, including comparing other optimization techniques such as bees, particle swarm, and grey wolf algorithms, could contribute a lot to the body of knowledge. There are also several algorithms to address class imbalance problems, dimension reduction, and data scaling. Thus, comparing several techniques for data resampling, dimension reduction, and data scaling could be considered in future studies. This study is also expected to support the development of advanced safety management strategies in construction projects through regulations.

ACCIDENT PREDICTION IN CONSTRUCTION USING HYBRID WAVELET-MACHINE LEARNING

This chapter was published in *Automation in Construction*, vol. 133, p. 103987 as a research article.

4.1 Introduction

One of the major concerns of the construction industry is workplace safety. It is among the most demanding working environments due to existence of poor physical working conditions, various stressors, and excessive time pressures. Therefore, construction workers are more prone to fatalities and injuries compared to workers in other industries [127]. Evidence present three to four times higher probability to experience fatality in developed countries, while it is three to six times in less developed countries compared to other industries [2]. Every year, approximately 100,000 workers lose their lives in the construction industry, accounting for nearly 35% of the global occupational fatalities [3]. The sector is one of the most dangerous sectors in almost every country with respect to fatal and non-fatal occupational accidents.

According to the Turkish Statistical Institute (TSI) [128], the number of workers in the construction industry in Turkey was 1.6 million accounting 5.6% of total number of employments in 2019 and more than 165,000 were registered as licensed safety professionals [129]. Construction industry solely accounted for approximately 32% of the total fatal occupational accidents with a total number of 49,710 accidents (including fatal and non-fatal) in 2019 [21]. Despite ongoing improvements in the safety management systems across the country, occupational safety in the construction industry is still in need of more progress. For this reason, analyzing past data to predict the number of future accidents could assist planning of necessary actions and safety measures to prevent potential fatalities and injuries in the future [4].

The existing literature suggests the time series analysis for future predictions, such that a time series is a sequence of data points, measured on the basis of time instants spaced at uniform time ranges or intervals [130]. On the one hand, stand-alone machine learning

(ML) methods were initially adopted for the projection of future values through time series in various field of researches. On the other hand, different decomposition methods, such as Fourier transform, wavelet decomposition, and empirical mode decomposition have gained considerable momentum in the last decade. These methods commit to investigate the original time series with respect to different frequencies for improving the accuracy of the models and provide a ground to assess the changes in time series at various scales [60]. Several researchers used time series analysis in other sectors to make projections. However, research efforts in the construction safety literature are limited in this context. This study aims to contribute to the body of knowledge of construction safety by coupling discrete wavelet decomposition and different ML methods to predict daily occupational accidents for different time periods. Results obtained from the study are expected to give insights to managers, decision makers, and inspection authorities in the development of necessary strategies to minimize the adversarial impacts of accidents through occupational health and safety (OHS) management practices.

4.2 Research Objective and Novelty

The objective of this research is to develop a prediction model of construction accidents for specific time periods using time series analysis aided by wavelet decomposition pre-processing technique. To perform a comprehensive evaluation, artificial neural network (ANN), multi-variate adaptive regression splines (MARS) and support vector regression (SVR), were employed to predict the number of daily occupational accidents, including injuries and casualties. Discrete wavelet transform (DWT), which is one of the commonly used pre-processing techniques, was adopted with stand-alone ML methods to increase the accuracy of the predictions. To achieve this goal, the original time series of daily number of occupational accidents was divided into different sub-bands by DWT. The dynamics of occupational accident frequencies were investigated in detail by creating a comparative predictive model based on three stand-alone ML algorithms. In hybrid modeling phase, the daily accident dataset was subjected to pre-processing along with DWT, and ML algorithms were employed to each decomposed band separately to improve model accuracies. A dataset consisted of 393,160 occupational accidents, which were recorded between 2012 and 2020 in Turkey, was analyzed to predict the number of accidents in short-term, mid-term and long-term time periods.

Recent studies in occupational accident in the construction industry have mainly focused on construction accident outcomes for workers individually based on predefined

attributes [2], [11], [39] or text mining subjected to accident narratives [4], [14], [97]. This study approached the occupational health and safety management in the industry from a different perspective, aiming to fill the gap in the pertinent literature in two ways. The first remark is that a dynamic national utilization plan is proposed based on hybrid DWT-ML analysis for the early detection of time periods that are expected to cause a high number of accidents in the future. The second one is related to the sample size of the data, such that 393,160 occupational accidents occurred in Turkey nearly over a decade are statistically investigated. The high number of recorded accidents enabled the use of DWT for the early detection of accident trends in the construction industry to improve safety performance of the industry with generic and proactive management procedures.

4.3 Background of Used Methods and Related Implementations in Construction Safety Literature

4.3.1 Time Series Analysis Literature

Time series present the variations specific to a particular time period, often with a seasonal or increasing and decreasing trends through a set of observations. Comprehending and predicting time series related to a variable have been carried out in many research fields stemming from the necessities for preliminary preparations of contingencies. A large and growing body of literature investigated the predictions of future values by using previously observed time series. As in every prediction studies, a user needs to define the input(s) and output(s) in the time series predictions. Extensive research has explored that the autocorrelation and partial autocorrelation functions that illustrate the relationship among the current value of a variable and its past observations are used to identify the input(s) of the prediction framework. To perform time series predictions, various statistical methods, such as regression and autoregressive integrated moving average (ARIMA) were initially employed by the scholars [131]. Then, ML methods were also introduced in addition to the statistical approaches to improve prediction accuracies recently [132]. ML methods are usually used as both stand-alone and hybridized with decomposition techniques that enable modeling trends, seasonality and cycling effects, which can be considered as the main patterns of time series [133]. The previous research findings show the variety of decomposition techniques, such as wavelet decomposition, variational mode decomposition and empirical mode decomposition, representing the features of original time series data set.

Predictions based on the time series decomposition techniques are not only accepted widely in various core research fields such as health, finance, energy and agriculture, but also performed in many disciplines of civil engineering such as structural, transportation, geotechnical and water resources engineering. For instance, Adeli and Jiang [134] coupled wavelet transform (WT) with neural networks based on two objectives, which are identifying the behavior of structures during earthquakes and improving the accuracy of structural system identification under extreme loads. Chen et al. [135] used empirical mode decomposition by hybridizing with the support vector machines to enhance the accuracy of traffic network predictions. In geotechnical engineering domain, Ding et al. [136] utilized the discrete wavelet transform to predict tunneling-induced ground settlement. Also, variational mode decomposition and WT are used widely in water resources domain for varying objectives such as predicting the monthly runoff forecast by He et al. [137] and projecting short and mid-term droughts by Özger et al. [60].

Despite a vast literature on other domains, construction accidents have not been investigated properly with time series predictive models. Although frequent implementations are reported in accident forecasting, a comprehensive literature survey showed that time series analysis has been used in few studies in construction safety. As a limited attempt, Marhavilas et al. [130] developed a risk assessment framework by using the harmonic analysis of time series of occupational accidents (based on Fourier superposition) and applied the model on electric power (for 1993-2009) and construction industry (for 1999-2007). The analysis, which was based on data with 1-month time resolution, showed that accidents could be characterized by the periodic factor, which constituted a permanent element for the OHS systems. Chen et al. [138] developed a prototype electroencephalography (EEG) safety helmet that allowed the collection of the neural data to estimate construction safety via mental workload. The EEG signal data was multidimensional including frequency, power, and phase data in addition to time series data, and used to reduce unexpected injuries based on mental conditions of workers. Zhou et al. [139] used complex network theory to characterize the time series of near miss accidents based on 1,124 time series accident data from a metro construction in Wuhan, China between 2011 and 2015. They found that inter-event time series of near-miss accidents were free of scale and occurred in a complex system with sharp fluctuations. The past literature shows that the studies related to use of time series in construction safety is more about time series investigation rather than prediction, such that neither

considered ML methods in time series analysis for future predictions, nor integrated wavelet decomposition technique to improve the accuracy of projections. This study contains the application of discrete wavelet transform for dividing the original time series of the number of daily nationwide accidents recorded in the construction industry into its spectral bands. Hence, with the methodology adopted in this study, an implementation of hybrid soft computing prediction framework in construction safety research is presented to bring a new perspective to the literature.

4.3.2 Wavelet Transform

Wavelet analysis is considered complementary to existing techniques, such as correlation and spectral analysis, in which the fundamental philosophy of the wavelet transforms (WT) is to divide a signal into different scale components [60]. WT operates similar to Fourier transform (FT), however, the basic elements of time series, such as trend, seasonality and cycling, are usually lost in FT [140]. The original time series is decomposed into consecutive sub-series in WT, and during this process, the main wavelength is expressed by shifting at a certain scale and time. All features of the original signal are consummately transferred to a sequential data. High and low frequency transitions that are included in signals can also be detected by means of the WT based on its time localized feature [141]. This feature secures the determination of all local and regional characteristics of the signal, enhancing the accuracy of predictions.

Existing literature highlights two different WT types: (1) continuous wavelet transform (CWT) and (2) discrete wavelet transform (DWT). In this research, the latter was used based on its time efficient feature and applicability to all type of data [142]. DWT is also defined more useful than CWT by not allowing the generation of excessively redundant data resulting from the calculation of wavelet coefficients in all possible scales [143]. Mallat [144] indicated a settled down approach in the determination of the scale and position parameters as dyadic scheme. In DWT, the original time series can be defined as the aggregation of several sub-series, such as details and approximations. Details (D) represent the high frequency bands, while the approximations (A) denote the low frequency bands. The number of details and level of approximation indicate the decomposition level of the original series, such that at each level, component A is divided into new A and D components [145]. It can also be noted that there is no strict rule in the literature for the determination of optimum level of decomposition. Despite its frequent use in other domains, there is a gap in the construction safety literature regarding time

series analysis with wavelet decomposition. In this study, DWT was used to decompose the original time series data of construction accidents into its sub-bands. This process enabled improvement in the accuracy of future projections and development of a more robust proactive risk management framework. In addition, the dynamics of occupational accident frequencies could be observed in detail with the decomposition of the original time series data into its sub-bands through DWT.

4.3.3 Machine Learning Background

In this study, ANN, MARS and SVR were used as non-parametric supervised machine learning algorithms to project daily occupational accidents likely to occur in different time periods. Each of these methods has different working principles and differs from each other in terms of several aspects. Even though these are commonly used algorithms in the ML literature, they also have different pros and cons such as the basis on which they functionalize, approaches they signify and their uses. Table 4.1 presents the comparative overview of three adopted ML methods. Their brief explanations and utilizations for the construction safety management can be found in the following.

4.3.3.1 Artificial Neural Networks

Artificial neural networks (ANNs) are formed by the combination of cells as in biological neural systems, while in mathematically, represents a dynamic system that can be modeled as a combined set of differential equations [150]. The smallest entity of an ANN is artificial neurons; in fact an ANN was initially suggested as a single layer neuron system. However, the multi-layer neural networks, such as multi-layer perceptron (MLP), were proposed later due to the challenges of the single layer structure in solving non-linear problems [151]. The multi-layer neural networks consist of minimum three layers: input layer, hidden layers, and output layer. The number of cells in the input and output layers varies depending on the information described in the problem [152]. The application of ANN is performed by two types of calculations: (1) Feed Forward and (2) Back Propagation. In the first step, the weights that are defined as the values that express the effect of input set are considered randomly, and the network's outputs are obtained for the given sample. In the second step, the weights are adjusted by introducing different samples to the network to find the optimum weights. In the model, where the sum of error squares is expected to converge to zero, the weights are updated through the backpropagation algorithm [153].

Table 4.1 Overview of the used machine learning algorithms

Methods	Basis	Approach	Ease of use	Strengths	Weaknesses
ANN	Non-linear functional approximation	Black-box	Moderate / Difficult	It provides de facto solution for non-linear regression problems [146]	It is prone to overfitting when the data is too noisy
				It has fast learning and reasonable generalization ability [147]	It provides less interpretability due to its black box structure
				It is a universal method by providing satisfactory results in a great deal of research domain	It is not capable of identifying the significant independent variables [147]
MARS	Knot-based additive model	Explicit mathematical representation	Easy	It requires a smaller number of parameters to be adjusted and has fast processing speed	It has potential to fail in detecting interaction effects if the basis function contains a pure high order interaction
				It provides importance of independent variables and interpretability among independent and dependent variables	It requires large enough dataset to make use of it worthwhile
				It has the capacity to mapping complex and high dimensional data sets [148]	It is characterized by a tendency to overfit through an unmodified generalized cross validation criterion when the spurious models cause a large bias
SVR	Decision boundary	Black-box	Moderate	It is appropriate for the dataset that has smaller sample size	It has less generalization ability when the sample size is lower than the number of attributes
				It provides the utilization of several kernel functions to decide non-linear boundaries	It requires extensive knowledge on the kernel functions [149]
				It has the ability to resist overfitting and provides enhanced generalization	It provides less understandability on its structure

There are several applications of ANN method in the construction safety literature. For instance, Ayhan and Tokdemir [63] investigated construction accident outcomes with respect to seven target attributes such as at risk behavior, near miss, incident with partial failure, first aid, medical intervention, lost workday cases, and fatality. They used conventional regression and ANN (with 13 different networks), and then implemented fuzzy inference system. The researchers found ANN producing better predictions regarding more severe accidents. Antwi-Afari et al. [39] used wearable insole pressure system to classify awkward working postures using several ML methods such as ANN, Decision Tree (DT), K-Nearest Neighbor (KNN), and SVM. From another perspective, Nath et al. [121] used neural network and DT algorithms to verify whether the workers are using personal protective equipment detected through three deep learning algorithms. The structure of the neural network included one input layer (including 4 features), two hidden layers (each comprising 8 nodes), and one output layer for a binary classification. The dataset containing 1,500 annotated images and 4,700 worker instances showed that the single convolutional neural network framework achieved the best performance. Apart from the construction safety management, other researchers used ANN either for safety predictions in other industries or other sub-topics of construction management such as construction cost prediction [154] and outcomes of construction claims [155]. Hence, the wide application of the method in different research areas is the primary reason to examine its performance in the time series analysis for the future projections of the number of occupational accidents in the construction industry.

4.3.3.2 Multivariate Adaptive Regression Splines

Multivariate Adaptive Regression Splines (MARS), which was introduced by Friedman [156], is a nonparametric method that does not rely upon the assumptions on the functional relationship between dependent and independent variables. Instead, the MARS algorithm creates a dynamic relationship among the variables based on basis functions extracted from the regression data. A flexible regression model is built by using basis functions that correspond to different ranges of independent variables. MARS can be deemed as a powerful algorithm, as accurate results can be obtained even if there are complex relationships (such as nonmonotonic) between dependent and independent variables. Multicollinearity, which is the interaction of independent variables with each other, is a counterwork in regression analysis due to the fact that a single function is utilized to define the relationship between variables [157]. However, it does not pose a

challenge in MARS method since it employs a piecewise polynomial function to determine the basis functions. The MARS model has two fundamental steps, forward selection and backward elimination. In the first step, a knot is described for each value for every regression slope changes, such as an end of one domain in the dataset, and so the beginning of the other [158]. In the second step, significant independent variables and mutual interactions among these variables are determined. Also, less important variables are pruned from the MARS model and the model, which satisfies the objective function (the sum of error squares as default) is defined as the most appropriate model.

Despite its several advantages, there is a gap in the construction safety literature about the application of the MARS method. Even the utilization of the method in any of the sub-disciplines of construction is limited. As an example of the very few attempts, Cao et al. [159] proposed a hybrid forecasting model that analyzed the effects of attributes to predict construction cost index in Taiwan. In another study, Chen and Yang [160] calculated an optimum surrogate model according to the dataset simulated with EnergyPlus, using MARS, multiple linear regression and SVM. Applications for a case study of a passively designed high-rise residential building were performed through several ventilation and thermal load scenarios. Guo et al. [161] analyzed the applicability of MARS and deep forest for the rockburst risk assessment, and compared their performances with some baseline models such as Random Forest (RF), eXtreme Gradient Boosting (XGBoost) and ANN. In safety domain, Sánchez et al. [162] used MARS algorithm for a dimension reduction before the application of SVM method to predict occupational accidents. The literature on MARS application shows that the use of MARS is highly limited, particularly regarding construction safety domain.

4.3.3.3 Support Vector Machine

Support vector machine (SVM) is based on the statistical learning theory and structural risk minimization principle [163]. SVM method is described as the universal structured learning process that can be used to learn various projections, such as neural networks (sigmoid activation), radial based functions, splines and polynomial predictors. In SVM, the class of data points is determined by means of a line, which is called as hyperplane. For non-linear problems, a hyperplane cannot be obtained without an error, such that the error is allowed during the entire process and the hyperplane that gives the lowest error is derived. Therefore the hyperplane, which maximizes the margin and minimizes the error is called optimal hyperplane [164]. Considering the fact that SVM cannot produce

a linear hyperplane for a nonlinear dataset, existing research recognizes the critical role played by kernel tricks that are used for non-linear problems in improving the learning ability of algorithm [165]. Several functions are proposed as kernels, such as radial basis function (RBF), polynomial function and sigmoid function. The selection of the kernel function has a pivotal role on the performance of SVM. Among the functions used for non-linear problems, RBF represents a much larger function space than other kernel functions. With its practical and flexible nature, the RBF is able to establish all relationships swiftly [166]. Also, the RBF kernel exhibits not only a fast convergence but also better generalization ability [167]. These facts make RBF kernel a universal function and it was used in this research based on these advantages.

Owing to its several advantages, SVR (particularly with RBF kernel) has widely been used in the literature, including construction safety management. In this context, Zhang [168] analyzed construction occupational accident classification using Word2Vec skip-gram model, which is mainly used for learning word embedding, and developed a hybrid deep neural network structure with respect to text mining. The researcher compared the proposed model with five commonly used ML methods: Linear Regression (LR), SVM (with RBF kernel), DT, naïve bayes (NB), and KNN. They concluded that the proposed model constituted with 1,280 cases had a substantial effect on the classification results based on several target output elements, such as falls, fires, traffic, exposure to chemical substance. Goh and Ubeynarayana [4] performed narrative classification of occupational accidents based on text mining techniques using 1,000 construction accident narratives. They evaluated six ML methods as SVM, LR, RF, KNN, DT, and NB, and found that the best performance was produced with both linear and RBF SVM models. From another perspective, Li et al. [169] tested several ML methods such as SVM, DT, KNN, boosted tree, and linear discriminant analysis for the mental fatigue classification of construction equipment operators, and found that SVM was the most suitable ML method in the field. Some of the other studies that adopted SVM method in the construction safety literature can be summarized as Cheng et al. [97] (for accident classification), Baker et al. [14] (for injury severity, injury type, impacted body part, and accident type), Zhang et al. [170] (for causes of accident). Hence, its wide and successful applications in the construction safety literature are the main reasons to use SVM method in this study. It is also important to note that despite SVM being used in both classification (SVC) and regression (SVR)

problems, majority of the studies in the construction safety literature considered SVC method, while SVR method (for a regression problem) is used in this study.

4.4 Research Framework

This research consists of comprehensive evaluation of different ML methods for short-term, mid-term and long-term predictions of the frequency of the occurrences of occupational accidents in the construction industry on a daily basis. To accomplish this goal, three different ML algorithms are employed and coupled with the DWT, which is one of the most effective tools in the decomposition of time series. As a first step, the lag-times that are used as input were obtained through the partial autocorrelation functions. Then, predictions were performed by means of both the stand-alone ML methods, ANN, MARS and SVR, and hybrid wavelet ML methods, abbreviated as W-ANN, W-MARS and W-SVR. Determination of the optimum level of decomposition for the hybrid wavelet predictions was diligently considered by means of a trial-error approach. The entire dataset was divided into two major groups as training and testing with the percentages of 70% (January 2012-February 2018) and 30% (March 2018-October 2020), respectively. Furthermore, the number of occupational accidents was predicted for three different lead-times (1-day, 7-day and 30-day ahead), while the results were also investigated by using different performance metrics to exhibit the statistical significance of the predictions. The flowchart of the analysis procedure is shown in Figure 4.1. Accordingly, a dynamic utilization plan based on the proposed prediction model is presented to aid in the development of national OHS management procedures.

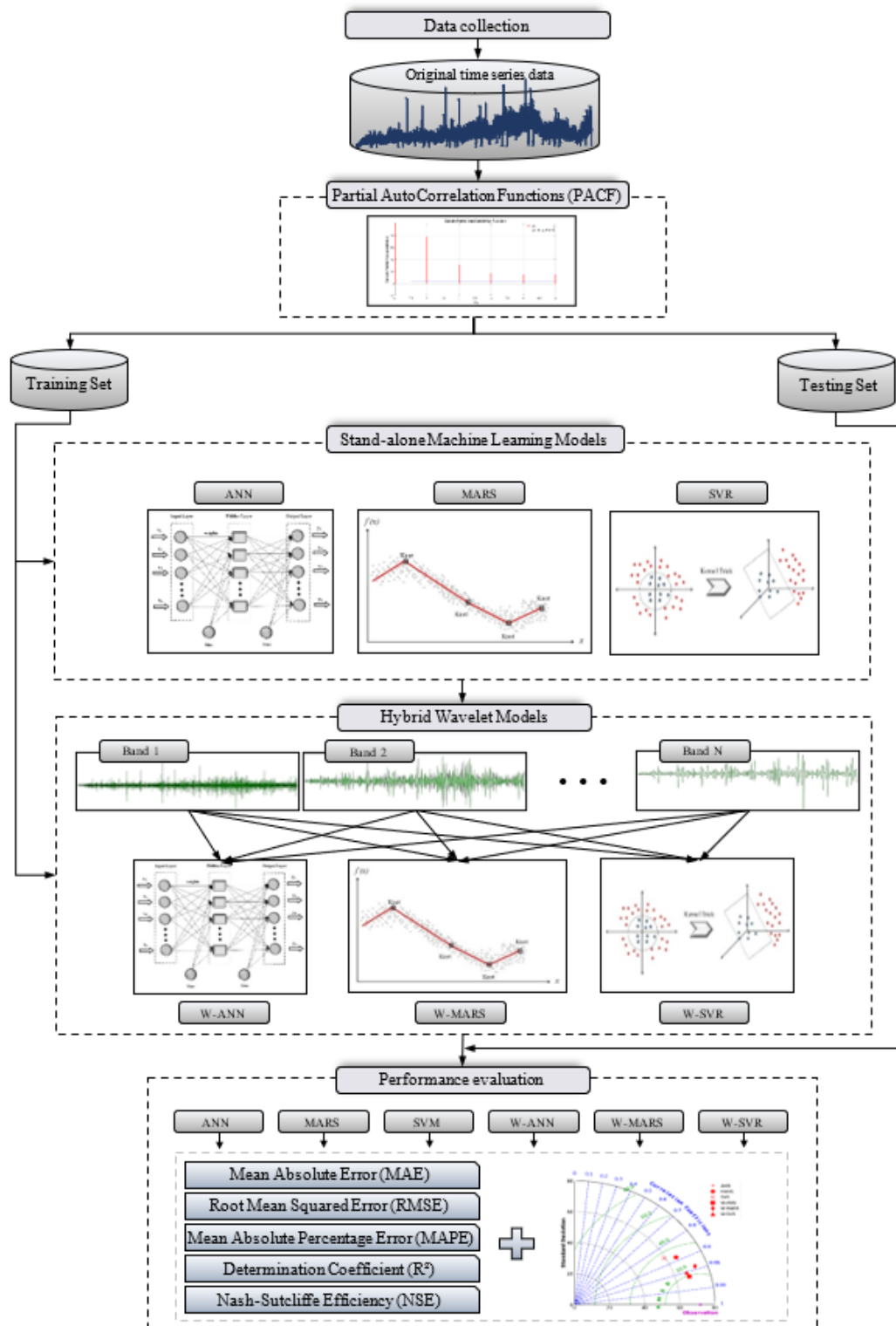


Figure 4.1 Analysis procedures of the proposed study

4.5 Materials and Methods

4.5.1 Data and Statistical Features

Occupational Health and Safety Law (Law no: 6331) [17] enacted in 2012 in Turkey resulted in a huge number of accident records between 2012 and 2020. The law states that employers must report occupational accidents to the Social Security Institution (SSI) within three days after an accident. Hence, some of the accidents experienced before 2012 were likely not to be recorded. According to the dataset acquired from the Republic of Turkey, SSI, there were only 425 accidents recorded between 1987 and 2011, while 393,160 recorded accidents were observed between 2012 and 2020 (up to October). Accidents recorded between 1987 and 2011 were not included in this study to minimize bias due to the law. The dataset obtained from the SSI initially included several variables such as accident date, time of the accident, location of the accident (city), gender, age and education level of the victim etc. According to the dataset 4,230 fatal and 388,930 non-fatal accidents were recorded, of which more than 30% took place in Istanbul, the most populated and urbanized city of Turkey. In addition, nearly 80% of the accidents were the first accident experience of the employees, while there was no workday loss in 74% of the accidents such that workers continued to work the same day. More than 75% of the accidents were experienced in small and medium-sized enterprises (SME) with number of workers less than 250. In addition, the majority of the recorded causes were referred to as “damage due to oscillating object”. It was observed that Monday was the day with the highest number of accidents and July was the most risky month.

Looking onto the literature, researchers tend to use ML models by using nation-wide accident datasets to predict measurable outputs such as accident outcome, injured body part, and accident severity. These models are usually developed to prevent future accidents by analyzing the reasons causing accidents or their severe consequences. In this study, a novel approach was adopted to explore the nation-wide dynamics of the occurrence of accidents. The accidents were classified daily and then summed up to calculate the total number of accidents occurred in each day throughout the country, which formed a daily-recorded time series data between 2012 and 2020 in the construction industry. Table 4.2 provides the details of the time series data, which is based on the statistical features. There is an increasing trend in the frequency of the occurrence of accidents between 2012 and 2018, while a decline can be observed between 2018 to 2020, mostly due to the recession in local construction industry. According to the Turkish

Statistical Institute [128], the average monthly employment in the construction sector increased from 2012 to 2018 at different rates, while it decreased by 4.9, 21.5, and 4.4% in 2018, 2019, and 2020, respectively.

Table 4.2 The statistical features of the time-series data

Years	Number of daily occupational accidents					
	Total	Maximum	Average	Minimum	Standard Deviation	Variance
2012	17,303	123	47.27	1	26.96	726.96
2013	27,727	376	75.96	0	30.81	949.00
2014	31,610	313	86.60	0	30.26	915.68
2015	37,233	453	102.01	0	42.60	1,814.37
2016	48,304	336	131.98	1	49.63	2,462.74
2017	66,656	471	182.62	9	69.29	4,800.95
2018	77,632	514	212.69	11	72.45	5,248.63
2019	49,710	288	136.19	3	45.49	2,069.17
2020	36,985	310	126.66	11	58.01	3,365.66
Grand Total	393,160	514	122.33	0	70.07	4,909.33

Some highlights can be pointed out related to the time series dataset used in this study. There were only three days without an accident in the dataset, while at least 11 occupational accidents occurred every day in 2018 and 2020 (Table 4.2). The maximum daily number was experienced in 2018 as 514 and the daily average was recorded as 212.69 in the same year. These implications signify the importance of the occupational accident time series analysis in Turkey. In this research, the occupational accident time series data was divided into two major groups as the training and testing sets to develop prediction models. The first 70% and remaining 30% of the data was considered as the training and testing sets respectively, based on the separation as a block [60] as illustrated in Figure 4.2, similar to the other studies using time series analysis in the literature [171], [172].

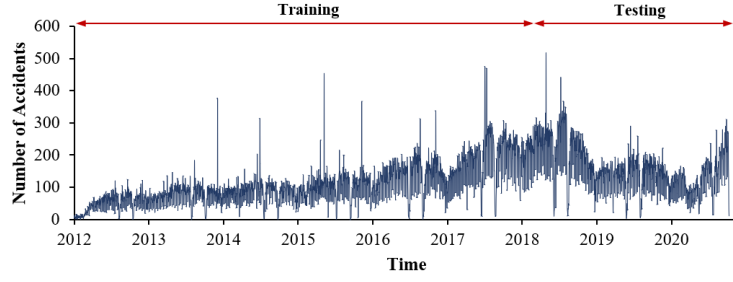


Figure 4.2 The time series of the daily occupational accidents

4.5.2 Data Preprocessing

4.5.2.1 Lag Times and Lead Times Identification

In time series predictions, lag times refer to the previous time domains that affect the current time domain. Also, lead times denote the time domains, which are to be predicted. Therefore, the effective time domains (lag times) are selected as inputs, while the lead times are used as outputs (Figure 4.3). The lag time determination is based on a settled down approach, simply correlation among the time domains, and the identification of lead times depends on the value attributed by a user for predicting the desired time domain. In this context, input-output structure of a prediction framework is constituted. As a rule of thumb, correlations between the time series and delays of them are obtained by using both autocorrelation functions (ACF) and partial autocorrelation functions (PACF). The similarity of the original signal over the time delays is measured via autocorrelation functions [60]. Thus, the similar previous time domains gain significance to be used as inputs to predict future time ahead.

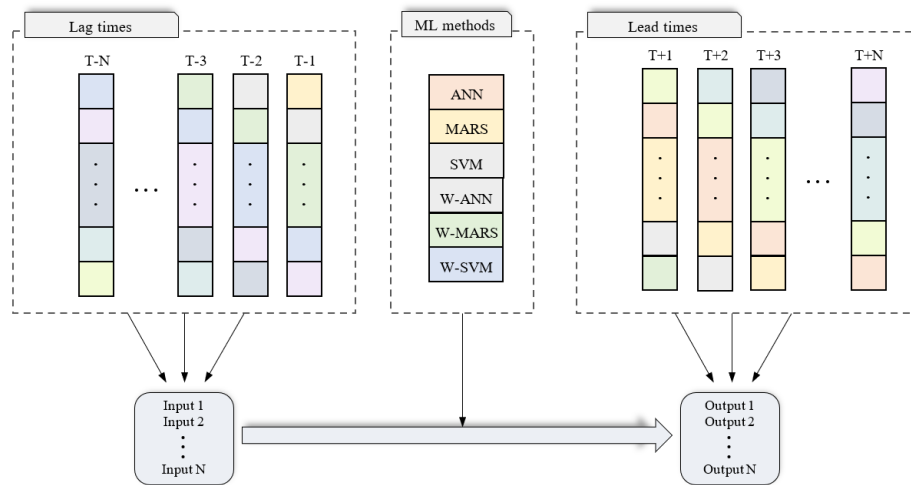


Figure 4.3 Illustration of lag and lead times

4.5.2.2 Optimal Decomposition Level Determination

In time series analysis, determining the optimum level of decomposition can be considered as an initial step once the original series are divided into different sub-bands with respect the various frequencies. In the discrete wavelet decomposition, which was used in this study, there is a common approach allowing users to employ the following formulation to determine the optimal decomposition level:

$$J_0 \leq \log_2(1.5 \times N) \quad (4.1)$$

in which J_0 is the maximum number of levels and N is the number of data samples. Within the boundaries of this approach, the determination of optimum level can also be performed by trial-error approach [60]. Thus, one can identify the maximum number of sub-bands using the Eq. (4.1) and use it as a boundary condition. The more decomposition levels, the more computational load. Hence, in this study, Eq. (4.1) was used to obtain the maximum number of sub-bands and then trial-error was performed to finalize the decomposition level determination process.

4.5.3 Machine Learning Model Setup

In this study, ANN, MARS and SVR were adopted as machine learning methods and integrated with discrete wavelet transform to predict the number of occupational accidents in the construction sector in Turkey with 1-day, 7-day, and 30-day ahead projections. ANN structure was developed with reference to the feedforward backpropagation, in which the number of hidden layers was determined as 8 and the maximum iteration was set to 2,000. For training, Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm was used and the selected activation function was tangent hyperbolic. In the MARS model, the parameters were determined based on the trial-error approach, such that the penalty parameter and the maximum degree of interactions were selected as 4 and 2, respectively. SVR setup was also adjusted through the trial-error method and the commonly used RBF was selected as kernel function. In addition, the hyperparameters of the SVR model were determined as 8×10^{-4} , 1×10^{-4} , and 80 for epsilon, gamma, and penalty, respectively.

4.5.4 Performance Evaluation Indicators

In statistical studies, providing a fair performance evaluation among the utilized methods plays a critical role in addressing the most powerful predictions. Therefore, a great deal

of performance metrics is utilized by the research society. In this study, the performance of proposed models were evaluated through five indicators; mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), determination coefficient (R^2), and Nash-Sutcliffe efficiency (NSE) index [173]. In MAE, RMSE, and MAPE, the smaller and larger values indicate higher and lower accuracies, respectively. In addition, positive and negative values are the subjects for R^2 performance evaluation metric. Positive ones indicate the direct proportionality of the recorded and predicted values and negative values refer the opposite (Table 4.3). Values of -1 and 1 represent the perfect match, while 0 means that there is no correlation between the observations and predictions. In the last performance metric NSE, value 1 shows an excellent performance. The performance of the utilized model is considered to be acceptable when the NSE value is greater than 0.50 [174]. In this study, the model(s) with the highest NSE and R^2 , and the lowest MAE, RMSE and MAPE values were selected as the best model.

Table 4.3 The equations of performance criteria

Performance Metric	Minimum	Maximum	Equation
Mean Absolute Error (MAE)	0	$+\infty$	$MAE = \frac{1}{n} \sum_{i=1}^n A_{pi} - A_{ri} $
Root Mean Squared Error (RMSE)	0	$+\infty$	$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (A_{pi} - A_{ri})^2}$
Mean Absolute Percentage Error (MAPE)	0	$+\infty$	$MAPE = 100 \times \frac{1}{n} \sum_{i=1}^n \left \frac{A_{ri} - A_{pi}}{A_{ri}} \right $
Determination Coefficient (R^2)	$-\infty$	1	$R^2 = \frac{\sum_{i=1}^n (A_{ri} - \bar{A}_r) (A_{pi} - \bar{A}_p)^2}{\sum_{i=1}^n (A_{ri} - \bar{A}_r)^2 \sum_{i=1}^n (A_{pi} - \bar{A}_p)^2}$
Nash-Sutcliffe Efficiency (NSE)	$-\infty$	1	$NSE = 1 - \frac{\sum_{i=1}^n (A_{ri} - A_{pi})^2}{\sum_{i=1}^n (A_{ri} - \bar{A}_r)^2}$

Note: A_r is the recorded number of accidents, A_p is the predicted number of accidents and $\bar{A}_r$ and $\bar{A}_p$ are the average of observed and predicted number of accidents, respectively. n represents the number of days.

4.6 Results and Discussion

4.6.1 Optimal Lag-times and Lead-times

In this study, partial autocorrelation functions (PACF) were used to determine the input variables. Since defining lag-times is highly important for the model accuracies, they were rigorously established with respect to confidence interval of 99%. Figure 4.4 depicts that 1-day (T-1), 2-day (T-2), 7-day (T-7) and 8-day (T-8) lag-times have significant correlations with the original occupational accident dataset. Findings highlight that the number of daily accidents was affected by historical accidents observed in T-1, T-2, T-7, and T-8 at most. The results show the importance of the “day of the week” in daily construction accidents, such that 7 days lag can be observed from T to T-7, and from T-1 to T-8. Choi et al. [2] also found “day of the week” as one of the most significant factors contributing construction accidents. Therefore, inputs were determined as a set of four different lag-times, namely T-1, T-2, T-7, and T-8.

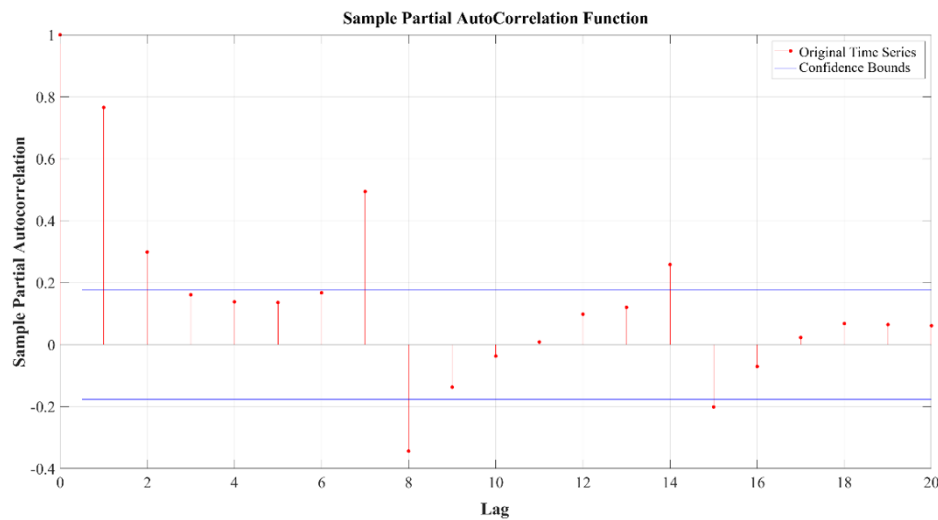


Figure 4.4 The partial autocorrelation functions

For the future predictions regarding the number of daily construction accidents, three different lead times were considered: 1-day ahead (T+1), 7-day ahead (T+7), and 30-day ahead (T+30) deemed as short-term, mid-term and long-term projections respectively. The main reason to select T+1, T+7, and T+30 in the prediction model is to support practical implications of the study for improved safety and project management applications. Health and safety reports are generally prepared as daily, weekly, and monthly documents, similar to other reports in construction projects. Zang et al. [175] developed a safety training scenario, measured safety training values daily, and

implemented to daily, weekly, and monthly safety training strategies. In addition, safety meetings are generally held daily, weekly or monthly to discuss safety concerns [176]. It is also a common practice that the communication between team members and departments usually rely on daily, weekly, or monthly periods in construction projects. With all these actual practice examples, inspectors, decision makers, practitioners, and safety departments of construction companies could be informed with valuable outcomes related to prediction of occupational accidents for 1-day, 7-day, and 30-day ahead (particularly by filling the gap between 1-7 days, and 1-30 days). The predictions could also support the planning of safety measures and daily, weekly, and monthly emergency plans.

4.6.2 Stand-alone Model Results

In this study, three different stand-alone ML methods, ANN, MARS and SVR, were employed to the daily occupational accident data for projections in short-term, mid-term and long-term time periods. According to scatter plots presented in Figure 4.5, the ANN model was well-scattered in the 1-day ahead prediction, while it mostly overestimated the number of occupational accidents for 7-day and 30-day ahead predictions. Findings illustrate that the best predictions were obtained through the MARS model for the short-term period with $R^2 = 0.7822$, which was the highest acquired determination coefficient among stand-alone models (Figure 4.5). The best models were obtained in short-term predictions in SVR model similar to the other two methods. A number of cross-sectional studies also suggest that the highest accuracies are obtained when the duration of the predictions are reduced [177], [178]. Thus, these findings broadly support the pertinent literature grounded on time series predictions. On the other hand, scatter plots provide an overview for mid-term and long-term predictions. Here, stand-alone models were not well-scattered against the recorded number of accidents and overestimated the number of accidents as the lead times increased. Brief statistics related to stand-alone model performances are provided in Table 4.4. Consistent with the scatter plots, the best models were obtained for 1-day ahead predictions, while the MARS model slightly outperformed other models in the short-term prediction with $NSE_{MARS} = 0.7810$. On the contrary, the worst performances were obtained with MARS model for 7-day and 30-day ahead predictions. An unconventional result was also observed in T+30 prediction acquired through the MARS, such that a negative NSE value was calculated. It can be concluded that the stand-alone models revealed good performances for short-term predictions, while

showing moderate performance in mid-term predictions except the MARS model. As a result, the accuracies of the stand-alone models were statistically unacceptable particularly for the long-term predictions, since the NSE values were below the threshold value of 0.5, according to the categorization introduced by Moriasi et al. [174].

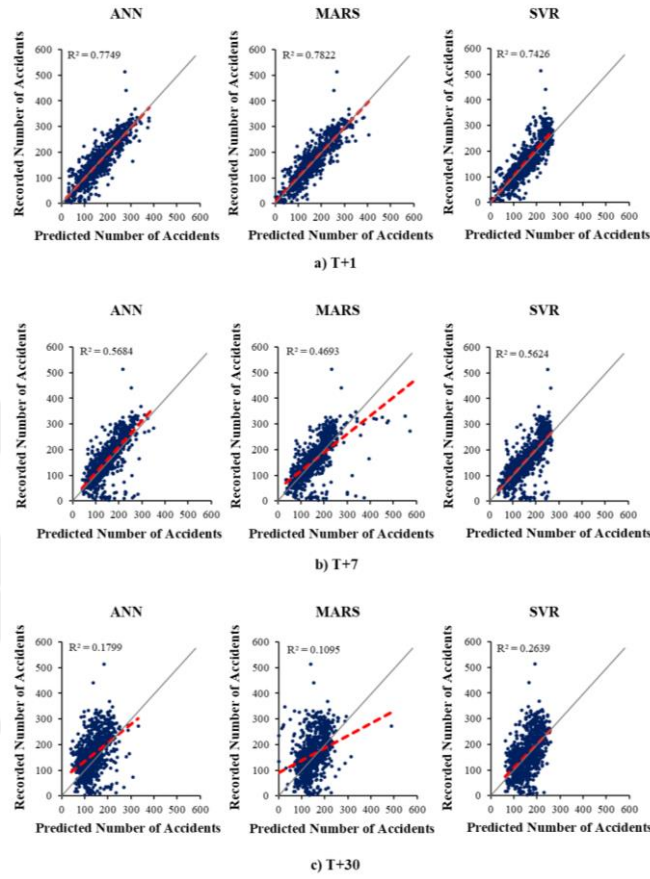


Figure 4.5 The scatter plots of the stand-alone models obtained for the test set (grey continuous lines represent 1:1 perfect lines and red dashed lines represent the trend lines)

Table 4.4 Performance comparison of stand-alone models

Output	Method	RMSE	MAE	MAPE	NSE	R ²
T+1	ANN	34.1111	25.0737	27.8534	0.7738	0.7749
	MARS	33.5665	24.4850	24.3910	0.7810	0.7822
	SVR	36.9322	25.8595	23.6113	0.7349	0.7426
T+7	ANN	48.3614	33.4341	52.3703	0.5455	0.5684
	MARS	55.8606	34.3971	54.6689	0.3937	0.4693
	SVR	47.5273	31.4054	53.3413	0.5611	0.5624
T+30	ANN	70.6311	54.2962	62.9440	0.0307	0.1799
	MARS	74.6251	53.0149	64.3824	-0.0820	0.1095
	SVR	61.8614	46.6379	65.5043	0.2565	0.2639

4.6.3 Hybrid Wavelet Model Results

In this study, the DWT method was used to improve the stand-alone model performances. Therefore, the original time series data was decomposed into five different sub-series beginning from the Level-4 to the Level-8 to identify the impact of the decomposition level on the prediction accuracies. To accomplish this aim, the maximum number of sub-bands was calculated as 12 using Eq. (4.1), and then different decomposition levels were tried starting from level 4. This investigation was carried out according to the one lead-time, such as 1-day ahead, via utilized ML methods, ANN, MARS and SVR. Table 4.5 shows the results obtained for each decomposition level regarding various performance evaluation metrics. Strong evidence of choosing appropriate level as Level-5 (5 details and 1 approximation) was found as the model performances reduced when the original series decomposed into more than 5 levels (Table 4.5). Thus, the optimum level was determined as Level-5, which also enabled an effective prediction with a decrease in computational load. The decomposition level was not increased over 8 since the accuracy of the model did not increase and choosing higher number of sub-bands was computationally expensive. The mother wavelet was selected as sym4, given the similarity of sym4 mother wavelet with the original time series. The spectral bands decomposed by DWT are presented in Figure 4.6, in which original series, approximation, and details are ordered from the top to the bottom with red, blue, and green lines, respectively.

Table 4.5 Determination of the optimum decomposition level (for 1-day lead time)

Method	Level	RMSE	MAE	MAPE	NSE	R ²
W-ANN	Level-4	19.3817	15.0322	14.2505	0.9270	0.9276
	Level-5	19.3165	15.0403	13.7579	0.9275	0.9287
	Level-6	19.3646	15.0192	14.2251	0.9271	0.9278
	Level-7	19.6524	15.2098	14.6907	0.9249	0.9275
	Level-8	19.5631	15.1499	14.5644	0.9256	0.9276
W-MARS	Level-4	25.2197	15.8757	14.8878	0.8764	0.8824
	Level-5	25.0199	15.7436	14.4088	0.8783	0.8843
	Level-6	25.0602	15.7558	14.6001	0.8779	0.8842
	Level-7	25.0536	15.7615	14.5989	0.8780	0.8842
	Level-8	25.0550	15.7610	14.6020	0.8780	0.8842
W-SVR	Level-4	23.2546	16.3885	16.3616	0.8949	0.8957
	Level-5	21.5436	15.9122	16.5264	0.9098	0.9098
	Level-6	21.8502	16.1513	16.5048	0.9072	0.9081
	Level-7	22.1430	16.3293	16.4436	0.9047	0.9065
	Level-8	22.1561	16.0394	16.5100	0.9046	0.9053

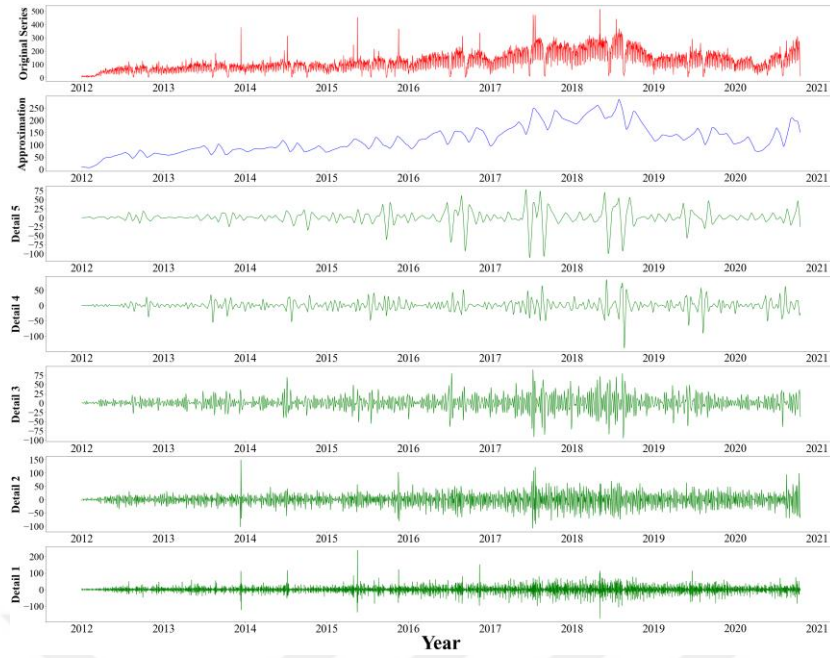


Figure 4.6 Spectral bands decomposed by discrete wavelet transform

Compared to stand-alone model results, the accuracy of the predictions increased significantly with the conjunction of DWT as can be seen from the scatter plots depicted in Figure 4.7. For three of the lead-times, improved results were obtained with hybrid models. It was observed that the predictions and observations were well scattered, particularly for short-term and mid-term periods. Even though the accuracies were substantially increased, overestimated and underestimated results still existed in the long-term projections. Performance comparison of hybrid wavelet models is shown in Table 4.6. The findings highlight the increase in the performance of ANN model with the integration of DWT. Accordingly, W-ANN provided the highest accuracies for T+1 ($NSE_{W-ANN}=0.9275$) and T+7 ($NSE_{W-ANN}=0.7678$) predictions. On the other hand, the performance of W-MARS ($NSE_{W-MARS}=0.4109$) was slightly better than W-ANN ($NSE_{W-ANN}=0.3903$) for T+30 predictions. The smallest MAE and RMSE values were obtained with W-ANN for short-term and mid-term and with W-MARS for long-term predictions. It can also be noted that the accuracy of the W-SVM models were slightly lower than other two hybrid models with $NSE_{W-SVR}=0.9098$, $NSE_{W-SVR}=0.7285$ and $NSE_{W-SVR}=0.3745$ for 1-day, 7-day, and 30-day ahead predictions, respectively (Table 4.6).

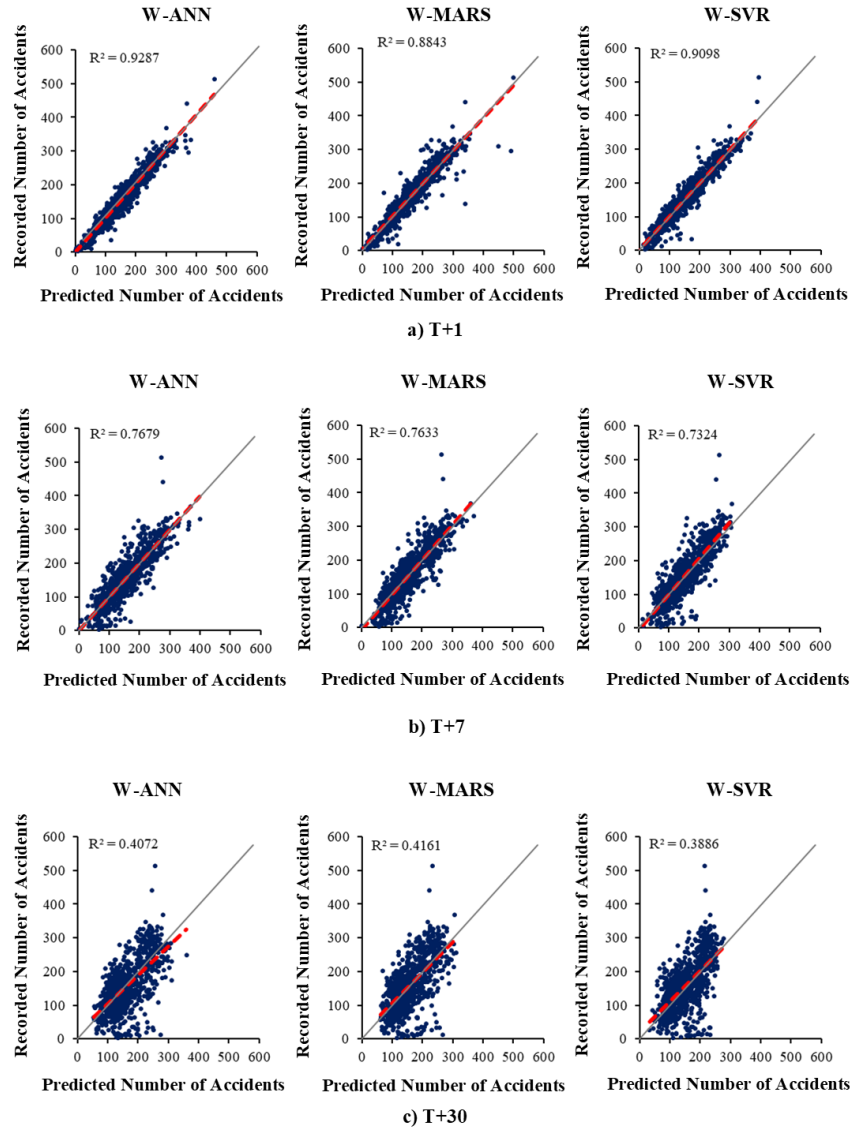


Figure 4.7 The scatter plots of the hybrid wavelet models obtained for the test set (grey continuous lines represent 1:1 perfect lines and red dashed lines represent the trend lines)

Table 4.6 Performance comparison of hybrid wavelet models

Output	Method	RMSE	MAE	MAPE	NSE	R ²
T+1	W-ANN	19.3165	15.0403	13.7579	0.9275	0.9287
	W-MARS	25.0199	15.7436	14.4088	0.8783	0.8843
	W-SVR	21.5436	15.9122	16.5264	0.9098	0.9098
T+7	W-ANN	34.5698	25.1946	30.4042	0.7678	0.7679
	W-MARS	34.9743	26.0570	30.6950	0.7623	0.7633
	W-SVR	37.3825	27.0467	34.0063	0.7285	0.7324
T+30	W-ANN	56.0173	41.4375	63.4305	0.3903	0.4072
	W-MARS	55.0657	40.4167	60.4311	0.4109	0.4161
	W-SVR	56.7382	42.0650	61.0664	0.3745	0.3886

Figure 4.8 shows the comparison between the observed and predicted number of accidents regarding the best models (W-ANN, W-ANN and W-MARS for T+1, T+7 and T+30 projections, respectively). In short term predictions (1-day ahead), the W-ANN performed well and captured both the extreme high and low values mostly. Although to a lesser extent, W-ANN still showed considerable performance in the mid-term predictions through capturing the trend in the time series, except performing great predictions for some of the higher extremes. To further probe into statistical evaluations of the predicted time series and provide concise summary and quick overlook for model performances, Taylor diagrams were introduced as shown in Figure 4.9. Based on versatile statistical evaluations, Taylor diagrams showed that hybrid wavelet models outperformed the stand-alone models for each lead-time. They also illustrated that the correlations between stand-alone model results and the observations were relatively low for T+1 and T+7 predictions and considerably low for T+30 predictions. The figure further depicts that the hybrid models had better match patterns with the observations according to the ratio of the variances.

Overall, it can be stated that the frequency of the occurrences of occupational accidents can be predicted through either nation-wide or within the scope of a mega project that has daily accident records for short and mid-term time spans with higher accuracy levels. For the long-term predictions, more investigations should be conducted to achieve improved accuracies. Comparing with the existing research in the construction safety literature, this study shows a uniqueness by predicting safety outcomes based on daily time series data consisting of 393,160 construction accidents. The utilization of the DWT contributed a lot in the improvement of prediction performances of ML methods in the number of construction accident prediction. By using the analysis procedures presented in this paper, inspection entities, safety managers, and decision makers can develop safety measures and policies to improve safety conditions, and act proactively for the preventive strategies. The following sub-section describes how this study can be used dynamically in a generic view to prevent future accidents.

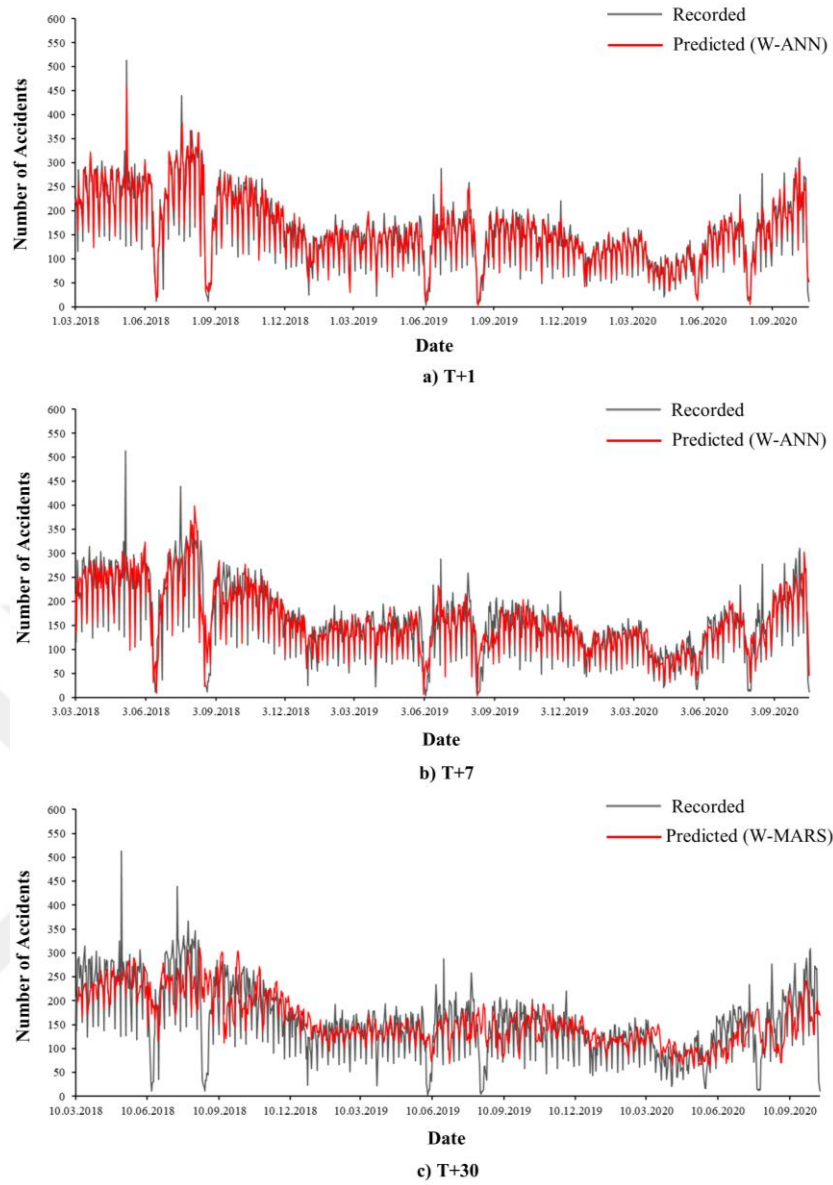


Figure 4.8 Recorded and best modelled time series with respect to the testing set. (a) T+1, (b) T+7, (c) T+30

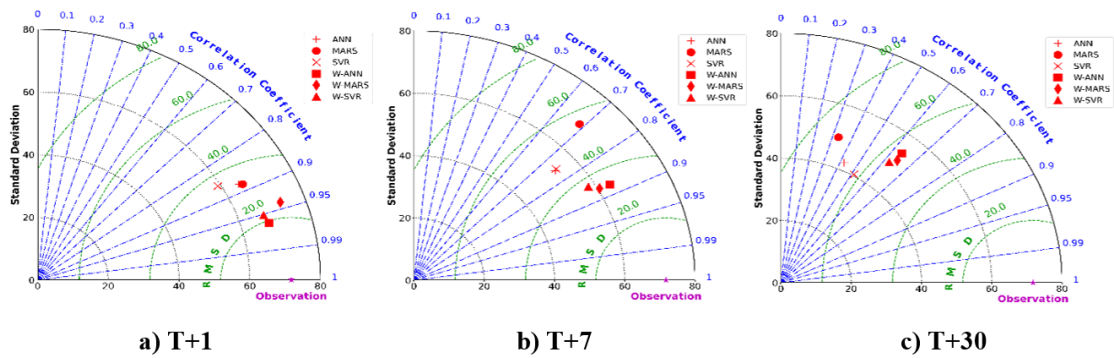


Figure 4.9 Taylor Diagrams. (a) T+1, (b) T+7, (c) T+30

4.6.4 Practical Use of the Study

This study offers a prediction model to determine the number of occupational accidents in the construction industry, with a focus on 1-day ahead, 7-day ahead, and 30-day ahead time periods. This approach allows the utilization of a dynamic prediction system in a discrete period. The proposed model can be used by related institutions, OHS companies, frontrunners of the sector, and safety professionals in mega projects (with a risk of high number of accidents) dynamically to predict the number of accidents for future projections. For a deeper insight, dynamic utilization plan is presented on the basis of 30 days predictions (Figure 4.10). By running the proposed model 30 days consecutively, related institutions and companies can predict results for the following 30 days, 1st day from the model of T+1 (Best model: W-ANN), 2nd to 7th days from the model of T+7 (Best model: W-ANN), and 8th to 30th days from the model of T+30 (Best model: W-MARS). These findings can provide insights to institutions and organizations with early detection capability of the number of potential occupational accidents. With the predictions, such entities can perform nation-wide or project-based inspections proactively, and maintain improved safety performances. Proactive risk management plans can be developed by construction firms and OHS departments, which could contribute to minimization of occupational accidents through the detection of most risky time periods.

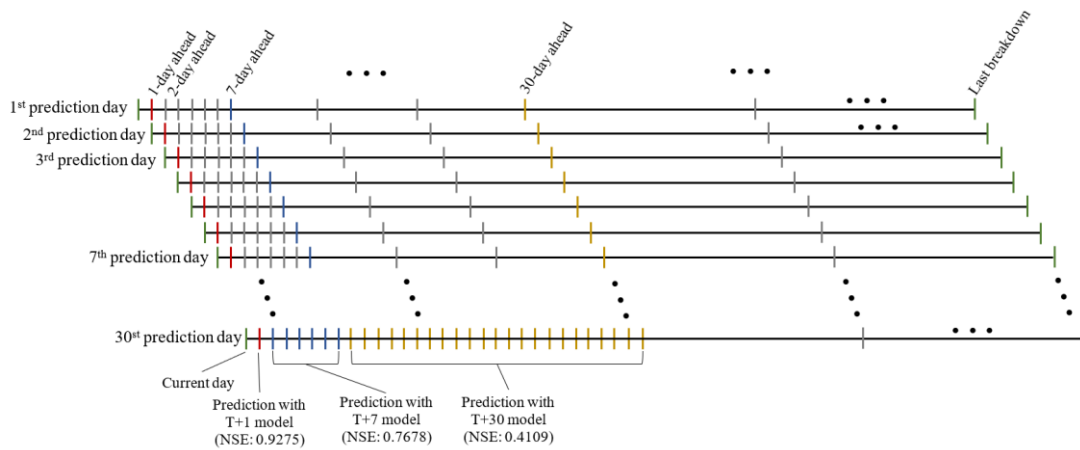


Figure 4.10 Dynamic utilization of the proposed framework

With the Occupational Health and Safety law (Law no: 6331) [17], the Occupational Health and Safety Regulation in Construction Works [19] defines actions that are required for construction works in Turkey. Measures that need to be taken by employers in construction works are monitored and controlled by inspectors of the Ministry of Labor

and Social Security, based on Labor Law (Law no: 4857) [179]. The role, authority and responsibility of the inspection board are regulated with the Ministry of Labor and Social Security, Labor Inspection Board Regulation [180]. Some of the duties of the board can be summarized as: (i) performing sector analysis regarding occupational accidents and work environment, (ii) giving priority for the inspection of some particular and more risky sectors, (iii) providing an advice for legislation when social changes or technological improvements are observed, (iv) providing recommendations to improve the work quality of employees according to the results of inspections, and (v) acquiring, analyzing and interpreting the statistics of the inspection results. Thus, the timely inspection of the working environment and analyzing the number of accidents that are likely to occur in the future are important for the improvement of safety performance of the companies. By implementing the proposed model for different sectors other than the construction industry, the sector-basis inspection frequency can be adjusted with the predictions. Human resource management for inspection (through inspectors) can be performed dynamically and effectively by inspection boards. This process can be integrated to other primary duties of the inspection board as illustrated in Figure 4.11.

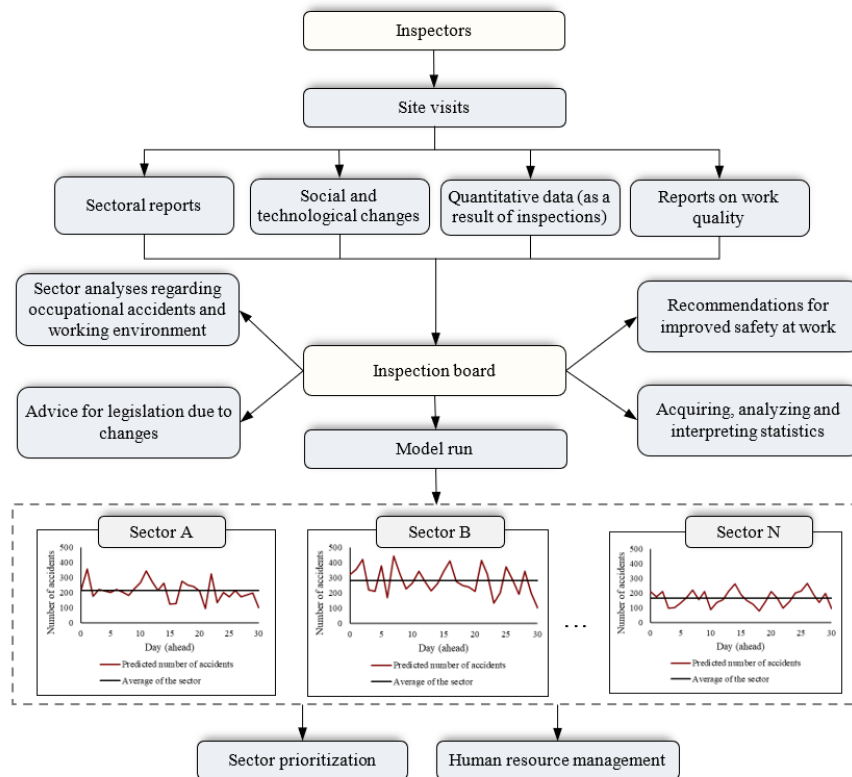


Figure 4.11 The role of inspection boards and sector-basis investigation

The integration of the proposed model with daily investigations and its dynamic use for proactive risk management is also illustrated in Figure 4.12. After implementing the proposed model 30 days consecutively (Figure 4.10) and projections for 30-day ahead are predicted, the days with high number of accident predictions can be more carefully monitored as a part of project management. For instance, if the number of accidents expected 1-day ahead (such as the one depicted in Figure 4.12) is predicted as high compared to the current trend (such as more than the average of the last month), then the OHS departments of the companies would have the opportunity to take necessary preventive actions and issue instructions to warn employees as a part of risk management.

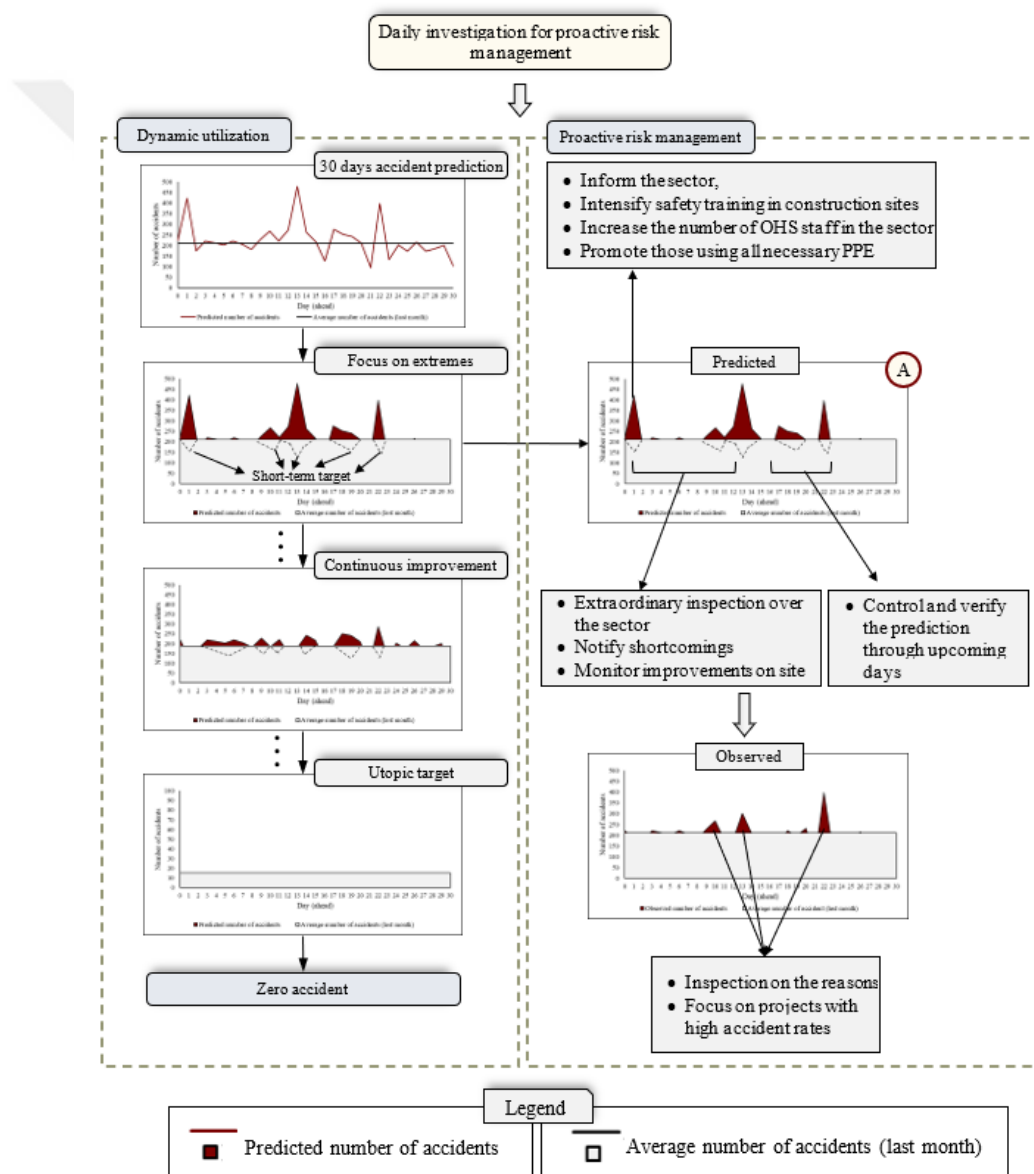


Figure 4.12 Proactive risk management and dynamic utilization of the study

The prediction of the days prone to more accidents would contribute to improve monitoring and controlling of required safety measures and trainings, and encourage the use of personal protective equipment (PPE) more carefully. To prevent severe accidents, the use of PPE is one of the centerpieces of safety management in construction. It is highlighted in the literature that reward is as effective as penalty in OHS management applications [181]–[183]. Therefore, workers (or corresponding subcontractors) can be rewarded if they are detected to be using all PPEs as required. As reported by many researchers like Mistikoglu et al. [1] and Xu et al. [184], the provision of safety training decreases the risk of fatality and safety trainings are influential on the safety performance, with particular focus on the personalized trainings. Hence, preplanned emergency safety training of employees could improve the safety performance of the construction projects. Another potential mitigation strategy can be considered as providing paid holidays or reducing the working hours particularly for those performing dangerous works as a reward mechanism, since employees who have less time for recreation due to long working hours are more prone to experience an occupational accident [185]. Evidence [186] shows substantial effect of long working hours, commonly encountered in the construction sector, on the safety performance as well as reducing the productivity and absenteeism on duty [187]. Therefore, workers' working periods, average working hours per week, and average holidays per month can be planned based on evidence from predictions, and corresponding mitigation measures can be implemented to minimize accidents due to long working hours and relevant stressors.

On the other hand, if high accident rates are predicted within a certain period (9th and 15th days as in the example, top right of Figure 4.12), those working days can be monitored with extra care and the frequency of the inspections can be increased as preventive or mitigation actions. Then, the inspection can be performed repeatedly to overcome the predicted risks and shortcomings can be rectified with additional monitoring and controlling activities on the construction sites. In case of a risky alert from the predictions throughout medium to long-term period (18th and 22nd days as in the example, Figure 4.12), then the activities can be monitored with more focus in the upcoming days with relatively higher prediction accuracy (such as 7-day ahead), which might also assist the human resource management plan. Figure 4.13 is drawn to illustrate how continuous monitoring can be conducted by inspection boards regarding long-term threats (with reference to “A” in Figure 4.12). The figure shows that short-term and long-term priority

levels of the construction sector (identified through sector-basis investigation) can be used as threshold values. The threshold values and predicted number of accidents can be compared daily as a decision-making parameter. Whenever the predicted values are calculated as higher than the corresponding threshold values, then the inspection decision (or expansion in inspection) can be considered.

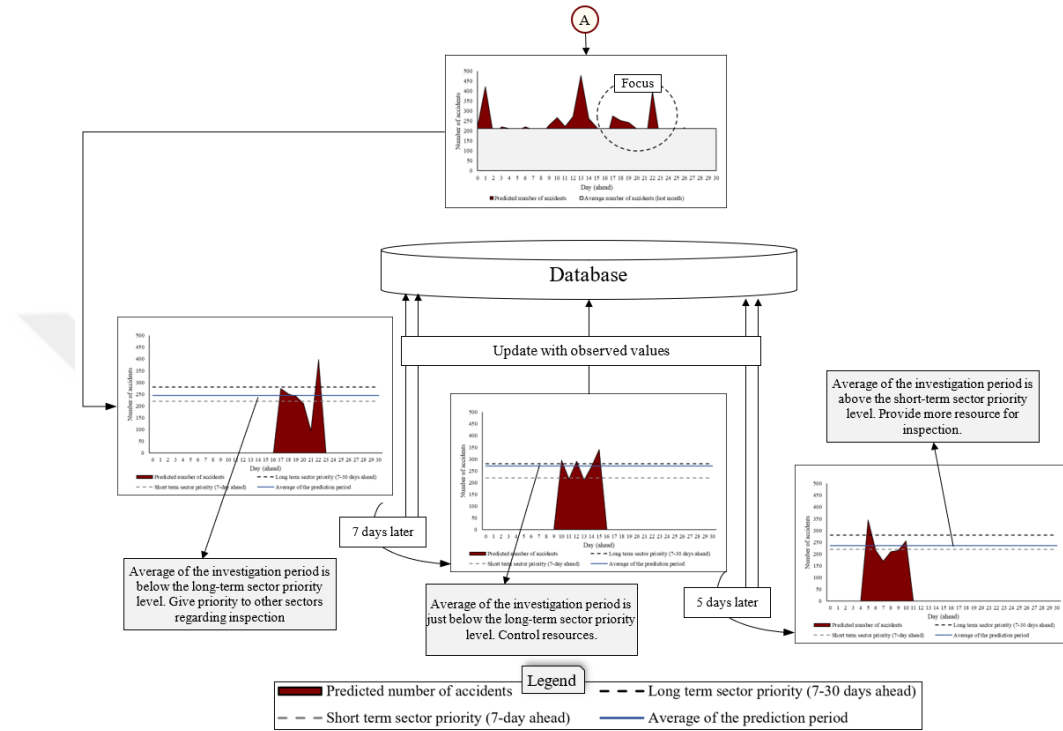


Figure 4.13 Example for continuous monitoring

The abovementioned procedures can be used dynamically by related institutions or construction companies to monitor, control, and abate safety risks generically, based on the national or company database. The ultimate and ideal objective of this study is to achieve accident-free construction industry or sectors in general through the proposed continuous improvement model with proactive safety risk management. For better insight into the practical utilization of the research, “five Ws and one H” method, mostly referred as 5W1H is also adopted, as illustrated in Figure 4.14. The answers to Why, What (purpose), Where, When, Who, and How are generally considered to be at the center of problem solving, which will aid in explaining the utilization of the proposed approach. Despite that the method is commonly used to describe news stories [188], the answers to the corresponding questions also play a vital role in project planning. Accordingly, this study is performed to highlight that safety is one of the major concerns in many industries, particularly in the construction industry due to its poor performance. It aims to improve

safety performance of countries, particular sectors, companies, or projects. The proposed model can be conducted in any countries where there is a daily accident database (number of accidents) available and can be run dynamically every day with the accumulated accident data. Inspection boards, OHS firms, construction companies, or OHS department of special projects can utilize the approach as a means to improve safety performances of the industries, companies or projects. Overall, the proposed model can be used with an integrative planning and implementation that is explained in subsequent illustrations in Figure 4.10, Figure 4.11, Figure 4.12, and Figure 4.13.

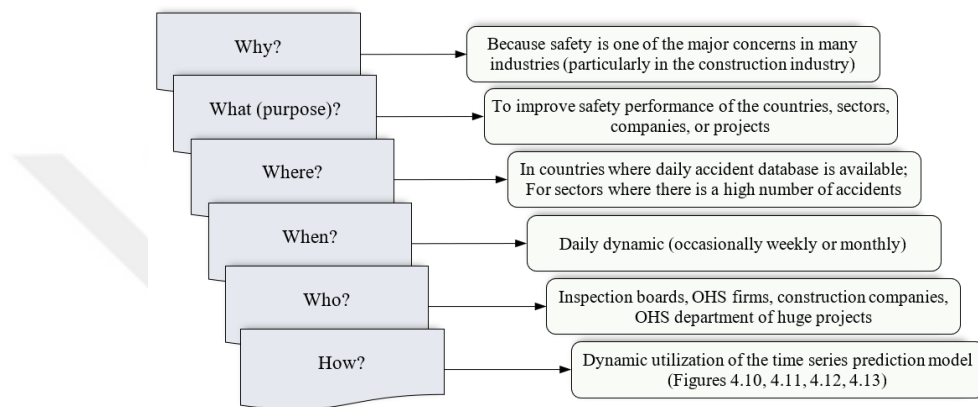


Figure 4.14 Answers to 5W1H

4.7 Conclusions

In this study, national daily time series data of 393,160 construction accidents between 2012-2020 were analyzed to predict the number of daily occupational accidents with respect to 1-day ahead (short-term), 7-day ahead (mid-term), and 30-day ahead (long-term) time periods. The main purpose of the study is to propose a predictive model to increase the efficiency of construction safety management and project management practices, for construction practitioners, safety managers and decision makers. Findings of this study also aim to assist safety professionals particularly during safety meetings to consider daily, weekly, or monthly occupational accident occurrence predictions to discuss safety concerns in construction projects. The proposed prediction framework can also be used by inspection boards dynamically, as scientific evidence developed by means of machine learning methods based on real data, to improve legislative power and safety conditions in the construction industry.

To achieve this objective, three different stand-alone machine learning methods, ANN, MARS and SVR, were employed to the daily national occupational accident dataset. To

improve the prediction performance of the presented methodology, discrete wavelet transform was integrated with the employed machine learning methods (W-ANN, W-MARS, and W-SVR) as a pre-processing technique. Six different models, including stand-alone and hybrids, were used to predict the number of construction accidents on a daily basis. Several performance evaluation metrics were utilized to compare the performance reliability of each model. The performance of W-ANN method was reported to show the best predictions for short-term and mid-term, while it was W-MARS for the long-term periods. The study shows that the discrete wavelet transform method increased the accuracy of all stand-alone models for all short-term, mid-term, and long-term time spans. Performances of the models demonstrate that the number of occupational accidents can be predicted with high accuracy rates, as in W-ANN method, which were 0.9275 and 0.7678 with respect to short- and mid-terms based on Nash-Sutcliffe efficiency index, respectively.

This study also provides valuable information about the generic view of occupational safety for decision makers. Still, it has some limitations, which needs be considered in future studies. Although the prediction accuracies were high for short and mid-term periods, the prediction performance was relatively lower for long-term consideration. Therefore, the developed predictive models in this study could be used for short-term and mid-term predictions, while long-term predictions require further investigations for improved accuracy in the models. To increase the predictive performance of the models, different decomposition techniques and machine learning algorithms may also be considered in future studies. In addition, even though the data related to occupational accidents were trained and tested, results could vary for different countries based on country-specific differentiations in the dataset. International comparative studies on predicting the number of occupational accidents would be useful to observe the impacts of country-basis conditions. Besides, projections with other national variables (in daily time series form), which are potentially effective in the prediction of the number of occupational accidents, can help policymakers understand in what national conditions more or fewer accidents occur. Finally, this study analyzed all occupational fatal and non-fatal accidents in the construction industry to develop prediction models in different time periods. Predicting the number of fatal and non-fatal accidents separately for daily, weekly, or monthly time periods (based on the amount and availability of dataset) would contribute a lot to the body of knowledge. Through an elaborated investigation and

improvement of the proposed model based on the nation-wide time series occupational accident dataset, a new safety management plan could be prepared by governmental authorities to manage safety issues in construction projects more efficiently.



SCENARIO-BASED AUTOMATED DATA PREPROCESSING TO PREDICT SEVERITY OF CONSTRUCTION ACCIDENTS

This chapter was published in *Automation in Construction*, vol. 140, p. 104351 as a research article.

5.1 Introduction

Occupational accidents not only cause severe injuries and fatalities, but also interrupt the orderly progress of work [189]. Construction projects, in particular, are executed under a turbulent and dynamic environment, which renders it one of the most demanding and dangerous industries [15], [190]. Construction sites can be characterized by their complex, uncertain and cluttered nature. A continuously changing environment is experienced nearly every day with movements of workers, materials, and construction equipment [191]. Thus, workers in the construction industry are at higher risk to experience severe accidents compared to workers in other industries in many countries [192]. Although serious efforts have been paid to minimize safety concerns in the industry, occupational safety in construction has experienced relatively slow progress [7]. Hence, improving workplace safety performance in construction projects is an important need for the industry.

Accident outcomes culminating in fatality and permanent incapacity are commonly considered as catastrophic and high-severe accidents [193]. Fatality and permanent incapacity rates account for very high percentages in the construction industry. Even though the employment rate is around 7% globally [3], the construction industry accounts for approximately 30-40% of occupational fatalities in most countries [3], [15]. The rates of fatality (32%) and permanent incapacity (37%) in the construction industry are also considerably higher than all the other industries in Turkey [21]. Thus, developing a model to predict the severity level of construction accidents is important to improve the safety of construction workers, provide confidence to project parties, and add value to the overall performance of the projects.

New ideas and opportunities have emerged in addressing new problems and/or solutions [194] with the continuously acquiring data and advances in machine learning (ML) applications. Historical data is the fundamental ingredient of the ML-based prediction models. Although most construction organizations and institutions have accident databases, they are hardly used regarding advanced analytics [66], since comprehensive and detailed preprocessing should be applied with a list of issues related to the quality of the data [195], [196] such as missing data problem, outliers, different types and scaling of features, and imbalanced class distribution [193]. Therefore, the requirement of multi-step comprehensive preprocessing is suggested in the literature [197] to minimize erroneous and misleading results. It is also well acknowledged that the majority of the time required for the entire process in ML applications was devoted to preprocessing [196]. These facts highlight that building a foundation for ML implementations with a robust data preprocessing phase could become as significant as developing a robust ML model [195]. Despite these, the effects of different preprocessing techniques on the performance of the proposed models with a scenario-basis investigation have not been explored comprehensively in the literature. Choosing the best preprocessing scenario has usually been undervalued [198], and more emphasis has been put on selecting the most appropriate ML method in the literature [8], [11], [14], [15], [199].

The main research focus of this study is to determine the combination of preprocessing techniques which yields the most accurate severity prediction for construction accidents. Accordingly, the aim is to develop a scenario-basis automated preprocessing model that identifies the most robust preprocessing pipeline automatically for accident severity prediction. The proposed model was tested with a state-of-the-art eXtreme Gradient Boosting (XGBoost) method due to its high performance in problem-solving, fast operation speed, effectiveness in calculations, and ability to achieve advanced results on complex datasets [48], [53], [54], [200], [201]. Overall, primary contributions of this study are listed below.

- One of the most prominent contributions of this study to the practice is enabling project and safety managers who aim to improve safety management practices in construction projects act proactively through the proposed prediction model, since it is an arduous task to identify events that could lead to severe accidents before they occur and detect the most effective preprocessing scenario.

- The utilized dataset includes 388,907 workers, who experienced occupational accidents in the construction sector in Turkey between 2012 and 2020, which enabled a statistical and empirical analysis on a national scale. Despite accident databases have widely been used to address construction safety issues [202], [203], they were merely used for accident outcome prediction purposes. On the other hand, existing literature on accident outcome prediction relies on smaller sample sizes and most of them focused on accidents in developed countries.
- A construction safety management framework is proposed, supported, and enhanced by the analysis results. This framework can be used as a detector for potential cases with high probability of occurrence values that might result in fatality and permanent incapacities. The proposed framework can be used daily as a warning mechanism about the workers that require particular attention in terms of safety measures.

The flow and structure of the remaining part of the study is shown in Figure 5.1. *Section 5.2* shows the results of literature review in terms of severity prediction of construction accidents (*Section 5.2.1*) and data preprocessing in ML (*Section 5.2.2*), and identifies research gaps. The national dataset used is described in *Section 5.3*, followed by model development stage in *Section 5.4*. In this section, initial preprocessing (*Section 5.4.1*), adopted comprehensive preprocessing steps (*Section 5.4.2*), proposed automated preprocessing model (*Section 5.4.3*), adopted ML method to test the model (*Section 4.4*), and considered performance evaluation metrics (*Section 5.4.5*) are explained. *Section 5.5* is focused on the results of preprocessing model (*Section 5.5.1*) and feature importance (*Section 5.5.2*). In discussion (*Section 5.6*), adopted methodology (*Section 5.6.1*) and feature importance results (*Section 5.6.2*) are evaluated and compared with the relevant literature, and how the proposed model can be used in construction sites is elucidated (*Section 5.6.3*). Finally, *Section 5.7* presents the recommendations, limitations, future research directions, and conclusions.

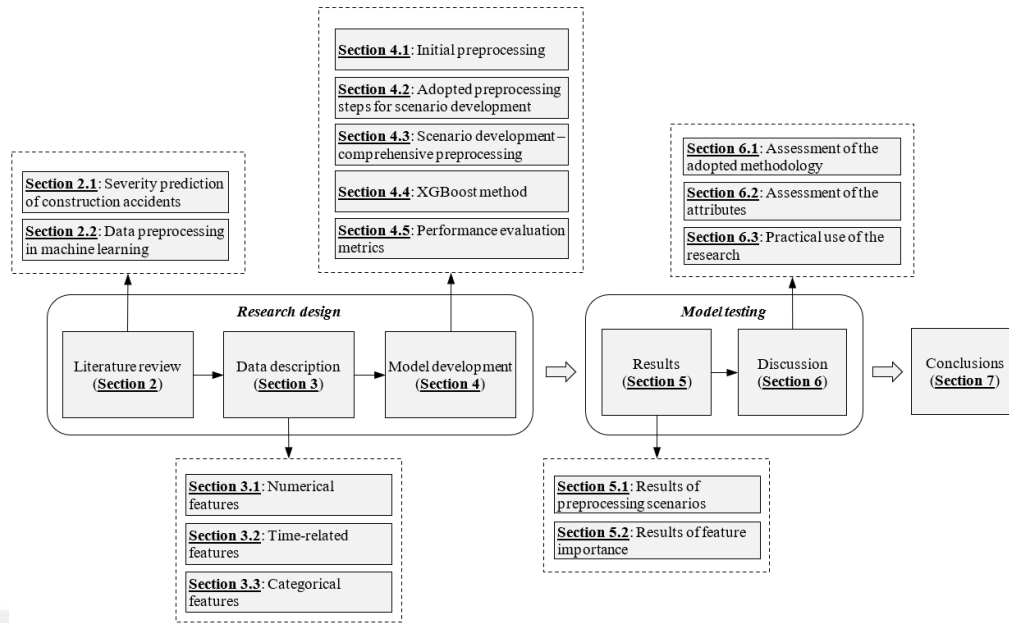


Figure 5.1 Flow of the research

5.2 Literature Review

5.2.1 Severity prediction of construction accidents

Conventionally, safety risks of workers are calculated by multiplying the probability of occurrence of an accident with its predicted severity [189]. Existing studies focus on the probability of occurrence due to the available data that corresponds to the characteristics of work environments. However, the severity is as equally important as the probability in risk quantification [7], [189]. Thus, estimating the severity level of occupational accidents has received considerable attention recently, particularly in the construction industry [6], [8], [14], [15]. In most of these studies, the accident severity was measured from the lowest severity level, such as near misses or first aid cases, to the highest severity level, such as fatalities.

In the construction safety literature, most researchers used binary classification to predict the severity level of construction accidents. Among the preliminary studies, Esmaeili et al. [204] developed a logistic regression (LR) model to predict the severity outcomes of construction accidents (fatal or non-fatal) by using a dataset collected from Occupational Safety and Health Administration (OSHA) comprising 1,436 accident instances. Based on the data collected from the same institution with 594 fatal and 819 non-fatal cases, Mistikoglu et al. [1] developed a decision tree (DT) model to predict the severity level of roofers with an accident type specific to falling from height. Similarly, Choi et al. [15]

used several ML methods such as LR, DT, random forest (RF), and AdaBoost to predict the severity outcome (fatal or non-fatal) of construction accidents based on national data (2,846 fatal and 137,323 non-fatal). Their study showed that the RF model achieved the highest classification accuracy. Ayhan and Tokdemir [6] classified the severity of accidents into two categories, namely low-severe (including at-risk behavior, near-miss, and non-day lost cases) and high-severe (including first aid cases, medical intervention, the day lost cases, and fatalities). The prediction model was developed with an artificial neural network (ANN) based on 4,109 cases collected from construction companies. Baker et al. [14] measured the severity of injuries with respect to the level of immediate medical care required, such as first aid and medical case. By using over 90,000 injury reports, the researchers developed XGBoost, RF, support vector machine (SVM), and XGBoost + RF models to predict severity level of construction accidents, as well as other outcomes such as injury type, body part, and incident type.

The severity level of construction accidents was also predicted within the context of a multi-class problem. For instance, Lee et al. [8] considered severity including three classes: slight injury, serious injury, and fatality. The researchers developed SVM and DT-based ensemble models subjected to data collected from a large construction company in Korea (with 963 accident cases). They predicted the accident severity and other outputs such as the cause of the accident and injured body parts. Tixier et al. [7] used 5,298 injury reports taken from contractors to predict energy source, injury type, body part, and accident severity. The researchers used RF and stochastic gradient tree boosting (SGTB) to develop prediction models in which severity level was categorized in three classes: pain/first aid, medical care, and lost work time. From another perspective, Antwi-Afari et al. [36] collected data from wearable insole pressure sensors to predict construction workers' loss of balance. They tested several classifiers, including ANN, DT, RF, K-Nearest neighbor (KNN), and SVM, with output classes of the simple walk, slip, trip, unexpected step-down, and twisted ankle. However, most of these studies praised the ML process more than designing the data before prediction with a comprehensive preprocessing. None of them examined the effects of different preprocessing scenarios, while most of them compared the prediction accuracies of various ML models such as [8], [14], [15], which constituted the point of departure of this study.

5.2.2 Data Preprocessing in Machine Learning

Preprocessing is one of the most crucial tasks affecting the accuracy of prediction models [65] as most datasets include missing values, different scales of features, and noisy data with outliers [87]. Researchers and data scientists usually spend on average 60% of their efforts in preprocessing for building the model as “no quality data, no quality results” [196]. Swarna-Priya [84] argue that most of the studies in the literature focus on building the ML model without paying the required attention to the preprocessing stage. Some of the most widely adopted preprocessing steps and their descriptions appear in Table 5.1.

In the construction safety literature, one of the most widely adopted data preprocessing methods is DRSM. The researchers used several resampling techniques such as ROS [15] and RUS [11] to deal with the class imbalance issue, which is inherent in accident datasets. There are also some studies that increased the dimension of the captured data by the number of elements of some categorical features with a one-hot encoding procedure to represent the categorical features in a more interpretable way [193], [197]. Besides, OUTH is frequently addressed as a preprocessing approach in the literature. In this context, some researchers select the most adequate ML model according to many criteria including their robustness to outliers [7], while others choose to treat outliers based on several outlier detection methods such as interquartile range (IQR) [4]. Similarly, rows including missing values are usually removed from the datasets [204], while some researchers use the median substitution method [207], mean substitution method [88], or ML algorithms for missing data imputation [65]. DBIN is another preprocessing technique applied to mainly numerical features, such as the age of worker or size of company [15]. On the other hand, DSCL methods, which are used for standardizing the numerical features, were mainly adopted in other research fields [86], [208]. Although researchers in construction safety used certain preprocessing steps, the pertinent literature lacks both a multi-step comprehensive preprocessing and scenario-basis automated preprocessing.

Table 5.1 Some preprocessing elements and their descriptions

Reference	Preprocessing step	Description	Potential scenarios
Walther et al. [205]	Missing data handling (MDH)	Most data include missing values due to several reasons such as human errors or data interruptions. MDH aims to deal with missing values in the dataset to improve prediction performances.	i) Remove missing data, ii) Consider missing data as mean or median for numerical features, iii) Consider missing values as "other" for categorical features.
Huang and Meng [206]	Data binning (DBIN)	DBIN is a method of dividing continuous features into a series of discrete ranges to prevent over-fitting. However, some information in the dataset is lost when binning the data.	i) Do not apply DBIN, ii) Apply binning to numerical features.
Erharter et al. [197]	Outlier handling (OUTH)	Outliers might reduce the prediction power of ML models. OUTH approaches deal with anomalies or outliers to represent the dataset in a more interpretable way.	i) Do not apply OUTH, ii) Remove outliers, iii) Cap outliers at the upper and lower limit of corresponding features.
Dietrich et al. [195]	Data scaling (DSCL)	ML models may not perform well if numerical features are scaled at different intervals. DSCL aims to scale numerical features in a similar sense by zooming in or out.	i) Do not apply DSCL, ii) Standard scaler, iii) Min-max scaler.
Sarker et al. [83]	Data encoding (DENC)	ML algorithms work well with numerical values while captured data might include categorical features. Thus, DENC aims to convert elements in the categorical features into numerical values.	i) Label encoding, ii) One-hot encoding.
Chi et al. [109]	Data resampling (DRSM)	If the dataset is imbalanced, ML methods may yield erroneously higher prediction accuracies if classify all the labels as majority class. Therefore, DRSM techniques should apply to the training set to minimize bias towards the majority classes.	i) Random under-sampling (RUS), ii) Random over-sampling (ROS), iii) Synthetic minority over-sampling technique (SMOTE).

Other than the construction safety literature, some researchers adopted a systematic preprocessing approach to boost the prediction accuracy of the proposed models in

different fields. For instance, Erharter et al. [197] used a sequential preprocessing approach as a foundation for cone penetration test classification. The researchers removed outliers, scaled the data between 0 and 1, balanced the soil types with SMOTE algorithm, and then applied the one-hot encoding scheme to deal with nominal-categorical data. Wang et al. [209] developed a prediction model for corporate misconduct detection and used three preprocessing techniques that were data binning, one-hot encoding, and SMOTE to deal with skewed variables, categorical variables, and imbalance classes, respectively. Li et al. [169] subjected several preprocessing techniques such as outlier and noise data elimination, missing data imputation and data resampling into an eye movement dataset to classify the mental fatigue of construction equipment operators. However, there are no distinct rules to decide when to use specific preprocessing techniques, which techniques to combine in a scenario, and how [210]. A complementary mixture of preprocessing methods and their combinations are suggested in the literature as they provide all the information linked to a distinct preprocessing step. Such a process takes the users out of the tedious loop of searching for the best preprocessing scenario [211]. In this vein, one of the preliminary studies that adopted a comprehensive preprocessing approach was performed by Dietrich et al. [195]. The researchers implemented an iterative preprocessing procedure including MDH, OUTH, DSCL, and DENC to predict the short-term electric load of production machines. They examined the data preprocessing strategies as follows: MDH with mean and median; OUTH with no treatment and median absolute deviation from the median; DSCL with standard scaler, min-max scaler, robust scaler; and DENC with one-hot encoding. However, up to the authors' knowledge, studies in the literature neither proposed a scenario-basis automated preprocessing model, nor integrated such a model with ML applications for predicting future events.

5.3 Data Description

The national dataset used in this study was collected from the Social Security Institution, in Turkey, which comprises 388,907 instances (after initial preprocessing, *Section 5.4.1*). These included 279,830 first aid (severity level=1), 96,159 temporary incapacity (severity level=2), 8,701 permanent incapacity (severity level=3), and 4,217 fatality (severity level=4) cases. Besides considering this classification in most data analyses, low severity level (severity level: 1 and 2) and high severity level (severity level: 3 and 4) were also emphasized during the data description and analyses. After diligent investigation of the

data, 15 features were identified for accident severity prediction (Figure 5.2). Some features that were accessible in the dataset, such as injured body part or incident type, were not included since they can be regarded as post-accident data and considered as outcomes rather than input features [66]. In this study, features were identified from pre-accident data to allow proactive safety management practices in construction sites [189]. Included features can be classified into three categories as numerical features, categorical features, and time-related features, which are explained in the following sub-sections.

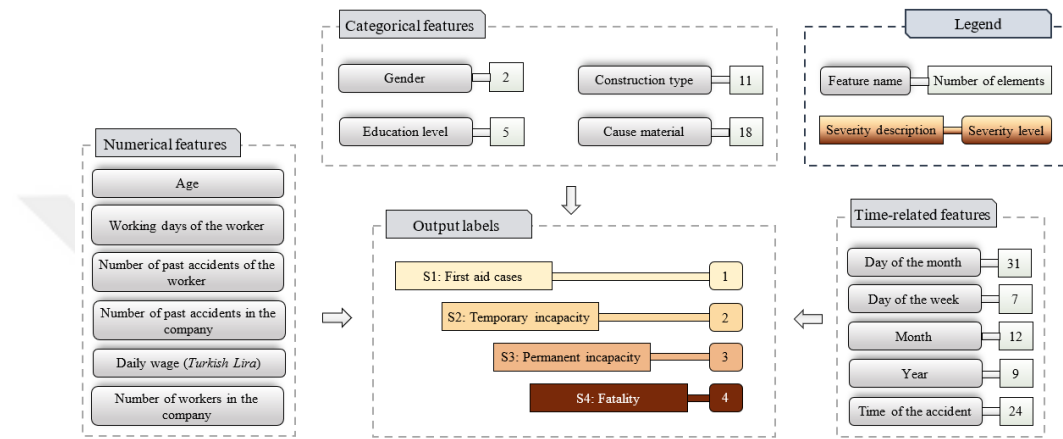


Figure 5.2 Features used in the prediction model

5.3.1 Numerical Features

Table 5.2 shows the severity level of construction accidents by each numerical feature. It was observed that with the increase in *the age* of the worker, the severity level of accidents also increased. Similarly, permanent incapacity and fatality rates of accidents increased as the age of the worker increased, from 0.93% to 4.06% and from 0.65% to 2.7%, respectively. *Working days of the worker* denotes the number of days passed after the first employment of the worker (in the corresponding company), and it is not surprising that the workers who have recently started (<10 days) had a higher probability of experiencing permanent incapacity (4.2%) and fatality (1.89%). Regarding the *number of past accidents of the worker*, it is essential to note that workers who experienced construction accidents before were more likely to experience accidents that would result in first aid only. 69.94% of the workers who had not experienced an occupational accident experienced accidents with first aid. The rate reached up to 98.31% for workers who had experienced at least four accidents. A similar tendency was observed considering the *number of accidents in the company* and *the number of workers in the company*. There were positive and negative relationships of the number of past accidents in the company

(from 1 to ≥ 100) with the first aid (from 63.80 to 88.65%) and fatality (from 2.13 to 0.17%) cases, respectively. Likewise, the highest mortality (2%) and permanent disability rates (3.87%) were experienced by workers working in a company employing less than 10 workers. Regarding *the daily wage*, there was a considerable difference in the percentages of workers who experienced accidents that led to temporary or permanent incapacity. Higher the daily wage of a worker is (from <50 to ≥ 150), less likely the worker needed medical intervention, in other words, experienced temporary (from 35.08 to 12.37%) or permanent incapacity (from 4.46 to 0.96%).

Table 5.2 Number of accidents by severity level with respect to numerical features

Numerical features	Breakdown	First aid cases (%)	Temporary incapacity (%)	Permanent incapacity (%)	Fatality (%)
Age	<20	79.13	19.29	0.93	0.65
	20 – 29	75.27	22.91	1.14	0.68
	30 – 39	70.61	26.08	2.41	0.90
	40 – 49	67.61	27.67	3.40	1.32
	≥ 50	67.98	25.27	4.06	2.70
Working days of the worker	<10	69.58	24.33	4.20	1.89
	10 – 49	73.30	23.72	2.06	0.92
	50 – 149	71.41	25.68	1.98	0.93
	150 – 249	72.25	25.07	1.75	0.93
	≥ 250	72.35	25.13	1.53	0.99
Number of past accidents of the worker	0	69.94	26.43	2.38	1.25
	1	77.77	19.96	1.80	0.47
	2 – 3	90.71	7.31	1.10	0.88
	≥ 4	98.31	1.31	0.31	0.08
Number of past accidents in the company	1	63.80	30.29	3.78	2.13
	2 – 3	68.68	27.65	2.51	1.16
	4 – 9	73.05	24.75	1.58	0.62
	10 – 99	78.50	20.03	1.13	0.35
	≥ 100	88.65	10.64	0.54	0.17
Daily wage (Turkish Lira)	<50	58.65	35.08	4.46	1.82
	50 – 74.99	64.26	32.21	2.42	1.11
	75 – 149.99	83.49	14.88	1.05	0.57
	≥ 150	85.60	12.37	0.96	1.08
	Missing data	89.85	7.81	0.94	1.39
Number of workers in the company	<10	63.51	30.63	3.87	2.00
	10 – 49	66.81	28.65	3.04	1.50
	50 – 249	72.23	24.94	1.93	0.89
	≥ 250	80.99	17.55	1.00	0.46
	Missing data	94.06	4.56	0.81	0.57

5.3.2 Time-related Features

Since accident outcomes are not independent of time, many researchers included several time-related features in the input layer such as hour, day of the week, month, and year [8], [15], [65] to predict the accident outcomes. In this study, five time-related features were included, which are day of the month, day of the week, month, year, and time of the accident (hour). The distribution of the severity level of accidents by time-related features is shown in Table 5.3. Highest percentages regarding permanent incapacity (2.35%) and fatality (1.17%) were observed during the first week of the month. The data classified by *day of the week* also showed that the severe accident rates were highest at the weekend by 2.45 and 2.44% of permanent incapacity, and 1.25 and 1.29% of fatality for Saturdays and Sundays, respectively. On the other hand, fatality and permanent disability rates were the lowest on Thursdays and Wednesdays, respectively.

According to the *month* classification, the mortality rate was the highest in September (1.18%) and lowest in February (0.94%). Higher percentages of the accidents resulting with temporary incapacity were observed during winter, while it was summer and fall regarding first aid cases. Considering the *year* of the accidents, the rate of the severity of accidents increased steadily in terms of first aid cases from 49.52% in 2015 to 98.42% in 2020. It was also realized that the rate of workers experiencing more severe accidents steadily decreased for temporary incapacity (from 45.37 to 0.37%), permanent incapacity (from 3.57 to 0.42%), and fatality (from 1.54 to 0.59%) between 2015 and 2020. Statistics depicting a clear decreasing trend in high-severe accidents can be related to the increasing occupational health and safety (OHS) expenditure in the country [212]. The last time-related feature was *the time of the accident*. Notably, the rates of permanent disability and fatality were the highest at night hours by 2.28% and 1.56% between 00.00-03.59, and 2.57 and 2.65% between 04.00-07.59, respectively.

5.3.3 Categorical Features

The severity level distributions of four categorical features, namely cause material, construction type, education level, and gender are shown in Figure 5.3, Figure 5.4, Figure 5.5, and Figure 5.6. Figure 5.3 shows that the safest materials were *dust, silt, grain, etc.*, such that only 0.14 and 0.78% of them resulted in fatality or permanent incapacity, respectively. In addition, the lowest permanent disability occurred in accidents with *office materials and equipment*. The construction workers in Turkey experienced the highest

rate of fatality when engaged in *construction machinery* (such as crane and excavator) (2.5%) and *stairways, ladder, and scaffold* (2.42%). Although the fatality rate of *gas, explosive and flammable materials* (1.27%) were not among the most dangerous five materials in this context, the rate of permanent disability was highest (4.53%) for those working with or exposed to these materials.

Table 5.3 Number of accidents by severity level with respect to time-related features

Time-related features	Breakdown	First aid cases (%)	Temporary incapacity (%)	Permanent incapacity (%)	Fatality (%)
Day of the month	1 – 7	71.50	24.98	2.35	1.17
	8 – 14	72.18	24.58	2.16	1.08
	15 – 21	72.58	24.17	2.19	1.05
	22 – 31	71.63	25.08	2.25	1.05
Day of the week	Monday	70.94	25.80	2.16	1.09
	Tuesday	72.07	24.64	2.25	1.04
	Wednesday	71.84	24.97	2.21	0.98
	Thursday	71.89	25.04	2.05	1.02
	Friday	71.98	24.72	2.23	1.07
	Saturday	72.76	23.55	2.45	1.25
	Sunday	72.76	23.51	2.44	1.29
Month	January	65.95	30.79	2.23	1.03
	February	66.01	30.92	2.14	0.94
	March	66.37	30.28	2.29	1.05
	April	68.88	27.75	2.29	1.08
	May	72.94	23.81	2.25	1.00
	June	72.94	23.75	2.19	1.12
	July	78.75	18.04	2.11	1.10
	August	74.09	22.43	2.35	1.12
	September	77.39	19.27	2.16	1.18
	October	74.56	22.20	2.13	1.11
	November	69.81	26.52	2.51	1.16
	December	70.01	26.71	2.21	1.07
Year	2012	88.44	4.78	4.50	2.27
	2013	51.86	41.88	4.36	1.90
	2014	57.71	36.96	3.62	1.70
	2015	49.52	45.37	3.57	1.54
	2016	56.24	39.87	2.70	1.19
	2017	59.46	37.52	2.11	0.92
	2018	83.66	14.20	1.45	0.70
	2019	96.30	2.29	0.81	0.61
	2020	98.42	0.57	0.42	0.59
Time of the accident	00.00 – 03.59	74.84	21.32	2.28	1.56
	04.00 – 07.59	72.71	22.07	2.57	2.65
	08.00 – 11.59	70.24	26.58	2.27	0.90
	12.00 – 15.59	72.11	24.56	2.26	1.07
	16.00 – 19.59	73.73	22.82	2.21	1.24
	20.00 – 23.59	77.96	19.46	1.56	1.03

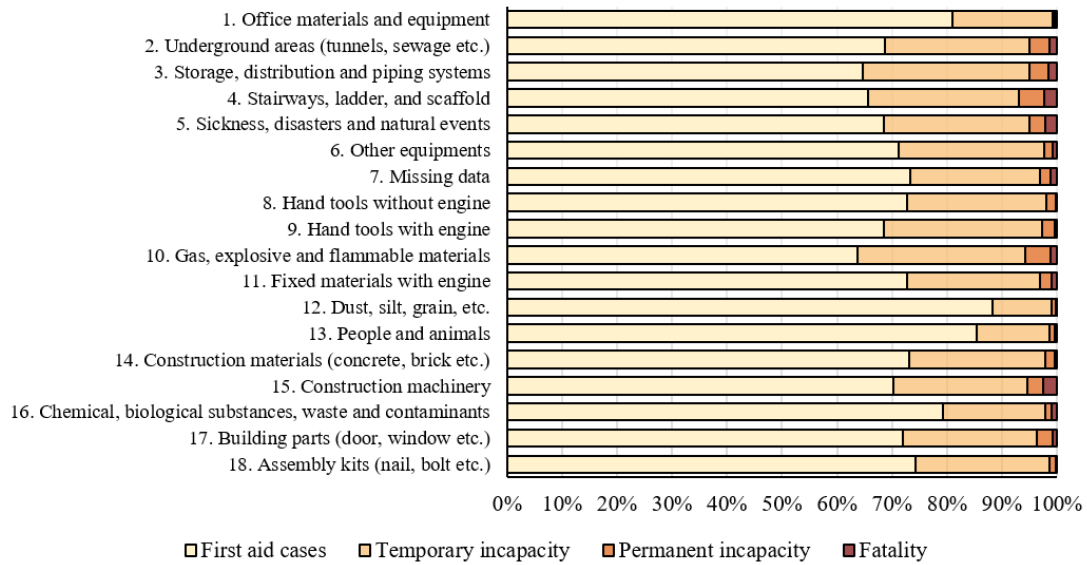


Figure 5.3 Cause material with accident severity level

Figure 5.4 shows 11 types of construction by accident severity level. Considering the fatality rates, workers were at more risk when working in *harbor and dam construction* by 1.8%. Additionally, construction workers have experienced the highest rate of permanent disabilities during *construction works in residential* (2.8%) and *non-residential projects* (2.6%). In terms of low severity level, the *earthworks* is perhaps one of the safest types of construction according to the first aid case rate (79.9%).

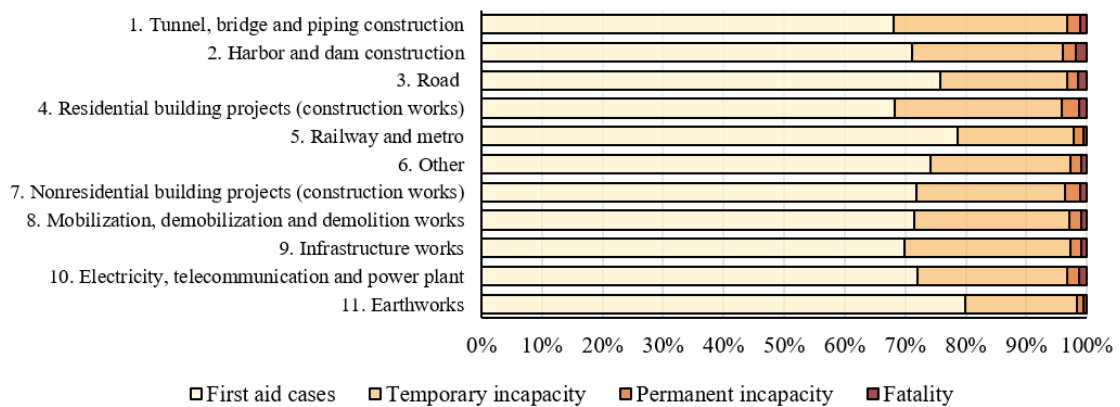


Figure 5.4 Construction type with accident severity level

The distribution of accident severity level regarding the education level is illustrated in Figure 5.5. Workers who were *illiterate or only literate* and with an education level of *elementary school* had higher probabilities of experiencing severe accidents, by 2.47 and 2.62% considering permanent disability, and 1.17 and 1.25% regarding fatality rates,

respectively. On the other hand, workers with an education level of *secondary school* experienced the lowest fatality rate.

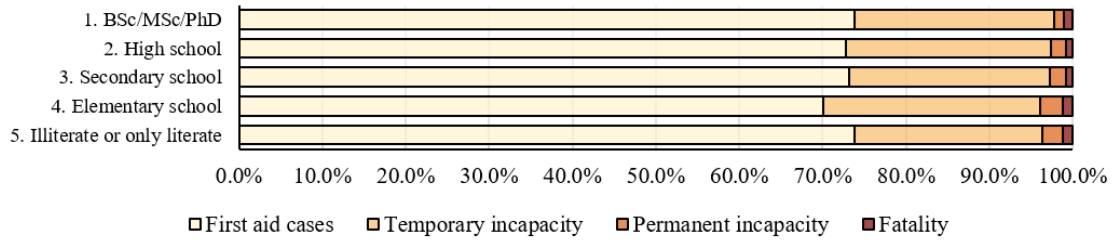


Figure 5.5 Education level with accident severity level

The last categorical feature included in this study was gender, with a significant imbalance between cases of *female* and *male* victims (Figure 5.6). Only 0.7% of the victims were females in the construction industry in Turkey. The figure also depicts that the male workers had a higher probability of experiencing permanent incapacity (2.25%) and fatality (1.09%) than female workers by 0.34 and 0.41%, respectively.

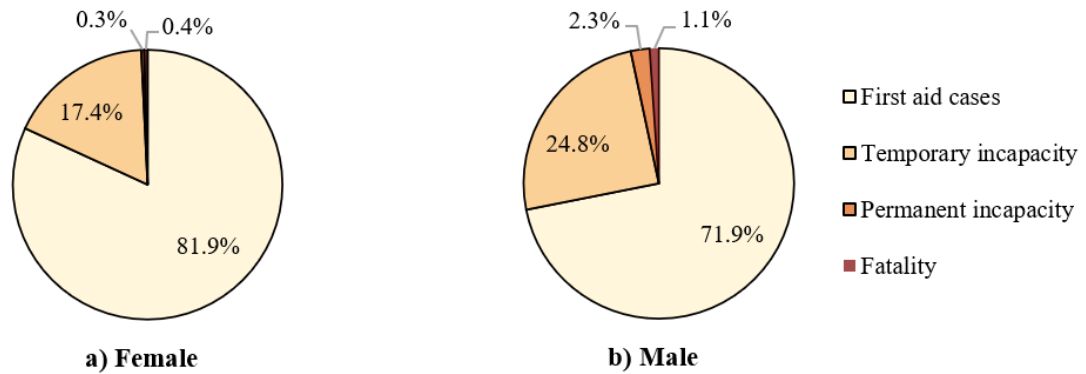


Figure 5.6 Gender with accident severity level

5.4 Model Development

Research steps conducted in this study are shown in Figure 5.7 and explained in the following sub-sections. All experiments were performed in *Python 3.8*, and the following main packages were used: *scikit-learn v0.24.1*, *numpy 1.18.5*, *pandas 1.1.3* for preprocessing, performance evaluation, and permutation importance; *xgboost v1.3.3* for ML application; *imbalanced-learn v0.7.0* for data resampling; and *seaborn v0.10.1* for visualization.

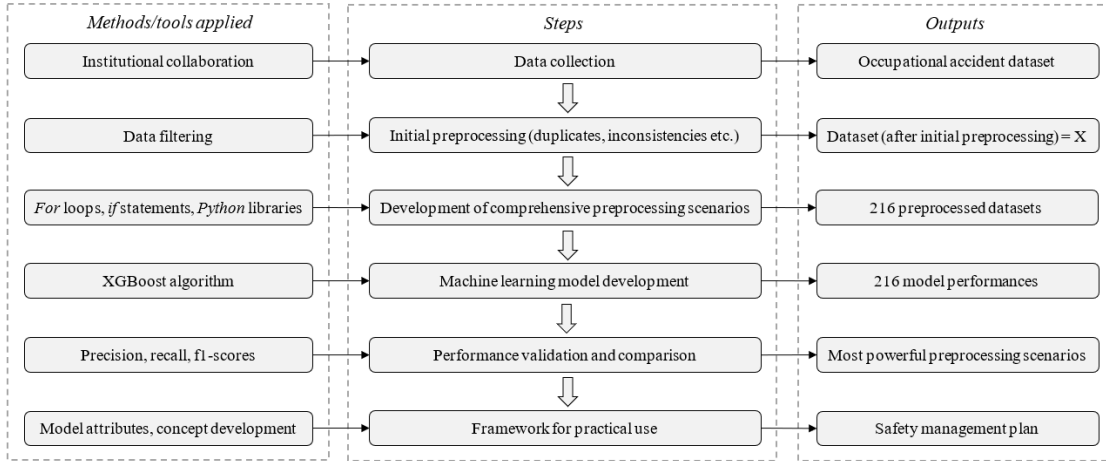


Figure 5.7 Research steps

5.4.1 Initial Preprocessing

The dataset initially included 395,777 instances (accident victims) that occurred between 1987 and 2020. However, based on careful examination of the dataset, five major issues were observed.

- There were just a few instances between 1987 and 2011.
- Missing values were detected in some of the instances regarding more than 50% of the features.
- There were some inconsistencies (such as fatal accident with permanent disability income or entries such as workers with negative age) in some instances.
- Duplicates were found in the dataset.
- Some elements in categorical features included less than 2% of the data.

Therefore, initial preprocessing (before automated preprocessing) was needed according to the issues mentioned above. Table 5.4 describes the context of the initial preprocessing steps. Correspondingly, the number of accident victims was reduced from 395,777 to 388,907 by removing 6,870 instances, which accounted for less than 2%.

Then, the dataset comprising 388,907 victims was investigated based on missing values and outliers. In this regard, four simple scenarios were considered in the preliminary step to check the adequacy of the research question (Table 5.5). It was observed that if the cases only with missing values were removed, there would be only 236,495 instances. When the cases with only outliers were removed, the size of the data was reduced to 250,940 and 251,833 while imputing missing numerical values as median and mean, respectively. By combining these, if the cases with both missing values and outliers were

removed, then the size of the data reduced to 151,882, which accounts for nearly 40% compared to 388,907 instances. These facts also highlighted the need to consider an automated preprocessing model to compare the prediction power of the combination of each preprocessing pipeline.

Table 5.4 Initial preprocessing steps before comprehensive preprocessing

Step no	Scope of the preprocessing
Step 1	Cases recorded before 2012 were removed from the dataset since there were just a few cases recorded in each year between 1987 and 2011.
Step 2	Any case that included missing values more than 50% of the features were removed.
Step 3	Cases with inconsistent and irrational values were removed.
Step 4	Duplicates were removed.
Step 5	Elements of categorical features were regrouped if they read less than 2% of the data.

Table 5.5 Number of output labels

Output labels	Severity level	Original labelled cases	Cases after initial pre-processing	If removing cases only with missing values	If removing cases only with outliers (missing values = median)	If removing cases only with outliers (missing values = mean)	If removing cases with both missing values and outliers
First aid cases	1	285,127	279,830	167,226	171,218	171,879	101,860
Temporary incapacity	2	97,476	96,159	61,060	69,288	69,483	43,376
Permanent incapacity	3	8,936	8,701	5,740	7,112	7,123	4,705
Fatality	4	4,238	4,217	2,469	3,322	3,348	1,941
Total	–	395,777	388,907	236,495	250,940	251,833	151,882

5.4.2 Adopted Preprocessing Steps for Scenario Development

In this study, six consecutive preprocessing steps were considered. These steps are MDH (3 options), DBIN (2 options), OUTH (3 options), DSCL (3 options), DENC (3 options), and DRSM (2 options). Preferred for their diverging objectives, the steps are referred as *a*, *b*, *c*, *d*, *e*, and *f*, respectively. These steps are shown in Figure 5.8 and briefly explained in the following sub-sections.

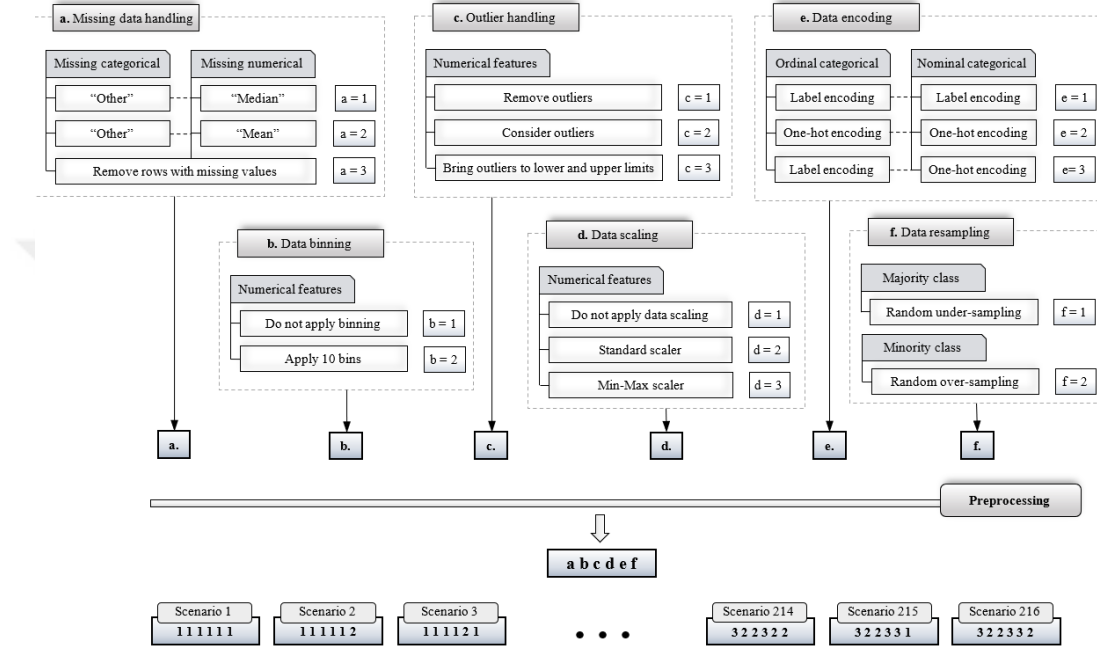


Figure 5.8 Scenario development process.

5.4.2.1 Missing Data Handling (MDH: a)

Missing value problem is one of the most common issues in developing ML models. The easiest way to handle the missing values is to remove the cases with missing data [83]. There are two other approaches to deal with missing values in numerical features. The first one is that missing values of a feature are searched and replaced with the mean value of the feature [88]. The second approach addresses imputing missing values with the median value. Meanwhile, missing values in categorical features can be coded as “other”, which aids in performing label or one-hot encoding [195]. All three potential options of MDH are addressed in this paper as shown in Figure 5.8: i) include categorical missing data as “other”, numerical missing data as median ($a=1$); ii) include categorical missing data as “other”, numerical missing data as mean ($a=2$); and iii) remove instances including missing values ($a=3$).

5.4.2.2 Data Binning (DBIN: b)

Splitting continuous features into a series of distinct ranges, referred to as bins, is called data binning [206]. Although it is not always referred to as “data binning,” this approach has widely been used in the construction safety literature. For instance, Choi et al. [15] divided the feature “age” into 14 discrete intervals, while Kakhki et al. [213] considered the age of the worker as a continuous feature. Two options of DBIN are addressed in this study (Figure 5.8): i) do not apply data binning ($b=1$), and ii) split the numerical features into 10 bins ($b=2$). Since the values corresponded to *the number of past accidents of the worker* and *the number of accidents in the company* were severely imbalanced, it was inappropriate to create homogeneous bins for these features, and thus all the numerical features except these two were included in DBIN option 2 ($b=2$).

5.4.2.3 Outlier Handling (OUTH: c)

In this study, the IQR method was adopted to detect outliers. Accordingly, the lower limit (LL) and upper limit (UL) of numerical features were determined as follows [4]:

$$LL = Q1 - 1.5 \times IQR \quad (5.1)$$

$$UL = Q3 + 1.5 \times IQR \quad (5.2)$$

where Q1, Q3, and IQR are the 25th percentile, 75th percentile, and inter-quartile range, respectively. Three options are considered for OUTH in this study (Figure 5.8): i) remove outliers that are out of the range between LL and UL ($c=1$); ii) consider outliers ($c=2$); and iii) consider outliers that are out of the range between LL and UL by replacing them with the value of LL and UL, respectively ($c=3$). All the numerical features except the *number of past accidents of the worker* were included in outlier treatment (OUTH option 1 and 3; $c=1$ and $c=3$), since handling all instances in this feature lead to no single case (if $c=1$) or same value for each case (if $c=3$) to run the ML model.

5.4.2.4 Data Scaling (DSCL: d)

Numerical features were scaled at different intervals in the obtained dataset. The most commonly adopted DSCL methods are *Standard scaler* and *Min-Max scaler* [85]. Numerical features are converted in a way that leads to a distribution with a mean value of 0 and a standard deviation of 1 in the *Standard scaler* method [87]. Alternatively, numerical features are scaled into a discrete range, generally 0-1, in the *Min-Max scaler*

method [86]. Accordingly, three options are addressed for DSCL in this study (Figure 5.8): i) do not apply data scaling methods ($d=1$); ii) apply *Standard scaler* ($d=2$); and iii) apply *Min-Max scaler* ($d=3$).

5.4.2.5 Data Encoding (DENC: e)

The dataset contained nominal-categorical (construction type and cause material) and ordinal-categorical (other categorical features except gender) features. The literature suggests label encoding and one-hot encoding approaches. Expressions in a categorical feature are encoded in an ordered way through label encoding [83]. Alternatively, new binary features are created by one-hot encoding for each element in a categorical feature [84]. This study considered three options regarding DENC (Figure 5.8): i) apply label encoding for both ordinal and nominal features ($e=1$); ii) apply one-hot encoding for both ordinal and nominal features ($e=2$); and iii) apply label encoding for ordinal and one-hot encoding for nominal features ($e=3$).

5.4.2.6 Data Resampling (DRSM: f)

The analyzed data was highly imbalanced. According to the dataset, the outcomes of construction accidents with first aid cases were 66, 32, and 3 times higher than that of fatality, permanent incapacity, and temporary incapacity cases, respectively. The most commonly used methods to deal with imbalanced class distribution are RUS, ROS, and SMOTE [15]. Instances of the majority classes are randomly removed from the data within the RUS algorithm [11], while the ROS algorithm is based on randomly repeating the instances of the minority classes [92]. On the other hand, cluster-based interpolation is performed to generate synthetic instances in SMOTE algorithm [93]. This study assessed two DRSM strategies (Figure 5.8): i) apply RUS ($f=1$); and ii) apply ROS ($f=2$). Since SMOTE performs interpolation on numerical features, it was not considered here, as the dataset included many categorical features [15].

5.4.3 Scenario Development – Comprehensive Pre-processing

In this study, different combination trials were generated based on several techniques in each preprocessing step as part of the automated preprocessing leading to 216 unique scenarios. For instance, the first scenario included the following preprocessing techniques: MDH: Include categorical missing data as other, numerical missing data as median ($a=1$); DBIN: Do not consider binning ($b=1$); OUTH: Remove outliers ($c=1$); DSCL: Do not consider data scaling ($d=1$); DENC: Consider label encoding ($e=1$); and

DRSM: Resample the data with RUS ($f=1$). This scenario was referred to as “scenario ID: 111111” by combining the values denoted as a , b , c , d , e , and f , respectively (Figure 5.8).

To perform the preprocessing, several functions/classes in *Python 3.8* were used in a closed for-loop structure. The details of the preprocessing are illustrated in Figure 5.9. First, a structure that included 6 nested for loops were coded to enable performing 6 consecutive preprocessing steps (i.e., a , b , c , d , e , and f). In this step, *Range* function was used between 1 and 3 for a (MDH), c (OUTH), d (DSCL), e (DENC), and between 1 and 2 for b (DBIN), f (DRSM). Then, the preprocessed dataset (after initial preprocessing, Section 5.4.1) was introduced to the code structure and denoted as *Dataset X*. After preparation of the code for the comprehensive preprocessing, several *if*, *elif*, *else*, and *break* statements were used and six new datasets were generated after implementing each preprocessing step (i.e., *Dataset A*, *B*, *C*, *D*, *E*, *F* for preprocessing steps a , b , c , d , e , f , respectively). For instance, if $a = 3$, then the categorical and numerical missing values were removed from the dataset using *dropna* function and a new dataset was generated as *Dataset A* (Figure 5.9, Step 1). After coding three *if* statements for MDH (i.e., $a = 1, 2, 3$), the next step was coded regarding data binning (Figure 5.9, Step 2). For example, if $b = 1$, then the new dataset (i.e., *Dataset B*) was coded as equal to the prior dataset (i.e., *Dataset A*) since $b = 1$ meant not applying data binning. Similarly, the *Dataset C* was formed in Step 3 based on 3 OUTH strategies, and the *Dataset D* was constituted according to the examined DSCL strategies (Figure 5.9, Step 4). In this step, *StandardScaler* and *MinMaxScaler* classes of *scikit-learn* library were used. In the next step, the categorical variables in the *Dataset D* were treated with one of the DENC strategies using *LabelEncoder*, *OneHotEncoder*, or both classes of *scikit-learn* library (Figure 5.9, Step 5), which formed the basis of the next dataset (i.e., *Dataset E*). At the final preprocessing step, DRSM strategies were applied using either *RandomUnderSampler* or *RandomOverSampler* classes of *imbalanced-learn* library (Figure 5.9, Step 6) and the final dataset (i.e., *Dataset F*) was generated accordingly. Note that the *Dataset E* was divided into two groups as the training and testing sets before the *if* statements in Step 6 and then DRSM strategies were applied to the training sets only as a rule of thumb. It should also be noted that there would be 324 preprocessing scenarios ($3 \times 2 \times 3 \times 3 \times 3 \times 2$) if all the combinations were run. However, as outlier treatment is inappropriate in the case of data binning, the only potential value of c was set as 2 when

b equals 2 (using nested *if* and *elif* statements). Thus, the number of scenarios generated was 216 $[(3 \times 1 \times 3 \times 3 \times 3 \times 2) + (3 \times 1 \times 1 \times 3 \times 3 \times 2)]$. With the preprocessed Dataset F , XGBoost method was employed using the *XGBClassifier* of *xgboost* library and the performance of each scenario was attained to perform comparative analysis. A brief description of the employed XGBoost method is provided in the following section.

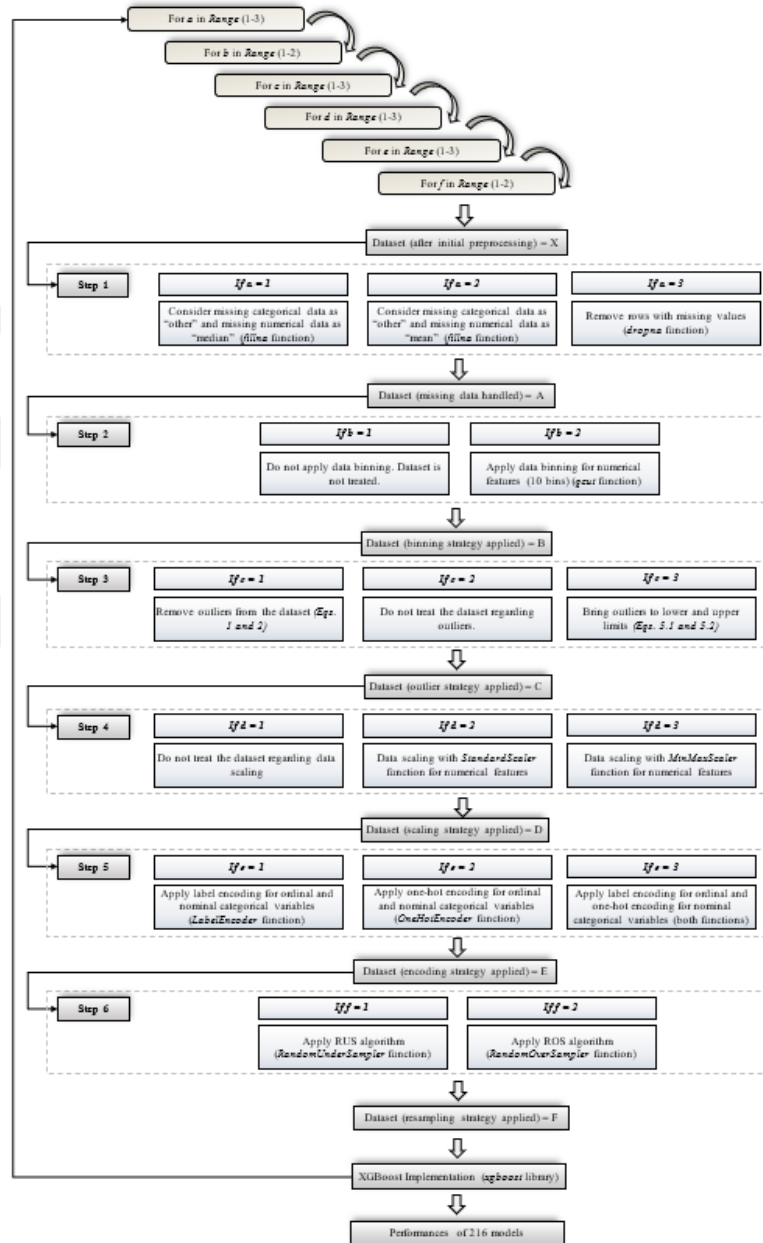


Figure 5.9 Illustration of preprocessing process

5.4.4 XGBoost Method

The proposed model was tested with the XGBoost algorithm. XGBoost was proposed by Chen and Guestrin [214] as an integrated learning method based on gradient boosting,

which aims to achieve an accurate classification via calculations of weak classifiers iteratively [48]. Basically, two main concepts define the objective function of the XGBoost method: the training error and regularization [54]. Accordingly, the decision rules that find an optimum value for the objective function are searched in the XGBoost method using gradient descent optimization method [215]. The key idea of the method is to learn and improve based on the prediction error [206]. Therefore, base learners are formed according to the generated decision trees, which are trained to straighten the errors through parallelization attribute of the method [216]. This parallel computing decreases the time required for learning and exploring the model [214]. In addition, XGBoost adds a regularization term to the loss function to penalize the complication of the model while implementing several optimization tricks to reduce computational time for training [14]. By searching the complexity of the model, the regularization function also addresses the issue of overfitting [54].

5.4.5 Performance Evaluation Metrics

This study evaluated the performances of each scenario using various metrics such as confusion matrix, precision, recall, and f1-score. Confusion matrix C is a square matrix of dimension $N \times N$, where N indicates the number of classes. The higher the diagonal values of the C , the better the prediction power. On the other hand, off-diagonal values are those that are predicted incorrectly. From the confusion matrix, precision, recall, and f1-score of each class can be computed [66]. Therefore, confusion matrices calculated according to the predictions of XGBoost method in each scenario were captured for comparative analysis. Precision is a measure of how accurate the positive predictions are, while recall measures how many of the actual positive instances are identified by the model [4]. Precision and recall are calculated through Eqs. (5.3) and (5.4).

$$Precision = \frac{C_{i,i}}{\sum_{j=1}^N C_{j,i}} \quad (5.3)$$

$$Recall = \frac{C_{i,i}}{\sum_{j=1}^N C_{i,j}} \quad (5.4)$$

where $C_{i,j}$ of the matrix C is the number of observations known to be in class i while predicted to be in class j . Finally, the f1-score is calculated by combining precision and recall values (through harmonic average of them) as follows:

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (5.5)$$

5.5 Results

5.5.1 Results of the Scenario-basis Preprocessing

This study evaluated and compared the performances of 216 preprocessing scenarios that were subjected to the XGBoost method. The parameters of the XGBoost were identified based on the trial-error approach, and the learning rate, number of trees, and minimum child weights were determined as 0.1, 90, and 1, respectively. Each preprocessing trial is evaluated with the same parameters to make the comparisons fair. Table 5.6 shows the scenarios which achieved the highest prediction accuracies with respect to precision, recall, and f1-score values, along with the first and last three scenarios to enable the interpretation of the process. The table also provides an insight into the scenario scopes to describe the automated scenario development process better. Analysis results show that the best scenarios were identified as 222321 and 313132, regarding precision (0.7854) and recall values (0.5447), respectively (Table 5.6). On the other hand, the scenario labelled as 212322, which was the 106th trial, achieved the highest f1-score by 0.6092. Since f1-score provides an overall assessment scheme especially when the class imbalance issue exists, the following preprocessing techniques are proposed for future studies predicting the outcomes of occupational accidents (scenario ID: 212322; trial ID: 106): i) Including categorical missing data as “other”, and numerical missing data as the mean of the feature (MDH: 2); ii) Not considering binning (DBIN: 1); iii) Neither implementing outlier treatment nor removing them from the data (OUTH: 2); iv) Considering Min-Max scaler as a data scaling (DSCL: 3); v) Adopting one-hot encoding for all the categorical features (DENC: 2); and vi) Resampling the data with ROS algorithm (DRSM: 2). Confusion matrices of the best scenarios (according to f1-score, precision, and recall) are also illustrated in Figure 5.10. The findings show that temporary incapacity (212322: 61.42%; 222312: 62.25%; 313132: 62.61%) and first aid cases (212322: 52.24%; 222312: 50.91%; 313132: 52.43%) were predicted by the best performing scenarios at higher accuracies compared to other and more severe output labels.

Table 5.6 Performances of first, best and last scenarios

Trial ID	Scenario ID	Scenario description	Scenario scope	Precision	Recall	F1-score	Support
1	111111	Fist scenario	MDH: Include categorical missing data as other, numerical data as median DBIN: Do not consider binning OUTH: Remove outliers DSCL: Do not consider data scaling DENC: Consider label encoding DRSM: Resample data with RUS	0.7365	0.4691	0.5423	75,282
2	111112	Second scenario	MDH: Include categorical missing data as other, numerical data as median DBIN: Do not consider binning OUTH: Remove outliers DSCL: Do not consider data scaling DENC: Consider label encoding DRSM: Resample data with ROS	0.7550	0.5064	0.5689	75,282
3	111121	Third scenario	MDH: Include categorical missing data as other, numerical data as median DBIN: Do not consider binning OUTH: Remove outliers DSCL: Do not consider data scaling DENC: Consider one-hot encoding DRSM: Resample data with RUS	0.6894	0.4484	0.5243	75,282
106	212322	Best scenario (f1-score)	MDH: Include categorical missing data as other, numerical data as mean DBIN: Do not consider binning OUTH: Consider outliers DSCL: Consider min-max scaler DENC: Consider one-hot encoding DRSM: Resample data with ROS	0.7778	0.5413	0.6092	116,673

Table 5.6 Performances of first, best and last scenarios (continued)

Trial ID	Scenario ID	Scenario description	Scenario scope	Precision	Recall	F1-score	Support
140	222312	Best scenario (precision)	MDH: Include categorical missing data as other, numerical data as mean DBIN: Consider binning OUTH: Consider outliers DSCL: Consider min-max scaler DENC: Consider label encoding DRSM: Resample data with ROS	0.7854	0.5333	0.6032	116,673
186	313132	Best scenario (recall)	MDH: Remove rows including missing values DBIN: Do not consider binning OUTH: Change outliers with lower and upper limits DSCL: Do not consider data scaling DENC: Consider one-hot encoding for only nominal features DRSM: Resample data with ROS	0.7677	0.5447	0.6080	70,949
214	322322	Last third scenario	MDH: Remove rows including missing values DBIN: Consider binning OUTH: Consider outliers DSCL: Consider min-max scaler DENC: Consider one-hot encoding DRSM: Resample data with ROS	0.6955	0.4879	0.5614	70,949
215	322331	Last second scenario	MDH: Remove rows including missing values DBIN: Consider binning OUTH: Consider outliers DSCL: Consider min-max scaler DENC: Consider one-hot encoding for only nominal features DRSM: Resample data with RUS	0.7526	0.4969	0.5748	70,949
216	322332	Last scenario	MDH: Remove rows including missing values DBIN: Consider binning OUTH: Consider outliers DSCL: Consider min-max scaler DENC: Consider one-hot encoding for only nominal features DRSM: Resample data with ROS	0.7679	0.5368	0.6043	70,949

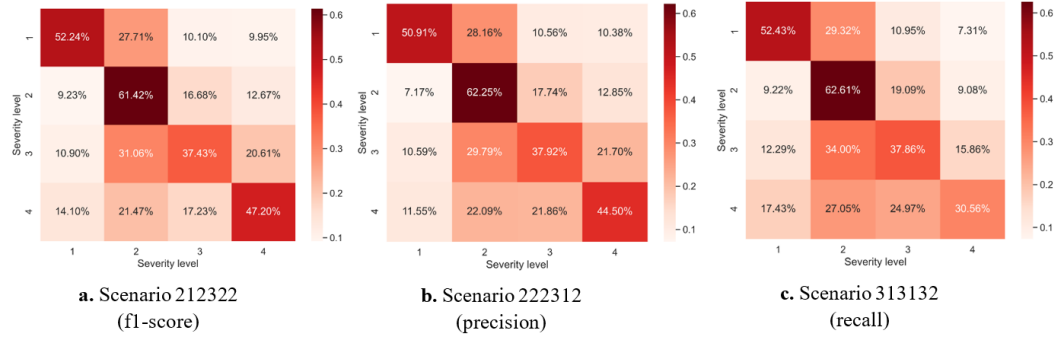


Figure 5.10 Confusion matrices of best scenarios

Figure 5.11 provides the prediction performance of 216 scenarios regarding f1-score (Figure 5.11a), precision (Figure 5.11b), and recall values (Figure 5.11c) to probe into the details of prediction performances. The results demonstrate that the scenario 212322 (trial ID: 106) yielded considerably higher f1-score compared to other scenarios, particularly if outliers were removed ($c = 1$) or the dataset was resampled with RUS algorithm ($f = 1$) (Figure 5.11a). Similarly, regarding recall values of 216 scenarios, not removing outliers ($c = 2, 3$) and resampling with ROS ($f = 2$) achieved higher accuracies (Figure 5.11c). On the other hand, there was not clear difference observed between precision values of the scenarios according to the adopted resampling strategy (Figure 5.10b). However, apart from removing outliers ($c = 1$), applying one-hot encoding for nominal and ordinal variables ($e = 2$) reduced the precision values considerably (Figure 5.10b).

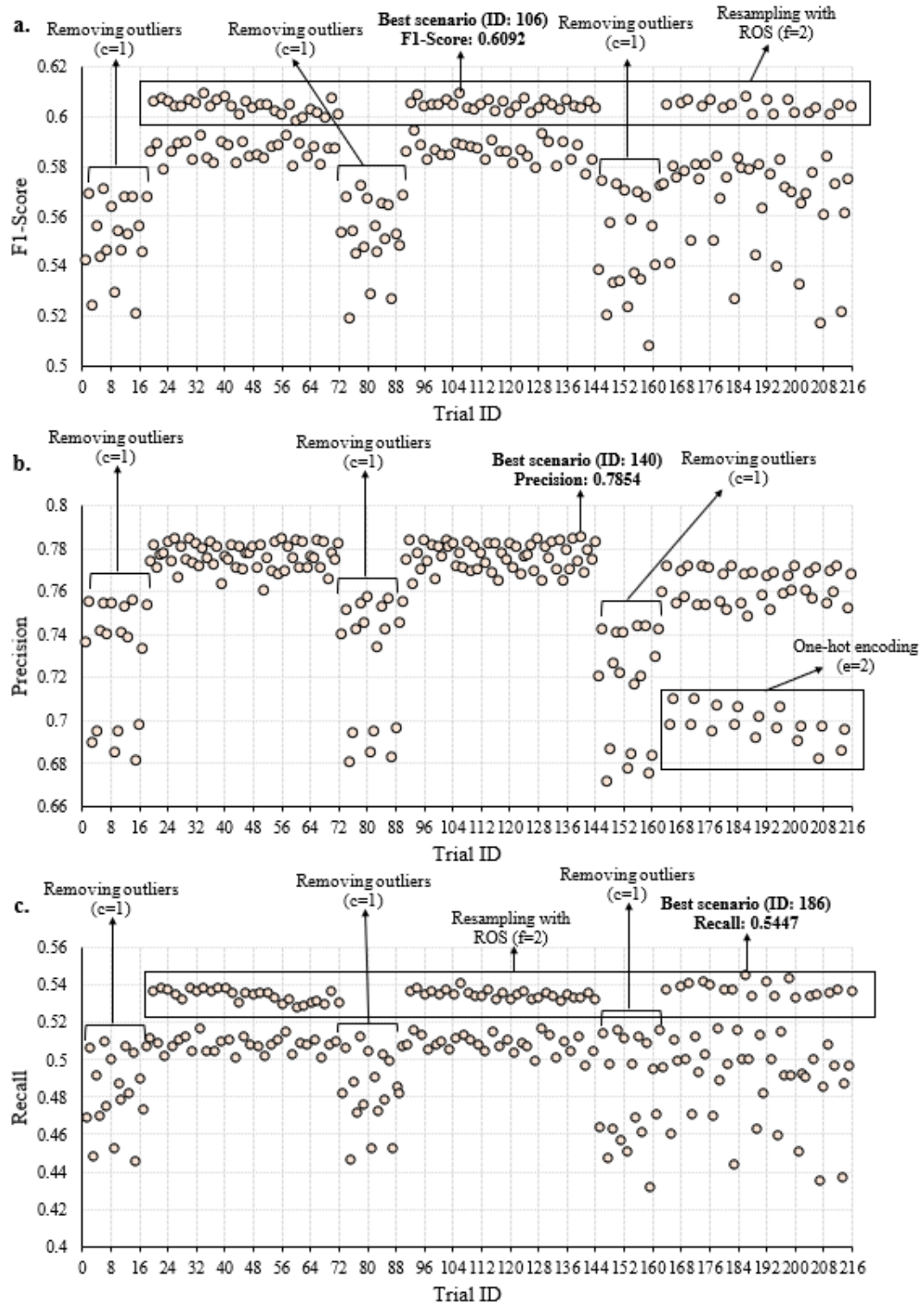


Figure 5.11 Performances of preprocessing scenarios in terms of (a) f1-score, (b) precision, (c) recall

Project managers and safety professionals may sometimes need foreseeing severe accidents, rather than predicting the severity level of accidents more accurately to gain an overall perspective in projects. To fulfill this need, the scenarios that led to the highest

recall values for each output label separately were searched as illustrated in Figure 5.12. These scenarios represent how many of the positive samples are classified as positive based on each severity level. The results show that the scenarios such as 313222, 311332, 313221, and 312211 achieved the highest individual recall values for classifying first aid, temporary incapacity, permanent incapacity, and fatality instances, respectively. Among them, temporary incapacity cases were predicted with higher recall value (64.86%) compared to other labels (311332; Figure 5.12b), similar to the best performances of the overall assessment schemes such as f1-score, precision, and recall values (Figure 5.10). The best scenario (312211) that predicted the most significant output label (fatality) classified 54.97% of the instances correctly (Figure 5.12d), while 54.78% of the first aid cases were classified correctly by the best scenario of the corresponding label (313222; Figure 5.12a). On the other hand, the best scenario (313221) that predicted permanent incapacity labels classified only 41.86% of the instances correctly (Figure 5.12c). The results show that high-severe accidents (permanent incapacity: **313221** and fatality: **312211**) can be predicted with higher recall values when removing rows with missing values (MDH:3), not applying binning (DBIN:1), scaling the data with standard scaler (DSCL:2), and resampling the data with RUS (DRSM:1) (bolded values). In addition, capping outliers to lower and upper limits (OUTH:3) and applying one-hot encoding for categorical features (DENC:2) were the best synthesis to predict permanently disabled cases correctly (not bolded values). On the contrary, it was not treating outliers (OUTH:2) and applying label encoding (DENC:1) for detecting fatality cases (not bolded values).

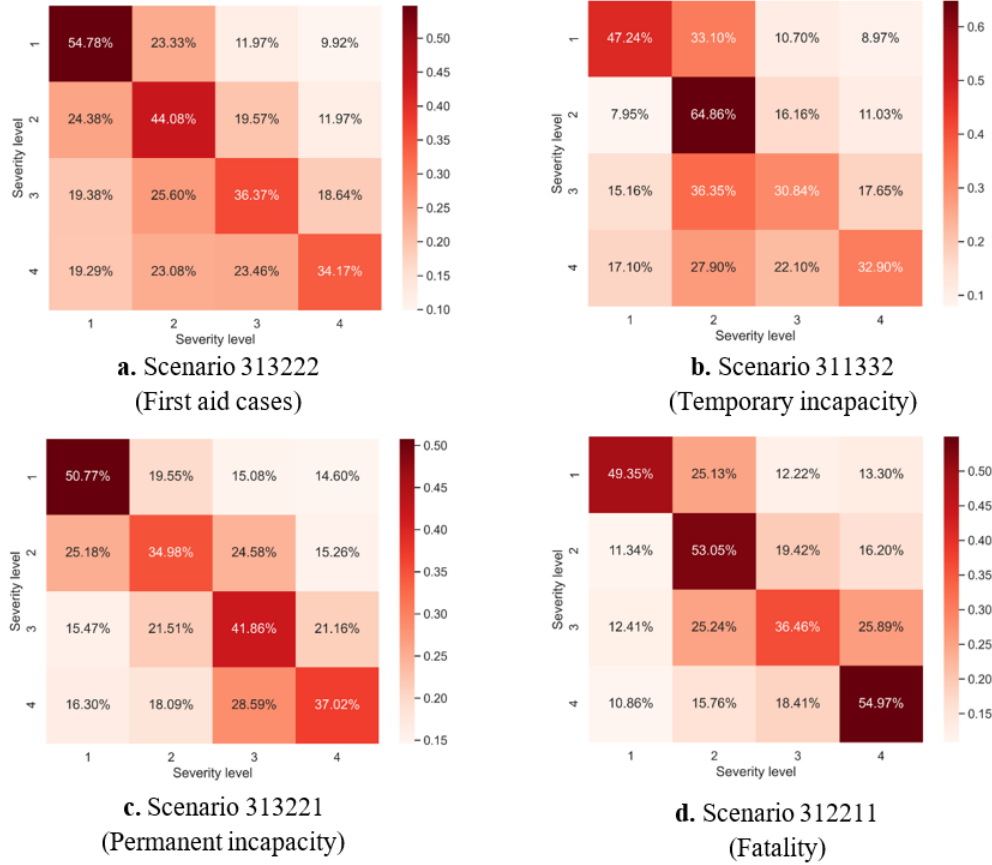


Figure 5.12 Confusion matrices of the best recall scenarios based on output labels

Box plot analysis regarding the recall values of 216 scenarios was performed based on severity levels to observe which classes were predicted more accurately by considering all the prediction models (Figure 5.13). The findings show that the proposed XGBoost models predicted cases in the temporary incapacity class better than other classes as the median line of the temporary incapacity was higher than the upper bounds of the other three classes. The range between the upper and lower bounds in fatality estimations was the greatest (0.2701-0.5467). In this context, the weakest two performances were observed through scenarios 311322 (0.2701) and 311122 (0.2781) to predict fatality cases. Since the best scenario for the fatality was 312211, particular attention needs to be given to OUTH, DSCL, and DRSM (bold values). Specifically, the ROS algorithm did not work well for detecting fatality instances, especially when the outliers were removed, and the dataset was not scaled with Standard scaler. Hence, practitioners and researchers are suggested to give more attention to the outlier handling, data scaling, and data resampling to predict fatality cases more accurately.

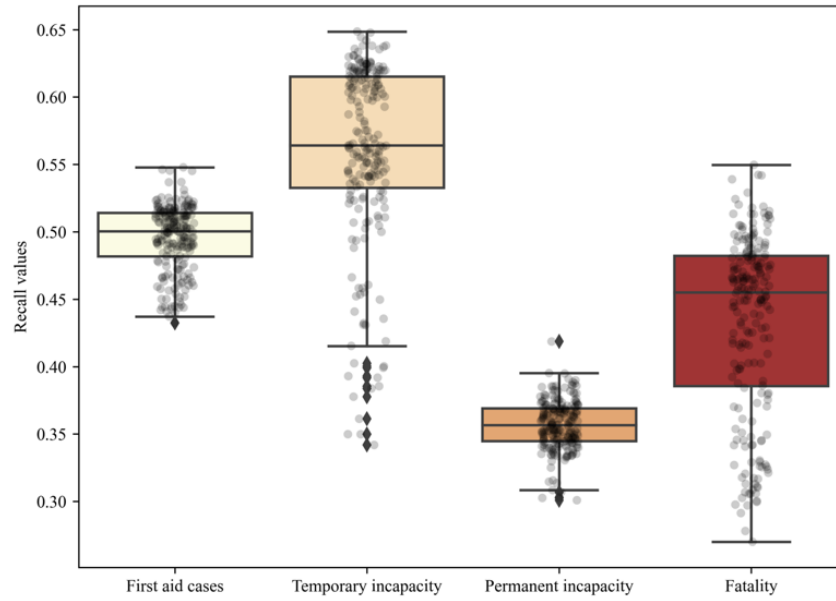


Figure 5.13 Boxplot of labels based on recall values

5.5.2 Results of Feature Importance

In order to explain features that are most influential to the severity prediction, identification of feature importance is required from the fitted models. The literature suggests the permutation importance technique, particularly in the case of high data dimensionality [217], [218]. Figure 5.14 shows the importance of the features computed based on permutation importance. To examine the importance of the features regarding the particular severity levels, four permutation importance values were calculated based on four scenarios, which were the most powerful scenarios for detecting each severity level separately (Figure 5.12). The results show that year of the accident was the most influential feature considering first aid cases, temporary incapacity, and fatality, while it was cause material to detect permanent incapacity instances. Overall, the year of the accident, cause material, age of the worker, and the number of past accidents in the company were identified as the top five features concerning any severity levels. In addition, the working days of the worker was the fourth most influential feature for predicting temporary incapacity instances, whereas the daily wage was found to be the fifth, fourth and fifth features for predicting first aid cases, permanent incapacity, and fatality, respectively.

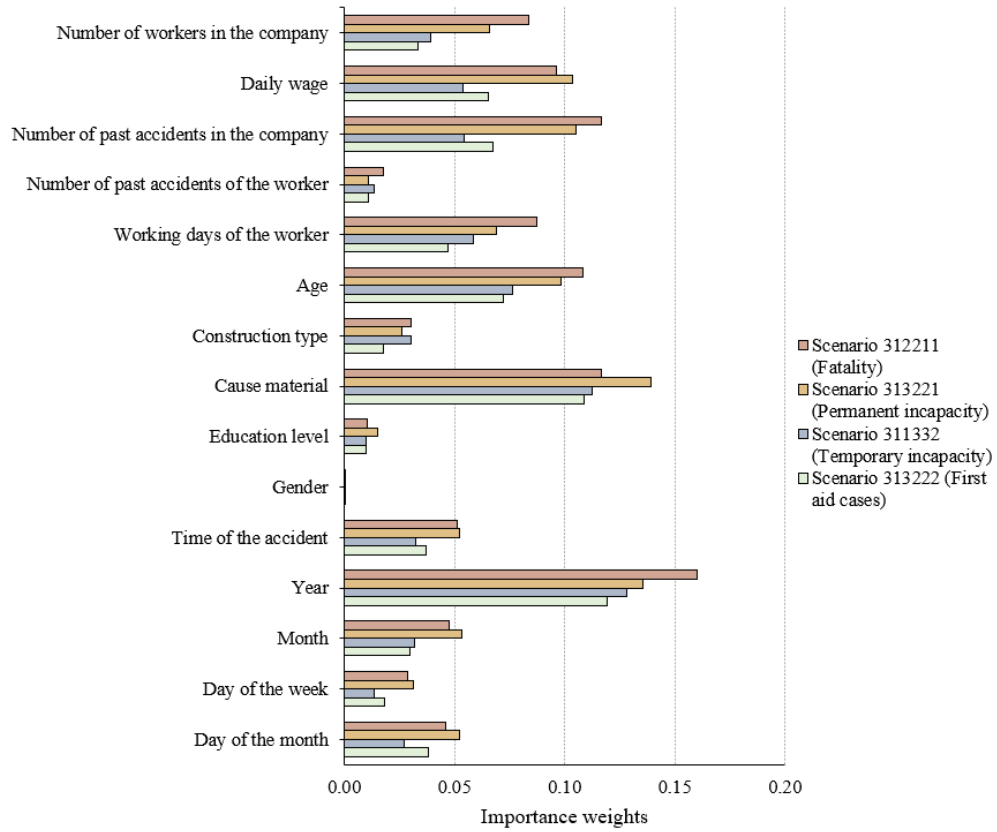


Figure 5.14 Permutation importance of features

5.6 Discussion

5.6.1 Assessment of the Adopted Methodology

The major contribution of this study to the theory is the validation of the automated preprocessing + XGBoost pipeline to predict accident severity using a national dataset. Existing literature on construction safety management that uses accident databases, proposes scenario-basis solutions, or develops a robust safety management plan mostly focuses on different approaches rather than predicting accident outcomes proactively [22], [202], [203], [219], [220]. Regarding ML applications, the effects of different combinations of preprocessing techniques on the prediction accuracy have not been investigated thoroughly yet. The findings reported in this study provided insight to data preparation before ML implementation by identifying the most effective preprocessing techniques for future projections. The results can be further used by construction companies to detect high severe accidents proactively, as part of continuous improvement, which might be enhanced through operational benchmarking as highlighted by Love et al. [203].

The scenario 212322 was found as the most powerful preprocessing combination to detect severity of construction accidents in this study. Several reasons support the proposed preprocessing scheme. One of them is highlighting “the value of each piece of information”. Careful investigation of the scenario (212322) showed that all the considered methods of the six preprocessing steps signified the captured information. Removing missing values [221], splitting the numerical values into discrete intervals [222], extracting outliers [223], not applying data scaling [197], adopting label encoding [83], and removing instances of majority classes by RUS algorithm [224] could cause some information loss. Even though there are missing values regarding some features, considering categorical missing values as “other” and numerical missing values as the mean of the corresponding feature could minimize the information loss with respect to other features. Many cases that include at least one missing value in the dataset could be related to the quality of the data. Still, other parts of these instances might be helpful for the model to predict the severity level of construction accidents. In addition, contrary to the view of categorizing some numerical features regarding DBIN [15], [206], this process could also worsen the quality of the data. Considering DSCL, not applying data scaling is not a common approach because of its manifesting effects on the quality of the captured information. Despite there is no shared view in the literature about whether the Standard scaler or Min-Max scaler provides higher data representation [195], Min-Max scaler provided higher prediction performance in terms of f1-score in this study. It is also recognized that label encoding could lead to information loss, particularly for nominal features, which cannot be represented orderly. In a similar vein, applying one-hot encoding to the ordinal features was found to be beneficial to improve prediction accuracies. Finally, ROS was found more appropriate method than RUS, probably due to the quality of the lost information and the severity of the class imbalance issue of the dataset.

According to the analysis results, the power of the prediction model was lower regarding permanent incapacity compared to other classes (Figure 5.13). It was observed that the prediction model mislabeled some of the permanent incapacity cases as temporary incapacity. In some scenarios, the rate of predicting temporary incapacity was even higher than the rate of predicting permanent incapacity, when permanent incapacity was the true class (such as Figure 5.12b). Despite the recall values of the permanent incapacity class were higher than temporary incapacity (regarding true permanent incapacity cases) in

most scenarios, unignorable misclassifications could easily be observed (such as scenarios depicted in Figure 5.10). The reason can be related to the fact that most permanent incapacity cases had been temporary incapacity before medical reports ensuring their permanent incapacity conditions [193]. Thus, it can be commented that the specific demographics or conditions attributed to the permanent incapacity can also be attributed to the temporary incapacity.

It is another significant finding that despite ROS was the most effective resampling strategy for overall performance (Table 5.6, Figure 5.11) and able to detect first aid (Figure 5.12a) and temporary incapacity cases (Figure 5.12b); RUS was better than ROS for detecting permanent incapacity (Figure 5.12c) and fatality cases (Figure 5.12d). The information lost during under-sampling could be associated with the majority classes (first aid and temporary incapacity), which did not affect predicting minority classes negatively [110]. Therefore, it is also sufficient to consider the RUS algorithm to predict minority classes more accurately, which also reduces the computational time. However, it should also be noted that the RUS algorithm may not always be the best choice if the dataset is not large enough, as this method is usually recommended for large datasets [95]. Besides, removing cases with missing values was the most effective MDH option for detecting single classes with higher recall values without considering the performances in other classes (Figure 5.12; $a = 3$ for each scenario). Hence, it can be deduced that information lost by removing cases with missing data did not affect classifying single classes, while it was ineffective regarding overall performance. In addition, not dividing the values in the numerical features into discrete bins is recommended, since most of the scenarios that performed best regarding several evaluation approaches did not consider data binning (Figure 5.10, Figure 5.12; $b = 1$ for each scenario except Figure 5.10b). Similarly, data scaling (either with Standard scaler or Min-Max scaler) is particularly needed to avoid information loss. Even though the combination of “considering outliers” and “applying Standard scaler” performed well in detecting fatality instances, this might also lead to erroneous scaling and information loss according to the characteristics of the data [197]. Since the change in the quality and characteristics of data can also change the best performing scenarios, the proposed automated preprocessing model, rather than the best performing scenario, can be beneficial to the interested scholars in the field.

5.6.2 Assessment of the Influential Factors

The results of permutation importance show both similarities and differences by comparing the accident outcome prediction studies in the literature. Similar to this study, Choi et al. [15] also found age and working days of the worker as among the most influential factors, and construction type and gender as the least important features according to the RF feature importance. As shown in Table 5.2 and highlighted in the literature [32], [225], most accidents leading to severe outcomes were experienced by older workers (such as 40-50 or older). In parallel, Leung et al. [226] noticed that construction workers who were over 40 perceived safety training as “an extra and unnecessary work-load”. It should be noted that the workforce in the construction industry is becoming increasingly older in many countries [227], which necessitates predictive models to save the lives of construction workers.

Contrary to this study, Choi et al. [15] identified the month and number of workers in the company as the two most significant attributes. These attributes have also affected the predictions of this study to a varying extent (Figure 5.14). However, other factors such as the cause material, year, number of past accidents in the company, and the daily wage outweighed these two attributes. There is an unignorable proportion of severe accidents that can be associated with non-compliance with safety regulations, lack of awareness, and imprudence [225]. By considering recent improvements in safety regulations and training programs in the construction industry, accidents that occurred recently (and probably in the future) are less likely to cause permanent incapacity and fatality compared to past accidents.

High impact of the “cause material” on the prediction can be explained by that some materials/equipment can particularly be associated with high-severe accidents. Similar to the analyzed dataset and the result of this study, Baker et al. [66] also found that high energy attributes (such as machinery and powered tool) were more influential to the high-severe accidents, while low-energy attributes (such as nail and small particles) were mainly associated with low-severe accidents. The dataset shows that construction machines, ladders, scaffolds, explosive or flammable materials, and power systems are the most dangerous material/equipment leading to severe accidents (Figure 5.3). Villanueva and Garcia [228] highlighted that most fatal accidents occurred with the involvement of transportation, elevation, scaffold, ladder, engines, and power generators. It was also addressed that the vast majority of struck-by-fatalities were accounted for

heavy equipment in the U.S. [229]. Hence, workers who use or are exposed to materials that can be associated with high energy should be carefully monitored in terms of safety measures to minimize severe accidents.

Daily wage was another influential factor on accident severity prediction. Similar to this study, Liao and Perng [230] found that the salary of the worker was among the most associated factors with fatal injuries. It can be stated that daily wages of the manual workers are lower compared to skilled and semi-skilled workers, and Villanueva and Garcia [228] showed that manual workers are more likely to experience fatal accidents. Therefore, safety training of workers with lower daily wages should be considered more diligently to raise their safety perceptions. Finally, as a unique attribute used in this study, the number of past accidents in the company was among the most determinant features. It was determined to be more significant feature for severe accident predictions compared to others (Figure 5.14). As also supported by the utilized dataset (Table 5.2), the higher the accident experience of the company, the less likely the workers to experience high-severe accidents in that company. The reason for this might be linked to the concept of lessons learned [231]. Therefore, it is of vital importance for newly established companies and companies with lower number of past accidents to invest more on the safety management practices in order to reduce the number of severe accidents.

5.6.3 Practical Use of the Research

To probe into the potential use of the study in construction sites, a proactive safety management framework was proposed (Figure 5.15). Typically, a worker provides the required information at the beginning of the work (such as age, gender, daily wage). This information can be captured automatically by the time a worker enters the construction site. Since *cause material* was determined as one of the most significant features, the worker can be asked to provide potential materials that the worker might use (or be exposed to) throughout the day (with tick marks). After controlling the information taken from workers, date, and database (regarding company-related features), instances that are aimed to be predicted can be developed. Considering that N_i is the number of *cause material* ticked by worker i , T_i is the work-hour time for worker i , and there are K number of workers entering the construction site, then there would be $N_1 \times T_1 + N_2 \times T_2 + \dots + N_{(K-1)} \times T_{(K-1)} + N_K \times T_K = N_{Trial}$ instances in the testing set to run the model daily. The training set, which is built based on historical data and actual cases, can be preprocessed with the proposed scenarios in this study. Accordingly, the workers who are predicted to

experience high-severe accidents in a specific condition (i.e., time or cause material) can be listed for site supervisors as a warning. This process can be performed every morning before the beginning of the morning shift to develop a dynamic and sustainable model. The outcomes of the instances regarding the previous day can also be stored in the database to contribute to the improvement of future predictions. By adopting this process, the worker and accident data can be collected dynamically, providing a proactive means to detect potential future accidents, some of which might be prevented.

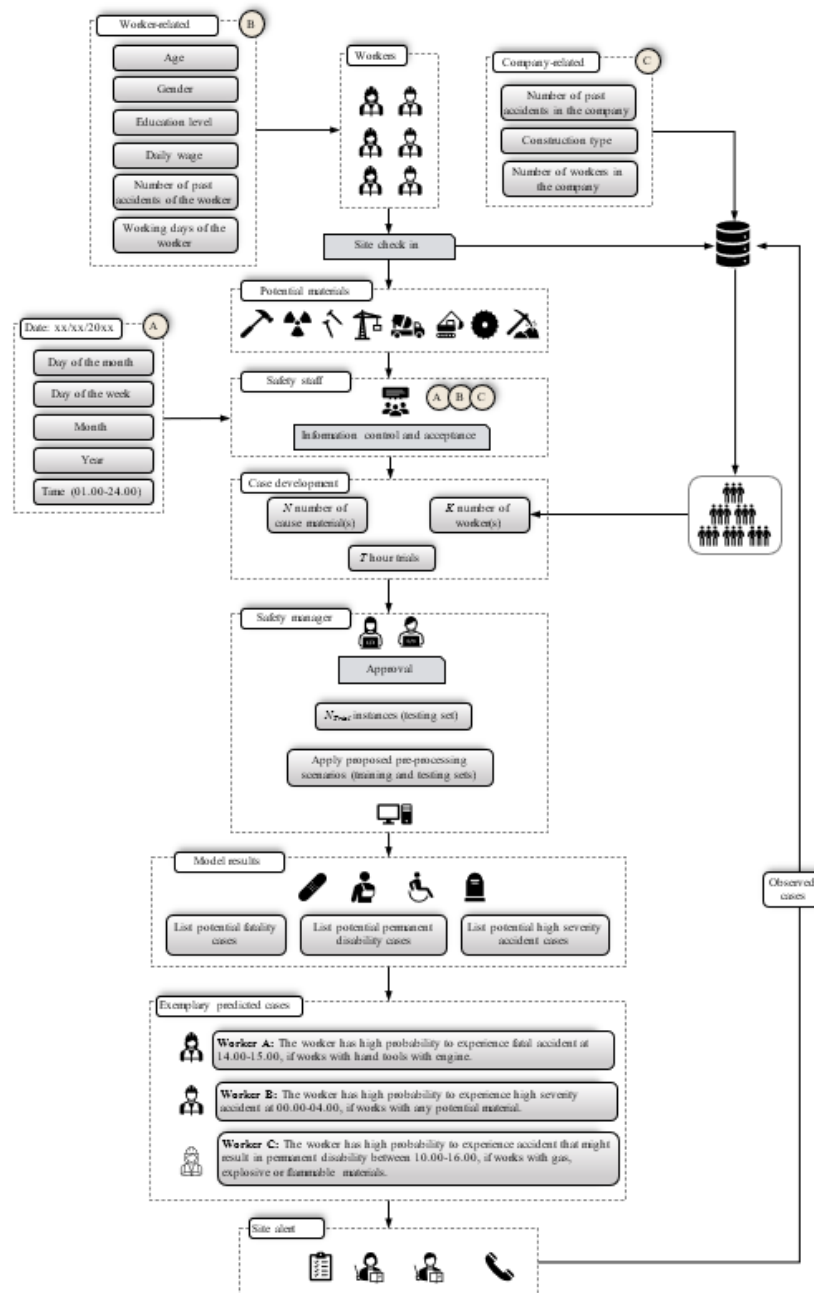


Figure 5.15 Example of a practical use of the research

5.7 Conclusion

This study produced a novel automated preprocessing model that compares the results of different combinations of several preprocessing steps to be used as a foundation for predicting accident severity levels in the construction industry. The applicability of the proposed model was validated by the XGBoost method, based on the data of 388,907 construction accident victims. A framework (Figure 5.15) was developed as a guide for the project and safety managers to adopt the proposed approach in construction sites and detect potential high-severe accident instances. Proposed framework can be used daily to specify the most risky conditions and workers based on time, material, company, and worker characteristics. The obtained results can contribute to the development of strategies in the prevention of severe accidents. Other major findings and corresponding contributions of this study can be summarized as follows:

- Analysis findings indicate that even a small piece of information can be critical in prediction models. Therefore, construction companies should record cases (as well as their descriptors) as much as possible to put them in use for predicting the potential accident outcomes.
- The findings showed that the highest prediction rate was 0.6092 in terms of f1-score attained by the scenario 212322. Therefore, (i) including categorical missing data as “other”, and numerical missing data as the mean of the feature, (ii) not considering data binning, (iii) not considering outlier handling, (iv) applying Min-Max-Scaler as a data scaling, (v) using one-hot encoding for all the categorical features, and (vi) resampling the dataset with ROS algorithm are suggested to the researchers for future predictions as these are determined as the most powerful preprocessing techniques.
- It was found that removing outliers and resampling the dataset with the RUS algorithm decreased the f1-score, precision and recall values of the prediction models. However, it is important to note that the permanent incapacity and fatality cases were classified with higher accuracies through the RUS algorithm. Therefore, the RUS algorithm can also be a proper resampling option for companies and/or datasets with higher fatality and permanent incapacity rates.
- Results of permutation importance addressed that the year of the accident, cause material, age of the worker, number of past accidents in the company, working days of the worker, and daily wage were the most influential attributes. These

findings may suggest the OHS managers to focus not only on the characteristics of the workers, but also on the job of the workers, the materials they usually use or exposed to, and accident history of companies when employing safety measures.

This study presents some limitations that can be considered in future studies. None of the specific preprocessing techniques was unanimously the best regarding all classes and evaluation metrics. Therefore, an ensemble approach, which includes the training results of several preprocessing techniques independently, might be beneficial to improve prediction accuracy. Other features can also be included in the prediction models to define the working environment of workers more precisely. For instance, unsafe conditions and the use of personal protection equipment are regarded as one of the main causes of construction accidents. Thus, although many of the existing studies (including this study) rely on the dataset captured from several institutions, it would add significant value to the theory and practice to incorporate new and descriptor features. Furthermore, this study focused on the occurred construction accidents and disregarded some incidents such as non-accident cases or near misses. However, inclusion of some instances that do not lead to construction accidents would contribute a lot to the theory and practice, despite the collection of such data requires further efforts given by construction companies. From the methodological point of view, this study evaluated the XGBoost algorithm as a tree-based ML method. Exploring the applicability of deep learning algorithms or other ML methods such as RF, ANN, and SVM in the proposed automated preprocessing model might help to show the method that works well with which preprocessing scenario. Finally, even though six preprocessing steps were assessed with some particular methods in each of them, there are some other preprocessing or feature engineering methods that can improve the proposed model. Therefore, this study can be extended by implementing other preprocessing steps such as feature selection and dimension reduction methods, not only to increase the model accuracies, but also minimize computational burden. Parameter optimization methods, such as particle swarm algorithm or genetic algorithm, can also be implemented in the model to identify the effects of different parameters of the ML methods on the prediction accuracy regarding developed preprocessing scenarios.

The extension and improvement of the proposed methodology can contribute a lot to the theory and practice, which could provide confidence to the researchers and practitioners to build a proper foundation for prediction models. The construction safety and research

communities in other fields would also benefit from this research to increase the accuracies of the prediction models in different domains. Living in the era of data, the importance of each piece of information cannot be ignored, regardless of its quality. Therefore, the study is also expected to aid governmental authorities and construction companies in recognizing the significance of safety information more extensively. In this way, lessons can be learned effectively instead of being forgotten, as a remedy to occupational accidents.



PREDICTION OF CONSTRUCTION ACCIDENT OUTCOMES BASED ON IMBALANCED DATASET THROUGH INTEGRATED RESAMPLING TECHNIQUES AND MACHINE LEARNING METHODS

This chapter was published in Engineering, Construction and Architectural Management as a research article.

6.1 Introduction

The construction industry is one of the most unsafe industries with a massive number of injuries and fatalities globally [232]. Huge damage to the well-being of workers, profitability and reputation of companies are some of the primary adverse consequences of construction accidents, which have been an everlasting issue under the consideration of policy makers over decades [233], [234]. The construction industry accounts for high percentages of all workplace fatalities across the world such as 22% in Ontario, Canada [235], 30% in Korea [2], 32% in Japan [236], and 40% in the United Kingdom [237]. These rates can become more serious for some countries with respect to workforce rate in the industry. For instance, the rate of employment in the construction industry accounts for 5% in the United States, however, 20% of the fatal occupational accidents occurred between 2003 and 2013, highlighting that the industry requires particular attention [238]. Similarly, accidents accounted for nearly 32% in 2019 [21] in Turkey, while the rate of employment was 5.6% in the construction industry [128]. With these facts, a growing interest in safety management practices has been given by scholars, employers, and policy makers for better construction safety management [235], [239].

Unfavorable image of the industry regarding the safety performance brings attention to the early diagnosis of construction accident outcomes in recent studies. Niu et al. [238] point out that the predictive approach is a promising research direction to improve construction safety performance within the context of smarter construction facilities. In this respect, past construction accident records can present a valuable foundation for the development of ML models, which can support the prevention of future accidents [4].

Data collection for providing a basis to predict occupational accidents is important to attain accurate results (Ayhan & Tokdemir, 2019a). Therefore, several researchers proposed different prediction/classification models based on the historical data collected from various institutions or companies [2], [8], [14], [240], [241]. However, a dataset consisting of fatal and non-fatal construction accidents can be severely imbalanced [2], [242]. Imbalance class occurs when the number of instances of one category is much higher than other categories, such that predictions under these conditions become biased toward the majority class [106]. In a construction accident dataset, when the number of non-fatal accidents is significantly more prevalent than the fatal accidents, the prediction rate may yield higher accuracies, compromising the accurate prediction of single minority class. Since most of the ML algorithms are based on the assumption of a balanced dataset, performing the pre-processing techniques to achieve balance in class distribution is essential to acquire accurate predictions of construction accident outcomes [2].

Safety management literature addresses the scarcity of research on the application of ML algorithms for predicting construction accident outcomes. Pertinent literature shows that the studies on construction safety are limited, particularly in developing countries (Zhou et al., 2015). Therefore, Turkey, as a developing country, can be a valuable case for the analysis of construction accident outcomes with a huge number of available accident data. A dataset consisting of 395,777 construction accidents (2011-October 2020) acquired from the Republic of Turkey, Social Security Institution (SSI) showed that nearly 31% of all nation-wide construction accidents took place in Istanbul, which is the most populated industrial, economical, and cultural center of Turkey (Ekmekcioğlu et al., 2020). Huge construction projects including several industrial and residential developments took place in Istanbul over the last two decades [244], [245], which brought a significant amount of employment opportunities to thousands of workers. As a result, 121,002 construction accidents were recorded in the city (120,277 non-fatal and 725 fatal) between 2011 and 2020.

A thorough literature survey reveals that there is a gap in accident predictions using an imbalanced dataset through integrated resampling techniques and ML methods in the construction safety domain. Construction accident databases [2] or performance comparison of resampling techniques in other domains [246], [247] have been studied separately in the literature. However, the impact of resampling methods on the performance of prediction models developed with construction accident datasets has not

been explored adequately. Therefore, the aim of this study is to develop a robust prediction model for occupational accident outcome classification in the construction industry suitable for an imbalanced dataset. The constructive and unique approach of this study is related to dealing with the imbalanced dataset (RAW) of construction accidents by comparing the performances of artificial neural networks (ANN), naïve bayes (NB), K-nearest neighbor (KNN), and random forest (RF) algorithms, which are integrated with three different preprocessing techniques; random under-sampling (RUS), random over-sampling (ROS), and synthetic minority over-sampling technique (SMOTE).

The study also highlights a city-basis investigation, which analyzes 121,002 construction accidents occurred in Istanbul. Past predictive studies on construction safety mainly rely on accident-basis [9], [113], project type-basis [240], company-basis [8], or country-basis [2], [11] implementations. However, metropolitan cities such as Istanbul have their own sector dynamics with the number and variety of projects. More than 20% of new buildings are constructed in Istanbul in Turkey every year [248] and 22.9% of the total number of safety professionals registered across the country work in the city based on the records of the Ministry of Labor and Social Security [249]. In addition, the majority of the mega projects (i.e., Istanbul Airport and Eurasia Tunnel projects) were completed in Istanbul intensifying construction works [250]. Such mega projects and major investments contribute significantly to the know-how transfer of effective safety management practices to the construction sector. Overall, this study is expected to contribute to proactive construction safety management even with the existence of an imbalanced construction accident outcome dataset. The proposed utilization plan can be used periodically (i.e., monthly) in construction sites to detect workers who are susceptible to fatal accidents. Hence, this study can provide a useful decision-making input for safety professionals to develop effective occupational health and safety (OHS) strategies.

6.2 Background

6.2.1 Accidents and Safety Issues in Construction Sites

Stemming from the dynamic and complex environment, dangerous working conditions, and several stress sources mostly related to time and cost constraints [251], [252], safety of workers is one of the major challenges in construction projects globally [2]. Hazardous situations due to insufficient measures and lack of knowledge/awareness of workers are important reasons for occupational accidents on construction sites [232]. Goh et al. [253]

reported significant impacts of intention and social norms on the unsafe behaviors of construction workers. Mohandes et al. [254] found that inappropriate safety guidelines, insufficient safety management system, lack of safety culture, poor safety knowledge/awareness, financial instability of firms, and corruptions as the underlying causal factors in the occurrence of accidents. Many researchers also addressed the role of time-related features such as time of the accident, month, day of the week etc. in accident occurrence or severity [2], [8], [199], [240]. Therefore, several safety issues in construction sites that can be associated with company-related, worker-related, project-related, and time-related factors not only cause occupational accidents but also elevate their severity levels. Particularly accident types such as falling from height, collision by objects, rollover, and falling objects occur frequently [11], causing severe results such as loss of lives and adversity in projects.

National occupational accident databases have been frequently used recently in the literature, mostly to highlight leading indicators of occupational accidents. For instance, Cheng et al. [255] analyzed the occupational accidents at small construction firms in Taiwan and reported that particular conditions or situations require significant attention to minimize occupational accidents such as (i) workers' first days, (ii) health and safety management system of companies, (iii) provision of necessary personal protection equipments (PPEs), (iv) adequate use of PPEs, and (v) perception of workers about hazard warnings. Shao et al. [27] examined accidents in building construction projects in China and addressed that more fatal accidents are observed in summer, on Monday, and during 10 am-11 am and 3 pm-4 pm. They also stated that the majority of fatal accidents are caused by falling from height and collapse/hoisting damages. Winge and Albrechtsen [256] found that falling, floor/platform, falling object, and contact with moving parts of a machine were the most frequently faced accident types in Norway.

Improvements have been made in Turkey over the last decades with the establishment of new standards and rules. However, occupational accidents rates are still considerably high, mostly due to the lack of enforcement, which led the industry facing many setbacks to perform adequate mitigation measures [199], [232]. According to the International Labor Organization [20], Turkey is the 9th most unsafe country in terms of fatality rates. A considerable portion of these accidents occurred in small and medium-sized residential projects [22] and Olcay et al. [25] reported that falling from height accounted for the vast majority of the fatal accidents in the country. Baradan et al. [257] underlined the

significant (i.e., negatively) relationship between the human development index and safety consciousness in the construction sector across 23 countries. They also found that Turkey ranked at the bottom in terms of the human development index with the highest fatality rate. Therefore, accident outcome prediction is of vital importance for the country. Detection of cases/workers, which have a high probability to experience fatal accidents can be effective in the employment of the very much needed mitigation measures.

6.2.2 Machine Learning in Construction Safety Literature

Recent improvements in ML algorithms encourage researchers to explore the applicability of artificial intelligence techniques in the construction safety management domain. Safety management processes can be improved through the investigation of construction accident outcomes using ML methods. Predicting accident outcomes could make major contributions in reducing the number of severe accidents. Therefore, researchers recently have been focusing on ML applications to predict accident-related outcomes. For instance, Yang et al. [9] developed an ML model for near-miss fall detection among ironworkers by using one-class support vector machine (SVM). Based on a worker's kinematic data, near miss fall experiments were performed in the laboratory and outdoor, which predicted the outcomes (i.e., near miss or normal walk) with 86.8 and 85.2% accuracies, respectively. Ray and Teizer [258] generated visibility maps to measure blind spots using the RF algorithm dynamically based on construction equipment operators' perspectives. By using range images and point cloud data for measuring, they showed the applicability of the proposed methodology as an intelligent alert system for construction equipment operators.

Tixier et al. [259] investigated unstructured injury reports and developed a natural language processing system to identify precursors and predict outcomes, which included energy source (chemical, motion, gravity, electricity, etc.), injury code (struck by an object, exposure to a harmful substance, etc.), and body part (head, neck, lower extremities, etc.), as well as injury severity (first aid, medical case, fatality, etc.). Alexander et al. [241] analyzed the precursors of fatalities in construction projects by principal component analysis and developed a regression model to predict construction accident outcomes such as successfully managed incidents, near-miss incidents, and fatal/disabling injuries based on the data consisting of 35 cases. Goh and Ubeynarayana [4] used SVM, logistic regression (LR), RF, KNN, decision tree (DT), and NB to classify the construction accident narratives into eleven labels indicating the reasons for accidents

such as the collapse of an object, traffic, falls, fires, exposure to extreme temperatures, etc. based on 1,000 construction accident narratives obtained from the U.S. Occupational Safety and Health Administration (OSHA). Ghodrati et al. [260] analyzed safety outcomes in the construction industry by means of the number of claims and entitlement claims. By implementing multiple regression analysis (MRA), they found more accurate results with models based on liable earning than the models generated using other attributes such as the number of employees.

More recently, Kang and Ryu [11] used the RF algorithm for predicting the construction accident types in Korea, such as fall from height, collision by an object, rollover, and falling object. The analysis results showed that the RF model achieved an accuracy of 71.3%, while assailing materials feature was determined to be the most important feature. Ayhan and Tokdemir [63] performed Delphi panels to collect construction accident data. Then, they performed multiple regression and ANN to predict construction incident outcomes by means of output elements such as at-risk behavior, near miss, partial failure, first aid, medical intervention, lost workday cases and fatalities. Choi et al. [2] proposed an ML predictive model based on national data to estimate construction accident outcomes (fatal or non-fatal) in Korea. They found that the RF method showed the best performance among several ML algorithms such as LR, DT and AdaBoost. Lee et al. [8] developed accident prediction models by using SVM and DT-based ensemble methods and analyzed the correlation between several variables affecting the construction accident outputs. The most influential variables were derived through several techniques such as clustering, chi-square, and Cramer's V to classify the output elements such as injury, death, and property damage. Zhang et al. [10] examined a construction accident report dataset consisting of 1,863 instances to classify the construction accidents obtained from the OSHA. They employed several commonly used ML algorithms such as convolutional bidirectional long short-term memory (C-BiLSTM), SVM, NB, and LR.

Despite a wide range of ML studies in the construction safety literature, their integrated applications with resampling strategies have not been investigated thoroughly. Coupling resampling strategies with ML methods for fatality prediction could contribute significantly to the improvement of safety conditions for construction workers by detecting the most undesired outcomes of accidents. The proposed approach can also guide practitioners and researchers about the type of the ML methods integrated with different resampling strategies, which can provide more robust predictions.

6.2.3 Class Imbalance in Construction Safety Literature

In classification problems, if the number of instances in one category (such as non-fatal) is much higher compared to other categories (such as fatal), then an important challenge called imbalanced class distribution emerges [189]. Under this condition, the prediction can be biased in favor of the majority class, which results in unreasonable but high prediction accuracy. In this vein, predicting each instance as a majority class might result in high accuracies. However, predicting the minority class is usually preferred in real-life problems (for example, fatality prediction in the current case). In most cases, imbalanced data is prevalent in construction accident datasets. Therefore, a balance in the class distribution is essential as most of the ML methods are based on the assumption of balanced class distribution (Chen et al., 2021). In this context, different resampling strategies could provide varying prediction accuracies with their divergent advantages.

A few researchers adopted some of the sampling strategies in construction safety studies, which drew attention to the class imbalance issue in construction accident outcome predictions. Ajayi et al. [240] adopted reweighting approach and added different penalty coefficients to training errors for the prediction of health and safety risk in power infrastructure projects. They assigned larger penalty coefficients for minority classes and smaller penalty coefficients for majority classes to deal with imbalanced data. Similarly, Baker et al. [242] used different weights inversely proportional to the frequencies of classes in the training set only to address the class imbalance issue before performing text mining to predict injury precursors from construction accident reports. Goh and Ubeynarayana [4] generated an SVM model by adopting One-vs-Rest (OVR) linear classifier to deal with the imbalanced frequency of different labels for construction accident classification.

Zhang et al. [10] aimed to make all categories more balanced by manually employing over-sampling technique to perform a fair construction accident classification. Kang and Ryu [11] adopted the RUS strategy to deal with the class imbalance problem for predicting construction accident types. They reduced the size of the dataset consisting of construction accidents from 9,795 to 6,374 by removing classes of less than 10%. Choi et al. [2] addressed that RUS, ROS, and SMOTE are the most frequently adopted resampling techniques. Since RUS might deteriorate the negative class distribution and SMOTE could generate artificial positive objects through interpolation that would be more suitable for numerical features, they used ROS algorithm to deal with imbalanced

class distribution. Koc et al. [193] used RUS algorithm to predict the permanent disability of construction workers prior to applying tree-based ML methods. The comparison of past studies addressing the class imbalance issue is provided in Table 6.1. Accordingly, this study fills the gap in the construction safety literature regarding comparative analysis of resampling techniques integrated with different ML methods with respect to scenario basis evaluation. This will help researchers and practitioners select the most effective hybrid resampling-ML scenario in analyzing the occupational accidents. In addition, the comparison of resampling strategies with the non-resampled model has not been performed previously in the construction safety management literature. It is also important to note that no significant performance differences between these resampling strategies would be expected if the class imbalance had been moderate. However, in case of a large imbalance in the class distribution (which is 1 over 166 in this study), the performance could be affected significantly [2].

Other than safety literature in the construction industry, some researchers have compared the performance of different resampling strategies in other sectors. Chi et al. [109] used SMOTE, cluster-based over-sampling (COS), cluster-based under-sampling (CUS), and cluster-based SMOTE (CBS), benchmarking with the raw data (without resampling) to evaluate the seismic capability of school buildings. They found that the RF algorithm achieved the best performance with the COS technique. Sarkar et al. [189] adopted SMOTE, majority weighted minority oversampling technique (MW-MOTE), k-means SMOTE (KM-SMOTE), and border-line SMOTE (BL-SMOTE) as minority oversampling-based imbalanced data handling techniques, to balance the dataset of accidents that occurred in the steel industry and found that KMSMOTE had better performance. Peng et al. [92] compared the performances of RUS, case control under-sampling (CCUS), ROS, and SMOTE in predicting real-time traffic crash risk, and found that coupling RUS algorithm integrated with Adaptive Boosting (AdaBoost) showed the best performance. Ghorbani and Ghousi [247] compared several resampling methods such as borderline SMOTE, SMOTE, SMOTE-Edited Nearest Neighbors (SMOTE-ENN), SVM-SMOTE, and SMOTE-Tomek, all integrated with ML methods such as KNN, ANN, XGBoost, SVM, DT, LR and NB for performance prediction of students. They found that the RF method yielded the highest accuracy if integrated with SVM-SMOTE resampling algorithm. Similarly, Iqbal et al. [246] examined the usability of several resampling techniques (RUS, ROS, and SMOTE) in software defect

classification. With experiments performed on 12 datasets, the researchers found ROS showing higher prediction performance.

Table 6.1 Studies addressing class imbalance issue in construction safety literature

Reference	Country of data source	Size of the dataset	Rate of majority class over minority class	Re-sampling approach	Employed ML methods	Aim of the study
Goh and Ubeynarayana [4]	United States	1,000	14	One-vs-Rest (OVR) linear classifier	SVM, LR, RF, KNN, DT, NB	Examining the effects of text mining techniques in accident report classification
Kang and Ryu [11]	Korea	9,795	3	RUS	RF	Developing accident type prediction model to prevent occupational injuries
Ajayi et al. [240]	United Kingdom	17,972	N/A	Re-weighting approach, added different penalty coefficients to training errors	DNN, GBM	Developing a deep neural network model for help and safety risk prediction
Choi et al. [2]	Korea	140,169	48	ROS	RF, AdaBoost, DT, LR	Developing a national prediction model to present the risk of fatality
*Baker et al. [242]	Different countries	69,596; 10,299; 14,616; 17447	3.4; 2.5; 2.5; 5.7	Weights inversely proportional to the frequencies of classes	CNN, HAN, TF-IDF, SVM	Comparing performances of several approaches to automatically learn injury precursors from test
Zhang et al. [10]	United States	1,000	N/A	Manually employing over-sampling technique	C-BiLSTM, SVM, NB, LR, CNN, LSTM, BiLSTM, C-LSTM	Classifying accident reports to understand and prevent occurrence of occupational accidents
Koc et al. [193]	Turkey	47,938	44	RUS	RF, XGBoost, AdaBoost, Extra trees	Integrating feature engineering, genetic algorithm and tree-based methods for permanent disability prediction
This study	Turkey	121,002	166	RAW, RUS, ROS, SMOTE comparative analysis	RF, NB, ANN, KNN	Examining the effects of resampling techniques in accident outcome prediction (fatal and non-fatal)

Sarkar and Maiti [106] conducted a literature review and science mapping on ML applications for occupational accidents and addressed that an in-depth investigation of imbalanced data handling is essentially required. By developing construction site accident classification models, several researchers suggest that data balancing techniques need to be considered as a valuable research direction to improve the predictive power of generated models in the construction safety literature [97], [170].

Although imbalanced data pre-processing techniques have been investigated in other industries and for different domains, the comparative analysis of resampling strategies has not been examined in the construction safety literature, disregarding the divergent characteristics of the occupational accident datasets in the construction industry. However, performances of resampling strategies are sensitive to attribute types, the number of instances, degree of imbalance, and the characteristics of the utilized dataset [2], [92], [110], [262]. Construction accident datasets might include several categorical, numerical, and binary attributes affecting the outcomes of occupational accidents, usually under a severe class imbalance condition. The dynamics of features defining an occupational accident could be different than features designating other issues such as traffic accident crash risk [92] or seismic capability of a school building [109]. In addition to pre-accident attributes, several post-accident conditions such as first aid provided, body part injured, or environmental considerations have impacts on accident outcomes. Even in the same domain, the essence and characteristics of the occupational accident datasets can vary in different countries with respect to the attributes included. Hence, the binary classification of occupational accident outcomes is a complex and unique problem.

The abovementioned facts highlight a need for careful examination of occupational accidents regarding prediction models, not only to analyze the performances of ML methods, but also to investigate the robustness of several resampling strategies. In this study, fatality is the minority class compared to injury class, which makes the examination of resampling strategies even more significant, since the primary objective of the classification problem addressed in this research is to detect fatal instances. This would enable opportunities to forestall adequate mitigation measures.

6.3 Research Methodology

The flowchart of the developed prediction model is illustrated in Figure 6.1 and analysis steps are explained in the following sub-sections. All tests were run in *Python 3.8* and the main packages used for model design were *scikit-learn v0.24.1* (for preprocessing, one-hot encoding, ML applications, feature importance), *imbalanced-learn v0.7.0* (for data resampling), and *matplotlib v3.3.2* (for visualization) libraries.

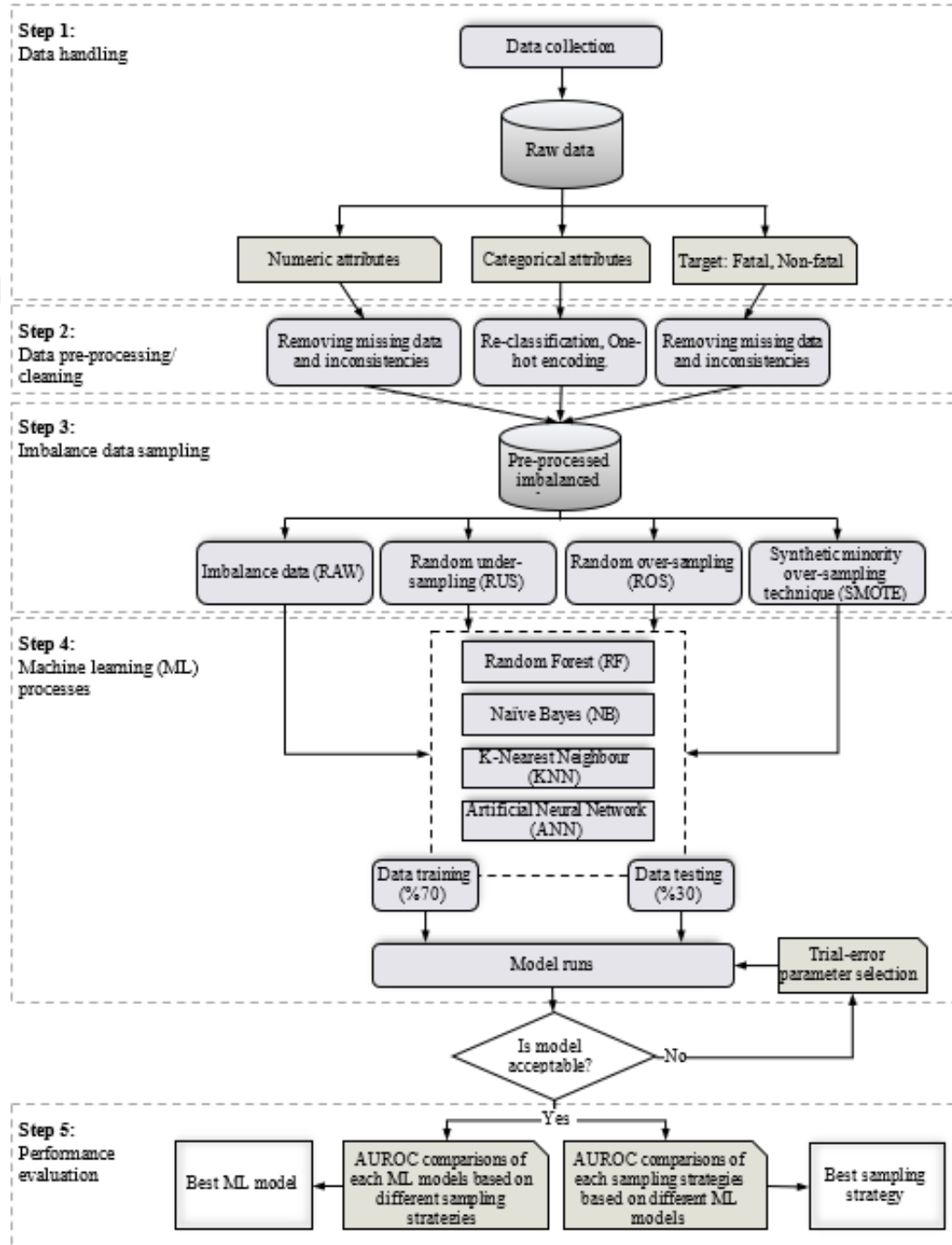


Figure 6.1 Flowchart of the study

6.3.1 Description of the Dataset

The dataset related to construction accidents occurred in Istanbul was obtained from the SSI. The dataset includes the number of fatal and non-fatal accidents recorded between 2011 and 2020 (up to October). Table 6.2 highlights the brief summary of 121,002 occupational accidents, of which only 0.6% resulted in a fatality. It can be observed that the number of recorded accidents continued increasing until 2018, then started decreasing. The records also illustrate that the employment rates in the construction industry increased from 2012 to 2018 and decreased afterward until 2020 (TSI, 2020), mostly due to the recession in the local construction sector.

Table 6.2 Imbalanced dataset of construction accidents

Year	2011 - 2012	2013	2014	2015	2016	2017	2018	2019	2020	Total
Number of accidents	3,406	5,718	7,496	10,495	16,366	23,247	28,788	15,859	9,627	121,002
Number of injuries	3,361	5,655	7,402	10,399	16,237	23,131	28,683	15,812	9,597	120,277
Number of deaths	45	63	94	96	129	116	105	47	30	725
% of death	1.32	1.10	1.25	0.91	0.79	0.50	0.36	0.30	0.31	0.60

Based on the dataset, 15 input variables were selected to estimate the target outcome (fatal vs non-fatal). Details of the attributes are shown in Table 3. First, a literature review was performed to identify the influential attributes of fatal and non-fatal accidents. The initial list of features was developed based on the studies of Choi et al. [2], Mistikoglu et al. (2015), Ajayi et al. [240], Tixier et al. [7], Ayhan and Tokdemir [63], Baker et al. [242], Alexander et al. [241], and Wu et al. [263]. Then, the results of the literature review and the attributes in the dataset were cross-checked. However, some features such as the use of PPEs, hazardous conditions, and unawareness of the workers could not be considered in this study due to the unavailability of the corresponding attributes. In addition, there were other attributes in the original data, such as injured body part, type of injury and reason for the incident, which were not used during analysis as they are post-accident features. Since the major objective of the current research is to provide insights for the development of proactive strategies, only pre-accident attributes were considered.

Table 6.3 Description of attributes

Context	Attributes	Type	Number of unique instances	Instance ranges
Time-related	Day of the month	Categorical	31	1, 2, 3, ..., 31
	Day of the week	Categorical	7	Monday, Tuesday, ..., Sunday
	Month	Categorical	12	January, February, ..., December
	Year	Categorical	10	2011-2020
	Time of the accident	Categorical	24	00.00–01.00, 01.00–02.00, ..., 23.00–00.00
Worker-related	Gender	Binary	2	Female, male
	Age	Numeric	56	18 – 73
	Working days of the worker	Numeric	2,387	0 – 15,684
	Education level	Categorical	6	Illiterate or only literate; elementary school; secondary school; high school; bachelor (two-year); BSc/MSc/PhD
	Daily wage (<i>Turkish Lira</i>)	Numeric	31,681	16.31 – 1,672.69
	Number of past accidents of the worker	Numeric	10	0 – 17
	Material used	Categorical	157 (16)*	Construction machinery; hand tools with engine; stairway, ladder, scaffold; office materials; assembly kits etc.
Company-related	Construction type	Categorical	322 (10)*	Road construction; residential building construction; railway and metro construction; infrastructure works etc.
	Number of past accidents in the company	Numeric	151	0 – 1,804
	Number of workers in the company	Numeric	1,434	1 – 9,980
Output variable	Accident outcome	Binary	2	Fatal, non-fatal

All the attributes were recorded as binary, categorical, and numeric values (Table 6.3). The number of instances varied according to attribute characteristics. For instance, the day of the month, day of the week and time of the accident were inherently grouped as 31, 7 and 24 instances, respectively. In addition, material used, and construction type consisted of a diverse range of instances with just a few cases in some elements. Thus,

elements comprising less than %2 of the instances in these two attributes were diligently examined according to the nature of the material and construction types. Accordingly, the elements of these two attributes were regrouped into 16 and 10 categories, respectively. It can be noted that all categorical data types were grouped as nominal and ordinal (Kang and Ryu, 2019). All the nominal attributes were subjected to one-hot encoding, since allowing the model to make assumption on a natural ordering among categories results in erroneous classifications (Pradeep & Kallimani, 2020; Wang & Chen, 2020).

6.3.2 Adopted Data Resampling Strategies

The dataset used to predict the accident outcome was highly imbalanced (120,277 non-fatal, and 725 fatal). Therefore, three class imbalance handling algorithms, (RUS, ROS and SMOTE) were considered as resampling strategies. However, it needs to be noted that, resampling was not subjected to imbalanced data to observe the prediction rate without any preprocessing related to imbalance class distribution in the first step. In other words, imbalanced dataset (RAW) was utilized for benchmarking. RUS, ROS and SMOTE were selected since they are the most frequently applied resampling strategies [2]. SMOTE works well with numerical attributes, while RUS and ROS might be preferred for different type of attributes [2]. Despite there are other under sampling-based (i.e., CUS and CCUS compared to RUS), over sampling-based (i.e., COS compared to ROS), and minority over sampling-based (i.e., MWMOTE and KMSMOTE compared to SMOTE) resampling strategies; RUS, ROS and SMOTE are the fundamental methods to address class imbalance issue. In other words, most of the other resampling approaches were developed based on the main principles of RUS (such as CUS), ROS (such as COS) and SMOTE (such as KMSMOTE) algorithms. Besides, adopted resampling algorithms have been used widely and successfully in various disciplines due to their less complex procedures and ease of applicability.

In the RUS strategy, the class distribution is balanced by randomly extinguishing elements in the majority class [11]. Despite the risk of information loss in the training data is regarded as the major drawback of RUS [189], there are several advantages of this algorithm. For instance, reduction of sample size enables avoiding computational expenses [94] and no artificial positive values are generated in RUS method. Since RUS is addressed as a proper alternative in case of a large dataset (Provost, 2000; Wang et al., 2017), it was included in this study.

The ROS was adopted as a method to repeat a number of minority classes to achieve a balanced class distribution [92]. Unlike under-sampling techniques, there is no risk of losing information in the ROS algorithm. However, in the case of ROS, an over-fitting problem can be observed [94]. Choi et al. [2] addressed that ROS is a proper alternative when the class distribution is highly skewed. Since the class distribution is severely imbalanced, the ROS algorithm was also applicable for the dataset. Finally, a minority-based over-sampling algorithm, namely SMOTE, which was proposed by Chawla et al. [93], was also considered as an algorithm that over-samples the minority class in the training dataset to generate synthetic samples by applying cluster-based interpolation [267]. Primarily, each minority class sample is processed along the line segment connecting the sample with its minority neighbor to generate new samples [189]. It is usually based on the combination of over-sampling the minority class and under-sampling the majority class in a simultaneous manner [93], [189]. SMOTE is highlighted to be a strong resampling alternative particularly if there are only a few dozen of minority instances [93], [268].

6.3.3 Adopted Machine Learning Methods

In this study, RF (tree-based), NB (probability-based), KNN (distance-based) and ANN (neural network-based) were used as ML methods for occupational accident outcome predictions in the construction industry. RF is an ensemble learning algorithm and a combination of DTs based on the values of a random vector that is independently sampled for each tree and has the same distribution for all trees in the forest [269]. RF creates independent bootstrap samples using the input data based on the bagging method [11]. In this algorithm, the best branched variable in each node of the DT is chosen from a small number of variables which are randomly selected among all variable sets. There are user-defined parameters in RF, which are the number of instances to be used on each node for the generation of tree structure and the number of trees to be produced. The user creates the decision forest with N number of trees. Thus, for each classification process, the dataset passes through the number of N decision trees. The trees are generated by using in-bag data and validated through the out-of-bag data. The average error rate of a model is obtained and the tree, which has the lowest error rate, is assessed with the highest weight. The lowest weights are assigned to the tree that has the highest error rate. Overall, considering that each tree has a vote, the voting is performed through an RF classifier to obtain the final prediction. Major advantages of the RF method are its ability to avoid

overfitting, provide feature importance, and work well in construction accident outcome predictions [2].

The NB method is mainly based on two assumptions. All dependent variables do not have any effect on each other and the independent variables affect the output equally [270]. These assumptions make the NB method a computationally time-effective algorithm. It examines the probability of a particular value in the target class according to conditional probabilities grounded on the Bayesian probability and allows samples to find the class to which they belong [271]. It provides the selection of the class with the highest probability of sampling to be classified according to the Bayesian rule. A model is created for a pre-defined case, which allows drawing certain results from the relationship between pre-defined case and actual observations. Since the calculation of posterior probability for a high dimensional dataset is challenging, Naïve assumption is frequently carried out, so that the values of features are conditionally independent of the others' values. The major advantage of the NB method is its efficiency with respect to both processing speed and parameter setting.

A distance-based algorithm, namely KNN, is one of the most widely used non-parametric ML methods in classification problems. It is based on the distance between the observations (or objects), and it utilizes the closest neighbor samples to fulfill the classification process [272]. During the application of KNN, each previously classified data point is stored, while the class of each point to be subsequently evaluated is decided through the comparison between the stored data and distances [273]. It is based on the principle of explaining the similarity of the samples by means of the distance between them [189]. The samples, which have similarities between each other, are close to (i.e., neighbor) and the ones not having similarities are far from the others. Therefore, the distance between two samples is regarded as a criterion that determines dissimilarity. Here, attaining a high classification accuracy with a relatively smaller number of hyperparameters and performing the classifications within a less computation time period has led many researchers to prefer using the KNN method in a variety of research areas.

ANN method emerged with the idea of mimicking the working principles of the human brain via computers. Therefore, the first studies of ANN focused on the mathematical modeling of biological cells such as neurons [274]. An ANN model consists of five different components; inputs, weights, net function, activation function, and outputs. Inputs represent the information provided to the network, while the weights indicate the

value of information coming onto a neuron. In this sense, the effectiveness and significance of the information in the system are provided by the weights. The net function allows the calculation of the net input that comes to the neuron, whereas the activation function provides the processing of the net input and determines the output. The learning process of multilayer neural networks has two stages: 1) Feed Forward and 2) Back propagation. In the first step, weights are randomly considered and the output of the network for a given sample is obtained [275]. In addition, according to the comparison of model results and expected values, the weights are updated based on the value of the gradient that is computed by backpropagation. The second step of ANN is grounded on the idea of reducing the sum of error squares through training [189].

6.3.4 Performance Evaluation

The robustness of a classifier is usually evaluated with four statistical elements in binary classification:

True positive (TP): It indicates the number of correct classifications with respect to positive instances.

False negative (FN): It indicates the number of incorrect classifications with respect to positive instances.

False positive (FP): It indicates the number of incorrect classifications with respect to negative instances.

True negative (TN): It indicates the number of correct classifications with respect to negative instances.

If there is a balanced class distribution, then the accuracy can be measured by means of the rate of correctly classified instances as defined in Eq. (6.1).

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (6.1)$$

However, there is no useful interpretation of the metric's value in case of an imbalanced dataset, since it may result in erroneously higher prediction performance [2]. In such cases, two performance evaluation metrics need to be considered, namely the true positive rate (TPR) and false positive rate (FPR) as defined in Eq. (6.2) and Eq. (6.3).

$$TPR = \frac{TP}{TP + FN} \quad (6.2)$$

$$FPR = \frac{FP}{FP + TN} \quad (6.3)$$

where TPR and FPR are the ratio of correctly classified positive instances to the true positive instances, and incorrectly classified negative instances to true negative instances, respectively. By taking TPR and FPR on the y and x axes, a receiver operating characteristic (ROC) graph, also called as ROC curve can be drawn, which is a valuable tool to visualize and compare various classifiers' performance based on TPR and FPR [276]. Hence, the larger the area under the ROC (AUROC), the better the prediction of a classifier. Despite there are some other performance indicators such as G-mean and F1 measure, AUROC was widely used recently in the literature (Ayhan et al., 2021; Chi et al., 2020; Choi et al., 2020; Hasanin et al., 2019), due to its ability to visualize the process aiding to interpret easily.

6.4 Results and Discussion

6.4.1 Classification Results

This study used an imbalanced dataset containing 120,277 non-fatal, and 725 fatal occupational accidents to predict the construction accident outcome. First, 121,002 accident dataset was divided into two major categories as training (70%) and testing (30%) sets using random hold-out approach [247]. RAW, RUS, ROS, and SMOTE resampling strategies were applied to the training set to perform fair comparisons. Resampled datasets were trained based on RF, NB, KNN, and ANN algorithms. Accordingly, performances of these algorithms were compared by using a blinded testing set, which was not subjected to resampling strategies on purpose.

For the configurations of the ML models, the trial-error approach was adopted by taking the imbalanced dataset into consideration. In the RF modeling, the number of estimators and maximum depth were sought as hyperparameters. The optimum values were searched in the range of 1 and 30 [11], and 100 and 1200 [14], respectively. Also, variance smoothing (10^{-x} , x between 3 and 9) was the only parameter to be considered in the determination of optimum model configuration for NB (Soni et al., 2020; Wu et al., 2020). In the identification of the KNN structure, three hyperparameters of the algorithm were scanned, the number of neighbors and the leaf size with the ranges of 1-30 for both parameters (Bykov et al. 2019; Zhang et al. 2020) and four different distance metrics, i.e., Chebyshev, Manhattan, Minkowski and Euclidean. The model configuration of the ANN

was also performed through trial-error considering three hyperparameters. In this regard, the momentum, learning rate, and the number of hidden layers were searched within the ranges of 0.01-1, 1-30, and 1-10, respectively (Sarkar et al. 2019; Son and Kim 2015). According to the results of trials performed based on the imbalanced dataset, the best performances were observed i) when the number of estimators and the maximum depth were set as 1100 and 20, respectively in the RF algorithm; ii) when the variance smoothing was set as 1×10^{-8} in NB algorithm; iii) when the number of neighbors was 30 and the leaf size was 1 in the KNN algorithm; and iv) when the momentum, learning rate, and hidden layer size were set as 0.12, 0.01, and 6, respectively in ANN.

The AUROC was used as the performance evaluation metric for the comparisons of each ML algorithm and resampling strategies for 16 scenarios in total. Figure 6.2 illustrates the ROC curves of ML algorithms subjected to RAW (Figure 6.2a), RUS (Figure 6.2b), ROS (Figure 6.2c), and SMOTE (Figure 6.2d) resampling strategies. The results imply that RF showed the best performance with 0.809 of AUROC when the training data was not subjected to any resampling strategy (RAW). The prediction rate showed significant improvements when ROS and RUS algorithms were applied, with AUROC values of 0.820 and 0.831, respectively. ANN and KNN methods also showed the highest improvements with resampled dataset through adopting ROS and RUS algorithms, from 0.500 to 0.785, and from 0.590 to 0.685, respectively. This shows that adopting the resampling strategies significantly improved the prediction accuracy. NB was the only method, which achieved nearly the best performance with a training dataset that was not subjected to resampling strategies (AUROC: 0.763 in RAW compared to 0.765 in ROS).

Table 6.4 provides the summary of evaluated scores and ranking of both ML algorithms and resampling strategies. Among the four ML trials performed with four different resampling strategies, RF pre-processed with RUS presented the highest prediction rate and the AUROC was found as 0.831. The RF model coupled with RUS exhibited small differences with other resampling strategies in the RF algorithm apart from SMOTE (RAW: 0.809, ROS: 0.820, SMOTE: 0.740). On the other hand, relatively large differences were observed with other ML methods subjected to RUS resampling strategy based on the AUROC (ANN: 0.727, NB: 0.746, KNN: 0.685).

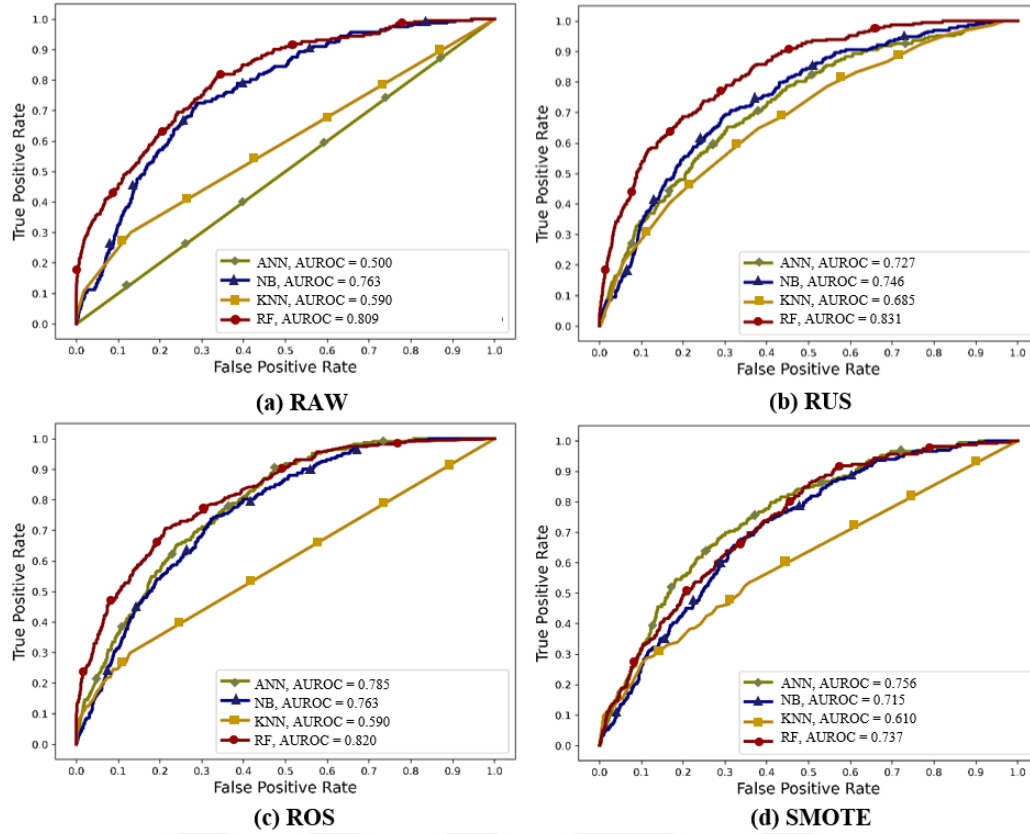


Figure 6.2 ROC curve of a. Raw models b. RUS-based models, c. ROS-based models, d. SMOTE-based models.

In addition, Table 6.4 presents the inner and classifier ranks. The inner ranks denote the success rating of a resampling strategy obtained for each ML algorithm and the outer ranks show the ranks of ML algorithms for each resampling strategy. With respect to the inner ranking of the ML algorithms, RUS was the best resampling strategy in RF and KNN, while it was ROS in NB and ANN models. Results also showed that SMOTE was not the favorable resampling strategy in any of the ML algorithms. In terms of classifier rank subjected to different resampling alternatives, RF was the best ML method in all scenarios (Table 4), in which the prediction rate was significantly higher than the others. The only exception was that ANN showed slightly better performance than RF when the dataset was resampled with SMOTE, with an AUROC value of 0.756 compared to 0.737. The results show similarities with other studies in the field of construction accident predictions in general. The RF was determined as the most accurate ML model in many studies, which compared the performances of several ML methods in construction accident predictions [4], [7], [11]. In line, Choi et al. [2] found the performances of RF as significantly higher than several ML algorithms such as DT, LR, and AdaBoost.

Table 6.4 Summary of evaluated scores

Model	Sampling technique	AUROC	Inner rank	Classifier rank
RF	RAW	0.809	3	1
	RUS	0.831	1	1
	ROS	0.820	2	1
	SMOTE	0.737	4	2
NB	RAW	0.763	2	2
	RUS	0.746	3	2
	ROS	0.765	1	3
	SMOTE	0.715	4	3
KNN	RAW	0.590	3	3
	RUS	0.685	1	4
	ROS	0.590	3	4
	SMOTE	0.610	2	4
ANN	RAW	0.500	4	4
	RUS	0.727	3	3
	ROS	0.785	1	2
	SMOTE	0.756	2	1

In this study, testing and training sets were identified based on random test/train splitting method with the randomly selected 30% and 70% of the dataset. However, the performance of the classifiers may change according to the separated testing sets, which needs to be considered for model validation. There are two commonly used cross-validation (CV) techniques; random hold-out (i.e., 30/70, 40/60) and shuffled 5-fold CV. These techniques examine how the results can be generalized when an independent dataset is introduced [247]. Therefore, 15 additional test/train splitting scenarios were considered in the context of CV to validate the performance of the RF-RUS pipeline. These scenarios are three folded: (1) 5 testing and training sets captured by a similar random splitting approach (i.e., 30/70), (2) 5 randomly generated test/train sets with a different but another commonly used test/train splitting ratio (40/60), and (3) 5 testing and training sets separated according to the 5-fold CV scheme. As explained before, only training sets should be balanced with resampling methods. Testing sets should remain imbalanced for practicality and satisfying real-life conditions. Therefore, during the 5-fold CV application, once the testing and training sets were separated, each training set was then resampled while using different and imbalanced testing sets. The results of test/train splitting scenarios are shown in Table 6.5. According to the findings, the RF-RUS model showed strong performance in each test/train approach such that the

performances of the models with 30/70, 40/60, and CV strategies were found as 0.829, 0.820, and 0.796 of average AUROC values, respectively. Therefore, RF-RUS can be regarded as a valuable method to implement in construction safety predictions, which can also be used to investigate attribute importance affecting construction accident outcomes.

Table 6.5 Alternative test/train splitting scenarios

Scenario ID	Test (Ts)/train (Tr) approach	Cross validation approach	AUROC	Sub-average	Standard deviation
Baseline 0	30/70	Random hold-out	0.831	–	–
1	30/70	Random hold-out	0.825	0.829	0.005
2	30/70	Random hold-out	0.832		
3	30/70	Random hold-out	0.830		
4	30/70	Random hold-out	0.836		
5	30/70	Random hold-out	0.824		
6	40/60	Random hold-out	0.811	0.820	0.012
7	40/60	Random hold-out	0.816		
8	40/60	Random hold-out	0.841		
9	40/60	Random hold-out	0.814		
10	40/60	Random hold-out	0.819		
11	CV (Tr-Tr-Tr-Tr-Ts)	Shuffle 5-fold CV	0.777	0.796	0.025
12	CV (Tr-Tr-Tr-Ts-Tr)	Shuffle 5-fold CV	0.791		
13	CV (Tr-Tr-Ts-Tr-Tr)	Shuffle 5-fold CV	0.840		
14	CV (Tr-Ts-Tr-Tr-Tr)	Shuffle 5-fold CV	0.783		
15	CV (Ts-Tr-Tr-Tr-Tr)	Shuffle 5-fold CV	0.788		

6.4.2 Attribute Importance

The RF classifier can present the importance of attributes based on several indices such as feature importance, Gini importance, and permutation importance [282], [283]. In this study, permutation importance, which is one of the most reliable importance indices especially when the attribute dimension is high [217], [218], was used. As being a model-agnostic method, permutation importance examines the impact of variations in the dataset on the model results rather than determining what is important in the model structure. The

permutation importance is based on the identification of the importance of a variable by replacing it with noise data that comes from the same distribution, instead of removing this variable from the model [269]. Thus, permutation importance relies on the accuracy difference between the baseline model (RF) and the updated model which shifts each input variable in the validation set [11]. The permutation importance utilizes the mean-decrease-accuracy strategy, and the importance values are highly related to the performance metrics (such as F1, AUC, and accuracy). This makes the permutation importance acceptable to use for ranking the attributes rather than calculated numerical values [11].

Attribute importance and ranking calculated through permutation importance with respect to four resampling strategies are shown in Table 6.6. Results showed that the top five attributes were “number of workers in the company”, “material used”, “age”, “number of past accidents in the company”, and “construction type” when the imbalanced training test (RAW) was used. It can also be noted that the importance of attributes marginally varied according to the adopted resampling strategy. For instance, in the RUS algorithm, “year” and “time of accident” were among the top five attributes instead of “number of workers in the company” and “construction type” compared to the RAW strategy. The same seven attributes (with respect to RAW and RUS algorithms) were also considered among the top five attributes in the ROS algorithm. The results were similar when the training set was subjected to the SMOTE, apart from “number of past accidents of the worker” feature, which was ranked among the least important attributes by adopting other resampling strategies (RAW: 14, RUS: 14, ROS: 13).

Some additional highlights could also be noted from the analysis results. The number of past accidents in the company was found to be the most significant attribute affecting the outcome of a construction accident based on three resampling techniques (RUS, ROS and SMOTE). This finding could be considered important, since, to date, this feature was not considered in previous studies for the prediction of construction accident outcomes. With this result derived in this study, it can be suggested that the prediction models proposed in the literature can be improved by including the “number of past accidents in the company” attribute, which is already known before the occurrence of accidents.

Table 6.6 Attribute ranks based on permutation importance

Attribute list	Attribute importance				Attribute rank			
	RAW	RUS	ROS	SMOTE	RAW	RUS	ROS	SMOTE
Day of the month	0.0030	0.0052	0.0351	0.0162	10	11	10	13
Day of the week	0.0028	0.0057	0.0279	0.0298	12	9	12	11
Month	0.0028	0.0029	0.0312	0.0421	11	12	11	9
Year	0.0033	0.0202	0.1871	0.1213	7	4	5	2
Time of the accident	0.0033	0.0122	0.1081	0.0094	8	5	7	14
Gender	0.0000	0.0000	0.0000	0.0002	15	15	15	15
Age	0.0036	0.0533	0.2459	0.0839	3	2	2	3
Working days of the worker	0.0033	0.0105	0.0924	0.0269	9	7	8	12
Education level	0.0019	0.0020	0.0070	0.0514	13	13	14	7
Daily wage	0.0034	0.0060	0.1600	0.0629	6	8	6	6
Number of past accidents of the worker	0.0011	0.0010	0.0179	0.0657	14	14	13	5
Material used	0.0037	0.0228	0.2070	0.0745	2	3	4	4
Construction type	0.0034	0.0055	0.0490	0.0474	5	10	9	8
Number of past accidents in the company	0.0035	0.0865	0.2680	0.1249	4	1	1	1
Number of workers in the company	0.0037	0.0107	0.2095	0.0405	1	6	3	10

“Age of the worker” was another attribute that was found significant. This attribute was also pointed out in earlier prediction studies with its potential impact on the occupational accident outcome (Ayhan and Tokdemir 2019a; Choi et al. 2020; Kang and Ryu 2019; Lee et al. 2020). This can be related to the fact that older workers, some of whom having lower physical and sensory capabilities (Chi et al. 2005), could consider some of the safety training programs as unnecessary work [226], which might eventually increase the probability to experience severe accidents.

The ranking of attributes also highlighted that “material used” was an attribute that has a significant effect on the outcomes of construction accidents. A similar finding was noted by Kang and Ryu [11], where they also found “original-cause material” as the second most influential attribute in predicting construction accident type, particularly related to scaffolding, stairways, ladder, edges and openings. These special works could be related to falls from height, which was also deemed as one of the main causes of occupational fatalities (Melchior et al., 2020; Wu et al., 2010).

Regarding time-related attributes, the year of the accident was found to be an important factor since it might represent the national or sectoral transformation and development of OHS systems over years, which was also found by Lee et al. [8] among the major variables in accident severity prediction. Another important time-related factor was determined as “time of accident”, which can be influential in fatality risk due to the changing environmental conditions, fatigue of workers and visibility in construction sites, as also addressed by Ajayi et al. [240]. Arditi et al. [286] addressed that construction works during night-time can be more dangerous than day-time, addressing the need to develop strategies to improve workplace safety at night shifts.

6.4.3 Discussions of the Adopted Techniques

According to the results obtained in this study, the RF coupled with the RUS algorithm showed the best performance to predict accident outcomes. Therefore, despite the common fact that the occupational accident dataset might be severely imbalanced, the RUS algorithm could be adopted to improve the prediction accuracy when the model is trained with the RF algorithm. In the pertinent literature, there are several studies illustrating the usability of the RF algorithm in occupational accident-related predictions such as Tixier et al. [7], Kang and Ryu [11], Choi et al. [2], and [14]. Among them, Tixier et al. [7] and Baker et al. [14] utilized construction accident datasets from different

countries, while Kang and Ryu [11] and Choi et al. [2] acquired datasets from the Korean national database. Considering the fact that the dataset collection procedures heavily rely on the country-based conditions, availability, and accessibility of the corresponding features, these studies included different accident attributes.

The RF model was also coupled with various resampling approaches in past studies. For instance, Tixier et al. [7] applied stratified over sampling, Kang and Ryu [11] coupled the RF with the RUS algorithm, Choi et al. [2] implemented the ROS algorithm, and Baker et al. [14] performed the calculations of weights inversely proportional to the frequency of each sample category to deal with the class imbalance issue. Despite the divergence in main objectives of these studies (prediction of injury type, energy type, body part impacted, injury severity, accident type, and accident outcome), they demonstrated the instrumentality of the RF algorithm with the resampling techniques. However, none of the abovementioned studies have examined the performance comparison through scenario-based evaluation that consisted of coupling ML methods with resampling techniques.

This study compared RAW, RUS, ROS, and SMOTE resampling strategies to predict the outcomes of construction accidents. Some of the factors were found more significant than others according to the feature importance results. Therefore, focusing on these features would improve the performance of more effective safety management applications. However, attained attributes might be different depending on the data collection procedure of companies or countries. For instance, some environmental factors (uneven surface, congested work environment, unstable support, etc.), project factors (project cost, project complexity, project duration, etc.), or OHS factors (experience of OHS team, number of OHS staff) can also influence the accident outcome predictions. In this study, the most powerful model (RF-RUS) yielded a powerful prediction accuracy (AUROC = 0.831) with the included factors. The prediction performance of the developed models and the context of the safety management applications could be improved with additional attributes. Since the applicability of the RF method was proven many times by other researchers in the field with various datasets as well, the RF-RUS model can be suggested as a useful and effective prediction tool for different accident datasets.

The results show that the performance of the SMOTE algorithm was relatively lower compared to RUS and ROS algorithms. Feature importance results were also significantly different in SMOTE algorithm compared to other resampling algorithms. For instance,

“number of workers in the company” was among the most important attributes according to the RAW (1st), RUS (6th), and ROS (3rd) algorithms, while it is 10th order of importance in SMOTE-based analysis. Several reasons can be related to the analysis results. First, SMOTE algorithm performs synthetic sample generation through cluster-based interpolation which can potentially results in biased estimations in categorical variables even though it works well with numerical attributes. Additionally, synthetic samples generated with the SMOTE algorithm can also be impractical and not representative in case of extreme values in the feature. This can be observed in some national datasets since they comprise accidents in small companies as well as accidents in mega projects.

To present more insight into resampling strategies, Figure 6.3 presents the distributions of the “number of past accidents in the company” vs. “number of workers in the company” in the training data that were subjected to RAW, RUS, ROS, and SMOTE algorithms, respectively. Here, it can be observed that SMOTE performed cluster-based interpolation based on all instances in the dataset regardless of the distribution of the feature values, which might be the reason for low feature importance of the number of workers in the company. In other words, synthetic samples generated with the SMOTE algorithm were not representative of the real observations, which can also be the reason for poor performance of SMOTE algorithm (Figure 6.2). Since many of the national occupational accident datasets include different sizes of companies with the number of workers as well as the number of past accidents, SMOTE algorithm is not recommended for accident outcome prediction studies that use national datasets, which are more likely to include extreme values. In order to investigate the applicability of the SMOTE-based resampling methods, researchers can consider comparing the performance of BL-SMOTE, KM-SMOTE, and SVM-SMOTE to represent training datasets in future studies. In addition, since many of the accident datasets include both numerical and categorical variables, performance of the SMOTE nominal categorical (SMOTE-NC) algorithm can also be evaluated in future studies.

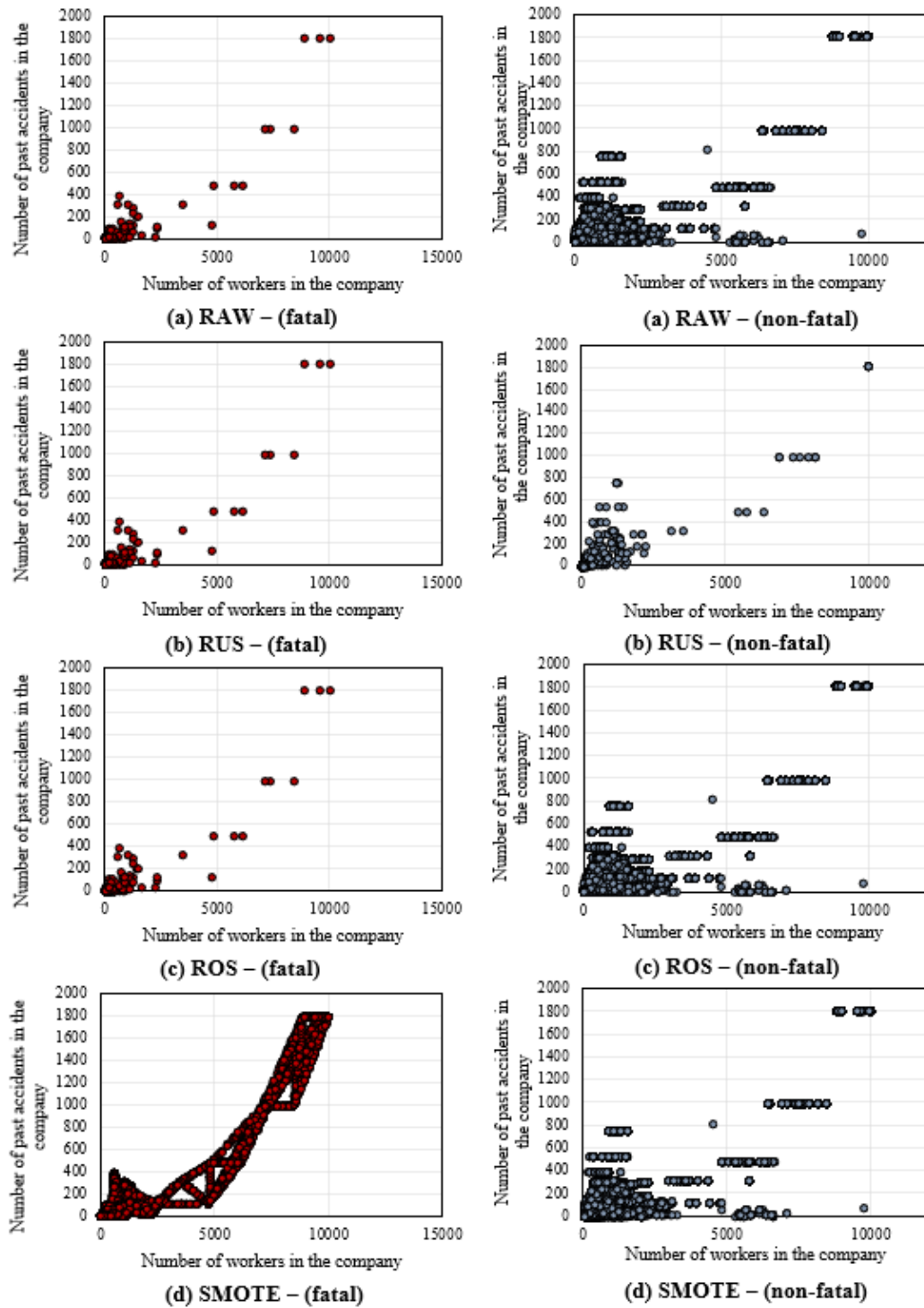


Figure 6.3 Distribution of number of past accidents in the company versus number of workers in the company based on a. raw data (RAW), b. resampled (RUS), c. resampled (ROS), d. resampled (SMOTE)

6.5 Implications for Safety Authorities and Practical Use of the Study

The most significant attributes considering the RUS algorithm comprised worker-related (age and material used), company-related (number of past accidents in the company and

number of workers in the company), and time-related (year and time of an accident) attributes (Table 6.5).

Among the worker-based attributes, “material used” was identified as a causal significant factor for the occurrence of fatal accidents. Therefore, additional safety training contents can be considered to raise the awareness of workers regarding such materials. Since some materials can be directly associated with dangerous environments (such as scaffolding or ladder with falling) [11], they can also be associated with severe accidents, thus increasing the probability of these workers to experience fatal accidents.

Workers’ age is another important class of the worker-related attribute. Several past studies have already addressed that the older the workers, more likely to experience severe accidents compared to younger workers [228]. This can be considered due to the decrease in physical capabilities of the elderly (Chi et al. 2005), which is further supported by Choi et al. [287] that workers having 50+ years old are accounted for more than 60% of the fatal accidents. Enhancing the coverage of the training sessions and increasing their frequency would improve the awareness level of older workers about actions and behaviors that would result in occupational accidents.

Apart from the worker-related attributes, company-related attributes, such as number of past accidents in the company and number of workers in the company would also contribute to expanding the extent of related studies. Many studies stated that the number of workers may be one of the indicators to define the size of a company. In this respect, the higher the number of workers in a company, the lower the fatal accident ratio of the company (Jeong 1999). This conclusion may be explained by the fact that workers in smaller companies are usually assigned to multiple tasks, in which they are less competent. This would compromise the safety management efficiency in many projects. For this reason, increasing the number of workers with specialized qualification for certain tasks could alleviate the safety-related risks in working environments.

In addition, one of the exclusive project-based attributes, number of past accidents in the company/project, plays a decisive role in elaborating the safety performance of the construction sites. Evidence suggests that companies that experienced occupational accidents in earlier projects are more likely to establish and perform safety management strategies with more caution and companies without past accident records could be at higher risk. On the other hand, young companies could be an exception in this regard,

having less experience in the business and limited possibility of exposure to occupational accidents. Such companies might need more efforts to be prepared against the occupational accidents by allocating the necessary budget for OHS investments.

The body of knowledge also recognizes the importance of the time-related attributes on the occurrence and severity of accidents [2], [8], [240]. For instance, managers or supervisors can be away from job sites for some hours during the day due to other responsibilities (meetings, visits, etc.), which might increase the fatality rates. Employment of temporary personnel to substitute the role of safety manager would support the maintenance of safe working environments in such cases. Furthermore, considering the fact that many occupational accidents occur during night shifts [286], companies are recommended to perform continuous monitoring of risks in construction sites and regulate working hours to decrease the probability of accidents.

Overall, practitioners can be suggested to develop strategies based on a dataset related to these attributes to prevent occupational fatalities. The necessary measures should be prepared and taken timely, since the time of an accident is an important factor in occupational fatalities. Many construction companies have their own occupational accident databases. However, these are merely considered in predictive models to improve safety management plans. Based on the key findings of this study, an explanatory guideline is presented in Figure 6.4 that can be included in safety management plans for construction sites. The proposed framework for the utilization plan of the guideline can also be seen in Figure 6.5 illustrating the data entry and prediction procedures, with reference to the “A” in Figure 6.4.

Safety managers can run the proposed model regularly (such as daily, weekly, or monthly). In this study, the applicability of this model to a monthly investigation was emphasized (Figure 6.5). Once the model training is performed based on the historical accident records; company, worker, and time-related attributes are taken from the construction site regarding all workers. In this context, assuming that there is k number of hours in the day shift, n number of days in the month, K number of workers working in the company, and N_i number of material groups that the workers will engage in during the month, then the size of the testing set can be calculated as $N_{Month} (k \times K \times n \times N_i)$. By running the proposed model, safety departments can perform worker-based (privately) inspections about the most risky conditions (i.e., materials, time of the day, day of the month, day of the week) for the workers with higher fatality risk of an occupational

accident. Such predictions can contribute to the implementation of timely and effective OHS practices on construction sites.

It is important to note that the features included in the proposed explanatory guideline (Figure 6.4 and Figure 6.5) are limited by the dataset used in this study. Hence additional features (i.e., the project and company-related factors, environmental factors, worker-related factors, and supervisor/OHS team-related factors) can be incorporated into the guide based on the availability of the data. With the assistance of a reliable and robust accident database, practitioners can develop an ML model for predicting accident outcomes as a part of their project management plan.

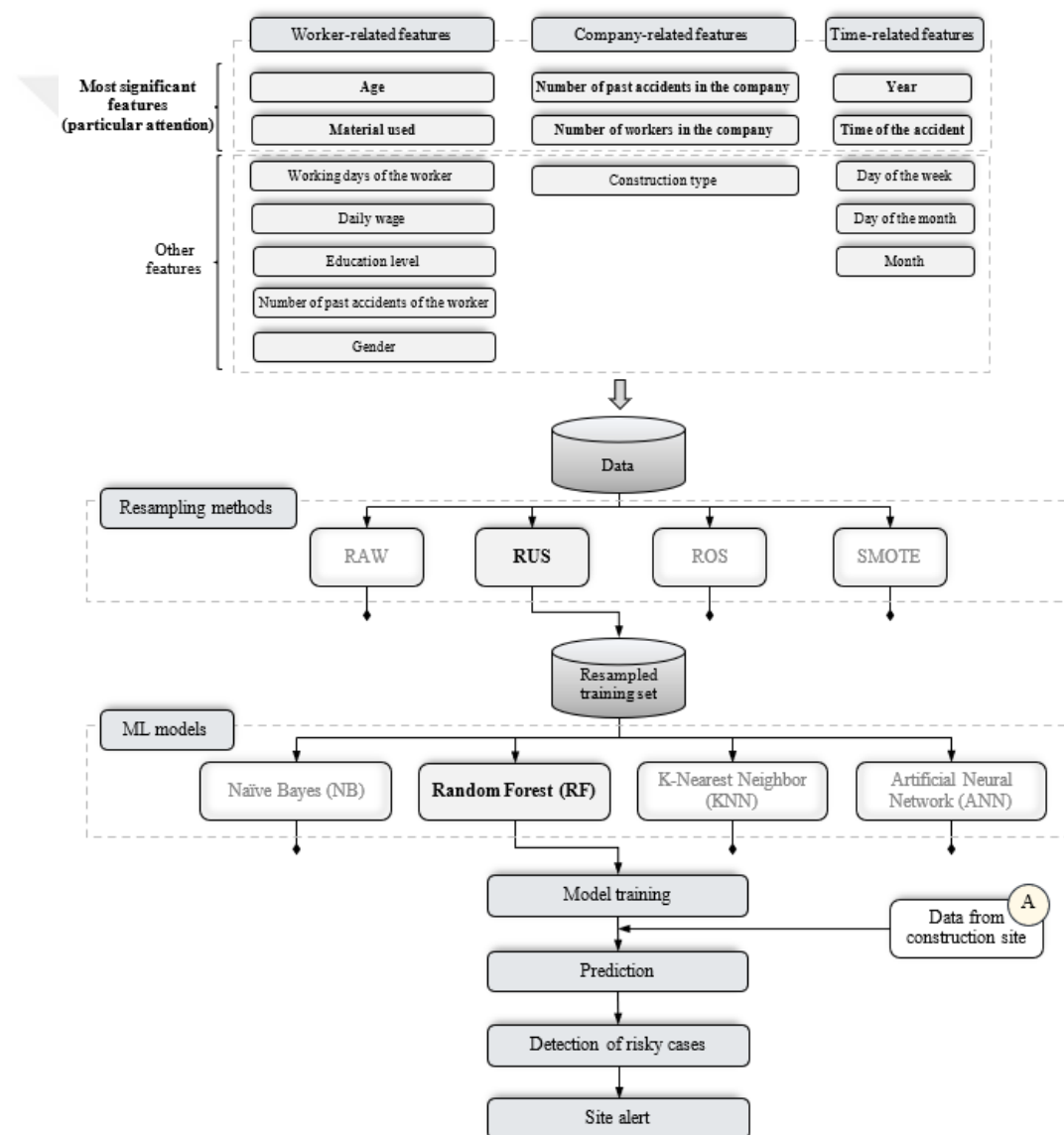


Figure 6.4 Practical use of the research

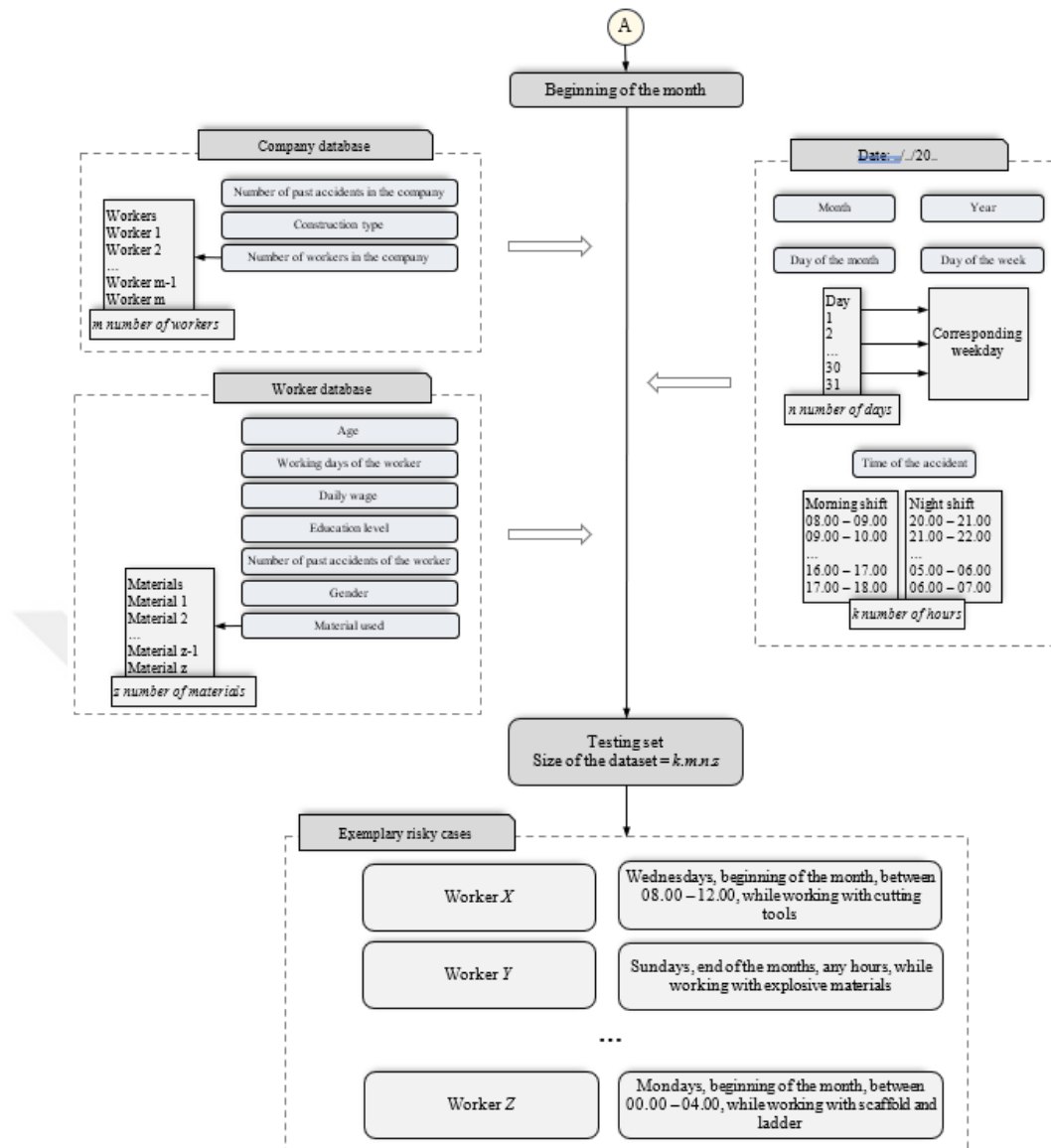


Figure 6.5 Example of a utilization plan

In order to ensure the feasibility of the proposed approach in real cases, five professionals including a safety personal (R1), OHS manager (R2), technical office engineer (R3), site supervisor (R4), and project manager (R5) in a construction company were contacted. The proposed model, attained results, and developed utilization plan (Figure 6.4 and Figure 6.5) are shared with them to further elaborate their practicality using a 7-point Likert scale (i.e., 1: Strongly disagree and 7: Strongly agree) based on semi-structured interview questions. The responses are shown in Table 6.7. Along with the findings presented in Table 6.7, the respondents also highlighted the applicability of the model as the attributes used in the model have been collected all the time due to SSI requirements. However, nearly all respondents also addressed that the proposed model would be more useful if “no accident” cases were also included, which was one of the limitations of the

study due to the lack of “no accident” information in the dataset. The results show that the respondents had positive views on the proposed ML model and the utilization plan.

Table 6.7 Respondents’ opinions about the proposed model

Questions	Respondents					
	R1	R2	R3	R4	R5	Average
Developed ML model can be used effectively to improve the safety performance of the company.	7	7	6	6	7	6.60
Developed model is applicable to practice in the company.	7	6	6	6	7	6.40
Generally, I agree with the most significant features calculated through the RF-RUS algorithm.	7	6	5	6	6	6.00
Utilization plan (Figure 6.5) is sufficient to detect workers who have a high probability to experience fatalities.	7	7	6	7	6	6.60
Utilization plan (Figure 6.5) can be used effectively in construction projects as an early alert mechanism.	6	6	7	6	5	6.00
Early detection of workers and specific conditions that are related to fatalities can help minimize fatal accidents on construction sites.	7	7	7	7	6	6.80
Monthly utilization plan is appropriate to use in construction sites as many important decisions about safety measures are made on a monthly basis.	6	7	6	7	7	6.60
I would like to use the proposed model in the project to detect the susceptibility of construction workers to fatal accidents.	7	7	7	7	6	6.80

Overall, findings are evident for the prediction of fatal accident cases as much as possible. Safety managers are suggested to focus on workers, material/equipment, and time zones that could potentially result in fatal accidents. In this context, real time data collected from construction sites can be very useful and adopted with the proposed model. Safety professionals could consider mitigation measures in advance on construction sites to prevent fatalities.

6.6 Conclusion and Future work

Machine learning based prediction models can be useful in enabling the identification of potential sources and outcomes of occupational accidents in construction projects. In practice, the existence of extremely imbalanced datasets (with a significantly higher number of injuries compared to fatalities) makes it difficult to predict occupational fatalities. In this context, the use of resampling techniques could provide substantial improvements in the accident outcome predictions. Hence, the purpose of this research was to predict construction accident outcomes using machine learning methods coupled with resampling techniques.

The results acquired from the ML models, which were pre-processed with resampling strategies, revealed significant findings. Among different resampling techniques coupled with four ML methods, variations were observed in model performances. RUS was found as the best resampling strategy when it was coupled with RF and KNN, while it was the third best technique with the ANN and NB implementations. In addition, compared to the other resampling techniques, SMOTE showed the worst performance in RF and NB models, whereas the RAW was found as the strategy with the lowest accuracy among KNN and ANN models. The most noticeable finding from this research was that RF outperformed other adopted ML methods according to the AUROC values for all resampling techniques (except SMOTE). Overall, the best performing model was obtained through the integration of RUS strategy and RF algorithm; in which AUROC was calculated as 0.831. Based on the attribute importance results, it can also be concluded that the number of recorded accidents in the company, age of the victim, material used, number of workers in the company, year, and time of the accident were the most significant attributes, which require careful examination to improve safety management initiatives.

Besides the crucial findings attained from the analysis, this study poses some limitations that need to be considered in future attempts. First, adopted features herein are limited by the national dataset acquired. Exploring the accident outcomes with more features defining the project characteristics, work environment, worker conditions, and OHS procedures more precisely can significantly contribute to the development of an improved proactive safety risk management system on construction sites. The prediction power achieved with a limited number of features is 0.831 (regarding AUROC), which highlights that including more exclusive features can help the detection of fatal accidents

at very high accuracies. Similarly, some of the post-accident attributes, (injured body part, type of injury, reason for accident, etc.) can also be predicted in future studies to illustrate the robustness and usability of the proposed framework. Besides, since there could also be “no accident” with the adopted attributes, dataset regarding non-accident cases can also be included in future studies to evaluate the applicability of the proposed model in detecting accident occurrences as well as its severity. Since “no accident” data is rarely recorded in construction sites, national data collection procedures should require cases/conditions with “no accident” in understanding the conditions that trigger the occurrence of occupational accidents. With a national accident/no accident dataset, the researchers can be able to perform improved accident occurrence predictions and boost the practical implications of the investigations in the field. From the methodological point of view, the performances of other under-sampling methods such as cluster-based under-sampling and case control under-sampling can be explored in future studies to assist practitioners and researchers select the most useful under-sampling approach for the prevention of fatal accidents. The use of SMOTE algorithm is questionable in future since its performance was low due to the impact of categorical variables and extreme values commonly exist in numerical variables in national datasets. Instead, the applicability of the BL-SMOTE, KM-SMOTE, SVM-SMOTE, and SMOTE-NC can be explored to observe whether they would work well with accident datasets.

Through diligent examination and prediction of construction accidents based on the imbalanced dataset, safety managers and industry practitioners can improve the safety conditions of individuals and projects by imposing necessary measures before the occurrence of any fatal accident. Resampling strategies investigated in this study showed remarkable improvements by adopting each ML model. Therefore, the prediction model proposed can be used, even when there is a severe imbalance between fatal and non-fatal accident records. The results emerged from the attribute importance evaluations suggest that both worker and company characteristics are strongly related to the outcome of construction accidents. These findings can contribute to the development of several strategies to prevent potentially fatal incidents in the construction industry. The current research is also expected to provide a visionary outlook for the safety professionals to always stay prepared, as the complex and uncertain nature of the projects and working environments in construction sites can easily result in fatal accidents.



This study investigates the applicability of ML methods in predicting occupational accident outcomes in several ways using national dataset acquired from the SSI. The primary focus of investigation was to explore the practical use of national datasets in future projections. The major motivation of the thesis is that a huge number of occupational accidents were collected by the SSI after the enactment of Law 6331, which states that employers must report all occupational accidents to the SSI. Hence, After the data collection step of this thesis, the dataset was investigated, and several prediction schemes were generated. In this thesis, four projections were presented:

- Post-accident permanent disability prediction was employed by integrating feature engineering, tree-based ML methods (RF, AdaBoost, XGBoost, ET), and GA for hyperparameter tuning (Model 1).
- Time series prediction model was generated by hybridizing DWT and three commonly used ML methods such as ANN, MARS and SVR for daily nationwide number of accident prediction (Model 2).
- A novel scenario-based comprehensive pre-processing model that comprise 216 preprocessing scenario was proposed for accident severity prediction using XGBoost algorithm (Model 3).
- Accident outcome prediction model was proposed by integrating one ML methods in each ML family (RF of tree-based family, NB of generative model family, ANN, of network-based family, and KNN of neighbor-based family) and several data resampling methods (RUS, ROS, SMOTE) (Model 4).

Several research gaps were addressed in each of the proposed models. First of all, there were a limited number of occupational accident predictions based on national dataset, and existing studies were mainly used datasets from developed countries. In this thesis, a huge dataset including 395,777 occupational accidents recorded in Turkish construction industry over a decade was statistically investigated. The findings emerged from the analysis results indicate that number of past accidents in the company, material used,

number of workers in the company, age of the worker, year, working days of the worker, and daily wage were strong precursors of occupational accident outcomes.

Secondly, the robustness of ensemble tree-based ML methods has widely been documented in the literature. However, their comparative analysis has not been performed in the construction safety literature. We found XGBoost as the most powerful tree-based ML method in permanent disability prediction, despite the performance of others were very similar to it. In addition, one of the overlooked tree-based method, namely ET, was the most powerful predictor regarding minority class instance.

Similarly, despite there has been several research on ML applications in construction safety management research domain, permanent disability prediction has been overlooked in the literature. This prediction scheme is socially crucial because construction workers who perform their jobs to meet variety of needs of future generations are at significant risk of not meeting their own needs. Hence, by predicting post-accident permanent disability status of construction workers, the precursors of permanent disability have been investigated in Model 1. Working days lost, and injured body part were found as significantly effective features in workers' permanent disability.

A comprehensive literature survey revealed that metaheuristic optimization algorithms were not employed in the construction safety management literature. However, hyperparameters of ML methods are known to have significant impacts on the performance of future projections. Therefore, for the first time in the construction safety literature, GA was employed to tune parameters of RF, ET, XGBoost and AdaBoost. It was found that hyperparameters of tree-based ML methods can be configured effectively by using metaheuristics.

Another gap in the literature was a detailed feature engineering. The need for comprehensive feature engineering comes from the fact that there could be several types of errors during data collection step (human errors or transaction errors) and collected data cannot be used directly. In other words, there might be missing data, outliers, different feature types and imbalance in the class distributions in the collected datasets. After these issues were observed in the dataset, another literature survey also revealed that such issues exist nearly all national datasets. Therefore, we proposed a multi-step feature engineering methodology including data encoding, data scaling, dimension reduction, and data resampling. For dimension reduction we also implemented PCA as

one-hot encoding resulted in a huge dimensionality. This becomes more crucial solution in Model 1 since in this model GA was also employed for parameter tuning. Without PCA application, the computational load would be significant burden. Another option could have been not employing one-hot encoding, but in this case, the misinterpretation of the model could have been a significant barrier (if label encoding was employed) regarding performance of the models.

With the recorded accident dataset, it was realized that the number of occupational accidents can be predicted through time series analysis. A comprehensive literature survey revealed that such a time series analysis has not been performed in the literature yet. Therefore, by using three commonly used regression models (ANN, MARS, SVR), we developed a time series prediction model in Model 2 for short-term, mid-term and long-term time periods, 1-day, 7-day, and 30-day ahead, respectively. By using the proposed model, safety authorities can develop nation-wide action plan before the most risky periods throughout a month. Stand-alone model results show that MARS, ANN, and SVR were the most powerful regressors regarding short-term, mid-term and long-term projections, respectively.

After performing time series analysis, it was observed that the prediction power was relatively low regarding any method (ANN, MARS, SVR) and projections (1-day, 7-day, and 30-day ahead). A literature review showed that DWT can be used effectively to divide time series data into its scale components. With DWT, seasonality, cycling, and trend in time series data can be considered and separated. With this insight, subcomponents can be predicted differently and then summed, leading to overall projections. This hybrid methodology has not been acknowledged in the construction safety literature. Findings show that W-ANN was the most powerful hybrid algorithm regarding 1-day and 7-day ahead, while it was that of W-MARS regarding 30-day ahead projection. Performance of the models were calculated as significantly high regarding T+1 and T+7 projections with NSE value of 0.9275 and 0.7678, respectively.

As mentioned before, there has been several issues related to the collected dataset. There are many preprocessing steps available and used in the literature. However, many options can be considered in prediction models. For instance, RUS, ROS and SMOTE are commonly used data resampling methods. Still, identifying the most appropriate data resampling method is essential. Therefore, comparative analysis of resampling methods was a proper research direction in this regard. On the other hand, there are also many

options in other preprocessing steps. For example, *MinMaxScaler* and *StandardScaler* are two commonly used data scaling options. Similarly, applying data binning and not applying data binning are two potential alternatives to consider numerical features. Moreover, there are various missing data handling or outlier handling strategies. Despite comparative analysis of these algorithms can be useful, a scenario-basis automated preprocessing model that identifies the best preprocessing pipeline is essential, and has not been considered in any domain. Hence, in Model 3, we developed a scenario-based automated data preprocessing model to identify the most effective preprocessing step in construction safety management. The findings highlight the following pre-processing pipeline: (i) including categorical missing data as “other”, and numerical missing data as the mean of the feature, (ii) not considering data binning, (iii) not considering outlier handling, (iv) applying *MinMaxScaler* as a data scaling, (v) using one-hot encoding for all the categorical features, and (vi) resampling the dataset with ROS algorithm. However, it was also found that RUS algorithm was more powerful in recognizing minority class instances. Since the main objective of such prediction studies was to detect minority class instances (such as fatality), RUS can also be used confidently for future accident projections.

Among these preprocessing steps, data resampling has always been one of the central topics that received considerable attention. Therefore, we developed a model to predict occupational accident outcomes based on national data using ML methods coupled with several resampling strategies. In this regard, one ML method in each ML family (RF, NB, ANN, and KNN from tree-based, probability-based, network-based, and distance-based families) was coupled with 3 most commonly used data resampling methods (RUS, ROS and SMOTE) resulting in 12 prediction results. In this way, each ML family and their performance in varying resampling methods was investigated. Findings highlight that RF algorithm was the most powerful ML method if the dataset is resampled with RUS or ROS algorithms. On the other hand, ANN was the most proper option for datasets resampled with SMOTE algorithm. Overall findings highlight that the RF coupled with RUS achieved the highest prediction power.

Findings emerged from this thesis indicate significant improvements in the safety management applications in the construction industry. The following conclusions, recommendations, and corresponding future research directions are derived from this thesis.

- The findings indicate that RUS and ROS are promising resampling methods, which can be applied in construction sites to identify the susceptibility of construction workers to occupational fatalities. It is important to note that imbalance in the class distribution might be significantly high in construction accidents and according to the dataset collected from the SSI, there are more than a hundred non-fatal accidents for each fatal accident. Therefore, data resampling needs to be carefully considered by construction professionals to predict potential accident outcomes of construction workers. In this vein, despite SMOTE showed low prediction accuracy, there are other SMOTE-based resampling methods. Therefore, a comprehensive and comparative study on the resampling methods would be very beneficial for research society and practice. Hence future research may consider comparing the following resampling methods: RUS, ROS, SMOTE, CUS, COS, CCUS, BL-SMOTE, SVM-SMOTE, KM-SMOTE, SMOTE-NC, CB-SMOTE etc. Such research would be very beneficial for the research society in addressing the performances of any resampling methods.
- It was found that tree-based algorithms, particularly RF and XGBoost were highly preferable compared to other methods. In addition to high performance of tree-based methods, they can provide attribute importance, which might enhance the visibility of the proposed models. Therefore, comparative analysis of tree-based algorithms would be fertile area for improvement not only in construction safety management domain, but also other research areas of construction management.
- As one of the metaheuristic optimization algorithms, GA showed effective configurations in the ML model setup. Therefore, performance of other metaheuristics such as particle swarm optimization, bees algorithm, grey wolf algorithm, cuckoo search algorithm, and ant colony algorithm would be an interesting future research are to observe which ML method can be optimized with less computational load with which metaheuristics. However, it needs to be mentioned that, computational speed is lowered once the dimension and instance in the dataset increases. In this case, applying PCA for dimension reduction and resampling the dataset with RUS algorithm can be a proper solution.
- Adopted features are limited to the dataset collected from SSI. However, exploring accidents with more features that define the work environment, hazardous condition, and OHS applications might increase the prediction

performance. Since high performances were already achieved within this thesis, it is expected that OHS departments of construction companies can detect severe accidents with very high accuracies before they occur with additional attributes. In this vein, policy-makers and decision-makers are suggested to collect more explicit features (such as workers PPE use, hazardous condition, risky events, last OHS training of workers, number of OHS staff per worker etc.) from construction companies. Additionally, probably more practicable prediction scheme would be accident vs non-accident classification approach. Because, despite preventing fatal accident is one of the most significant performance indicators of safety management applications, preventing accident occurrence were always desired by safety managers. Therefore, accident vs non-accident classification would be highly interesting and contributing research area. However, no accident data is not recorded in construction projects. Hence, SSI might require construction companies to report a non-occupational accident instance for each recorded occupational accident. In this way, we can observe in what situations occupational accidents occur. We believe that such a prediction model would contribute a lot to the body of knowledge.

- Although distinctive progress has been made in safety management applications over the last decades, construction industry still accounts for a considerable percentage of all workplace fatalities across the world. In this vein, proactive prediction models (whether daily or monthly) may help detect potential risky cases and/or workers and take corresponding preventive measures. Therefore, using prediction models to detect potential accidents can be regarded as a prequalification criterion in public projects (as it is done for BIM utilization in some countries) to improve their safety performance. Particularly in mega projects, where workers and reasonably occupational accidents are higher than other projects, prediction models not only provide OHS professionals with important feedbacks regarding dangerous environment, time period, particular working conditions etc., but also present improvements in national or firm-level safety know-how.
- Time series analysis results show that number of occupational accidents can be predicted based on past accident records. As a future research direction, other national variables that are recorded in a discrete intervals can be used to predict

future accident records. In this way, we can learn which national statistics, indices or variables are more influential in occupational accidents.

- The results also addressed pre-processing steps have significant impacts on the ML model performances. Therefore, despite some particular options were investigated regarding six consecutive preprocessing steps, the practicality of other potential methods such as *RobustScaler* in data scaling, applying 5 bins or 20 bins in data binning step, imputing missing data with ML methods in missing data handling step, and resampling the dataset with other resampling methods (CUS, COS, BL-SMOTE) can be explored with improved methodological background. In this thesis, a novel comprehensive pre-processing approach has been proposed and improvement in this model would contribute a lot to the data science.
- Living in the era of data, it was found that each piece of information is important in terms of safety outcomes, regardless of their quality. Therefore, policy-makers should develop strategies to support construction companies to record and inform all accident instances. Particularly SMEs in developing countries are thought not to record all accident instances. Here, examining the reasons of why some companies do not record safety-related data might be another research area fertile for expansion.
- Material-used, time of the accident, and age of the worker are significant precursors of occupational accidents. Therefore, construction workers should be trained based on specific materials they use and particular time period they work. Additionally, it was well documented in the literature and recognized in the dataset that older workers are more prone to experience severe accidents compared to their younger counterparts. Main reason for this was addressed to be lowering sensory and physical capabilities of older workers. Therefore, periodic safety training of older workers should be more frequent than present.
- In this thesis, four prediction models were presented such as permanent disability, daily accident, severity, and fatal vs non-fatal. However, there are other classification outputs that might be used proactively and confidently. For instance, the most risky month, the most risky time period, and the most risky working days of the workers can be predicted to improve safety management applications. Still,

it is important to mention that in order to develop a prediction model with high accuracy, we need additional and more representative features in this regard.

Overall, this thesis showed the potential of using national dataset in safety management applications in the construction industry. Several data-driven safety management frameworks were proposed to be used in construction sites in different time horizons (daily or monthly). In the era of big data, data-driven safety management applications are expected to increase, and proposed frameworks can be regarded as milestones in effective safety management application. automation in construction. By adopting the proposed methodologies, the worker, accident, and no-accident data can be acquired dynamically, which can present a proactive means to detect future accidents, some of which might be prevented.

REFERENCES

- [1] G. Mistikoglu, I. H. Gerek, E. Erdis, P. E. Mumtaz Usmen, H. Cakan, and E. E. Kazan, "Decision tree analysis of construction fall accidents involving roofers," *Expert Syst. Appl.*, vol. 42, no. 4, pp. 2256–2263, 2015, doi: 10.1016/j.eswa.2014.10.009.
- [2] J. Choi, B. Gu, S. Chin, and J. S. Lee, "Machine learning predictive model based on national data for fatal accidents of construction workers," *Autom. Constr.*, vol. 110, no. September 2019, p. 102974, 2020, doi: 10.1016/j.autcon.2019.102974.
- [3] Y.-H. Chiang, F. K.-W. Wong, and S. Liang, "Fatal Construction Accidents in Hong Kong," *J. Constr. Eng. Manag.*, vol. 144, no. 3, p. 04017121, 2018, doi: 10.1061/(asce)co.1943-7862.0001433.
- [4] Y. M. Goh and C. U. Ubeynarayana, "Construction accident narrative classification: An evaluation of text mining techniques," *Accid. Anal. Prev.*, vol. 108, no. May, pp. 122–130, 2017, doi: 10.1016/j.aap.2017.08.026.
- [5] F. Zhang, H. Fleyeh, X. Wang, and M. Lu, "Construction site accident analysis using text mining and natural language processing techniques," *Autom. Constr.*, vol. 99, pp. 238–248, Mar. 2019, doi: 10.1016/j.autcon.2018.12.016.
- [6] B. U. Ayhan and O. B. Tokdemir, "Accident Analysis for Construction Safety Using Latent Class Clustering and Artificial Neural Networks," *J. Constr. Eng. Manag.*, vol. 146, no. 3, p. 04019114, 2020, doi: 10.1061/(ASCE)CO.1943-7862.0001762.
- [7] A. J. P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, "Application of machine learning to construction injury prediction," *Autom. Constr.*, vol. 69, pp. 102–114, 2016, doi: 10.1016/j.autcon.2016.05.016.
- [8] J. Y. Lee, Y. G. Yoon, T. K. Oh, S. Park, and S. I. Ryu, "A study on data pre-processing and accident prediction modelling for occupational accident analysis in the construction industry," *Appl. Sci.*, vol. 10, no. 21, pp. 1–23, 2020, doi: 10.3390/app10217949.
- [9] K. Yang, C. R. Ahn, M. C. Vuran, and S. S. Aria, "Semi-supervised near-miss fall detection for ironworkers with a wearable inertial measurement unit," *Autom. Constr.*, vol. 68, pp. 194–202, 2016, doi: 10.1016/j.autcon.2016.04.007.
- [10] J. Zhang, L. Zi, Y. Hou, D. Deng, W. Jiang, and M. Wang, "A C-BiLSTM approach to classify construction accident reports," *Appl. Sci.*, vol. 10, no. 17, pp. 1–16, 2020, doi: 10.3390/APP10175754.
- [11] K. Kang and H. Ryu, "Predicting types of occupational accidents at construction sites in Korea using random forest model," *Saf. Sci.*, vol. 120, pp. 226–236, 2019, doi: 10.1016/j.ssci.2019.06.034.
- [12] D. Jahed Armaghani, E. Tonnizam Mohamad, M. Hajihassani, S. V. Alavi Nezhad Khalil Abad, A. Marto, and M. R. Moghaddam, "Evaluation and prediction of flyrock resulting from blasting operations using empirical and computational

- methods,” *Eng. Comput.*, vol. 32, no. 1, pp. 109–121, 2016, doi: 10.1007/s00366-015-0402-5.
- [13] D. Kim, M. Liu, S. H. Lee, and V. R. Kamat, “Remote proximity monitoring between mobile construction resources using camera-mounted UAVs,” *Autom. Constr.*, vol. 99, no. November 2018, pp. 168–182, 2019, doi: 10.1016/j.autcon.2018.12.014.
 - [14] H. Baker, M. R. Hallowell, and A. J.-P. Tixier, “AI-based prediction of independent construction safety outcomes from universal attributes,” *Autom. Constr.*, vol. 118, p. 103146, 2020, doi: 10.1016/j.autcon.2020.103146.
 - [15] J. Choi, B. Gu, S. Chin, and J. S. Lee, “Machine learning predictive model based on national data for fatal accidents of construction workers,” *Autom. Constr.*, vol. 110, no. November 2019, p. 102974, 2020, doi: 10.1016/j.autcon.2019.102974.
 - [16] B. U. Ayhan and O. B. Tokdemir, “Predicting the outcome of construction incidents,” *Saf. Sci.*, vol. 113, no. November 2018, pp. 91–104, 2019, doi: 10.1016/j.ssci.2018.11.001.
 - [17] Occupational Health and Safety Law, “Law No: 6331,” *Resmi Gazete*, 2012. <https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=6331&MevzuatTur=1&MevzuatTertip=5> (accessed Jan. 21, 2021).
 - [18] S. Bilir and G. E. Gürcanlı, “A Method For Determination of Accident Probability in Construction Industry,” *Tek. Dergi*, pp. 8537–8562, 2018, doi: 10.18400/tekderg.363613.
 - [19] Occupational Health and Safety Regulation in Construction Works, “Law No: 6331,” 2013. <https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=18928&MevzuatTur=7&MevzuatTertip=5> (accessed Jan. 21, 2021).
 - [20] International Labor Organization, “Statistics on safety and health at work,” 2020. <https://ilostat.ilo.org/topics/safety-and-health-at-work/>.
 - [21] Social Security Institution (SSI), “Statistics,” 2019. <http://www.sgk.gov.tr/wps/portal/sgk/tr/kurumsal/istatistik/> (accessed Jan. 21, 2021).
 - [22] G. E. Gurcanli, S. Bilir, and M. Sevim, “Activity based risk assessment and safety cost estimation for residential building construction projects,” *Saf. Sci.*, vol. 80, no. 112, pp. 1–12, 2015, doi: 10.1016/j.ssci.2015.07.002.
 - [23] M. Yılmaz and L. O. Uğur, “Comparison of estimated occupational health and safety costs with actual costs in maintenance and repair projects of public buildings,” *Int. J. Occup. Saf. Ergon.*, vol. 27, no. 2, pp. 633–639, 2021, doi: 10.1080/10803548.2019.1660528.
 - [24] M. Uzun, D. Öztürk, and G. E. Gürcanlı, “Mimari Restorasyon ve Konservasyon Projelerinde İşçi Sağlığı ve İş Güvenliği Uygulamaları,” *Tek. Dergi*, pp. 10275–10290, 2020, doi: 10.18400/tekderg.590368.
 - [25] Z. F. Olcay, A. E. Sakalli, S. Temur, and A. Yazici, “A study of the shift in fatal construction work-related accidents during 2012–2019 in Turkey,” *Int. J. Occup. Saf. Ergon.*, pp. 1–11, Apr. 2021, doi: 10.1080/10803548.2021.1900503.
 - [26] H. B. Başağa, B. A. Temel, M. Atasoy, and İ. Yıldırım, “A study on the

- effectiveness of occupational health and safety trainings of construction workers in Turkey,” *Saf. Sci.*, vol. 110, no. September, pp. 344–354, 2018, doi: 10.1016/j.ssci.2018.09.002.
- [27] B. Shao, Z. Hu, Q. Liu, S. Chen, and W. He, “Fatal accident patterns of building construction activities in China,” *Saf. Sci.*, vol. 111, no. June 2018, pp. 253–263, 2019, doi: 10.1016/j.ssci.2018.07.019.
 - [28] C. Q. X. Poh, C. U. Ubeynarayana, and Y. M. Goh, “Safety leading indicators for construction sites: A machine learning approach,” *Autom. Constr.*, vol. 93, no. October 2017, pp. 375–386, 2018, doi: 10.1016/j.autcon.2018.03.022.
 - [29] F. J. Forteza, J. M. Carretero-Gómez, and A. Sesé, “Occupational risks, accidents on sites and economic performance of construction firms,” *Saf. Sci.*, vol. 94, pp. 61–76, 2017, doi: 10.1016/j.ssci.2017.01.003.
 - [30] Y. Feng, S. Zhang, and P. Wu, “Factors influencing workplace accident costs of building projects,” *Saf. Sci.*, vol. 72, pp. 97–104, 2015, doi: 10.1016/j.ssci.2014.08.008.
 - [31] I. Kamardeen and R. Rameezdeen, “Modelling Accident Severity in the Construction Industry,” in *Proc. of the 32nd CIB W78 Conference 2015, 27th-29th October 2015, Eindhoven, The Netherlands*, 2015, no. 1, pp. 384–392.
 - [32] J. Amissah, E. Badu, P. Agyei-Baffour, E. K. Nakua, and I. Mensah, “Predisposing factors influencing occupational injury among frontline building construction workers in Ghana,” *BMC Res. Notes*, vol. 12, no. 1, pp. 8–15, 2019, doi: 10.1186/s13104-019-4744-8.
 - [33] I. F. Mohd Kamar, A. C. Ahmad, M. Mohd Derus, M. Y. Kasiron, and M. A. A. Bakar, “An exploration of the relationship between the safety and health cost dimensions and the accident costs incurred by the employers of urban rail infrastructure projects,” *Malaysian Constr. Res. J.*, vol. 8, no. 3 Special issue, pp. 188–198, 2019.
 - [34] U. B. Celebi and F. Yilmaz, “The Importance of Safety in Construction Sector: Costs of Occupational Accidents in Construction Sites,” *Bus. Econ. Res. J.*, vol. 6, no. January 2016, pp. 25–37, 2015, [Online]. Available: <http://www.berjournal.com/the-importance-of-safety-in-construction-sector-costs-of-occupational-accidents-in-construction-sites-2>.
 - [35] H. O. Adeyemi, S. B. Adejuyigbe, S. O. Ismaila, and A. F. Adekoya, “Low back pain assessment application for construction workers,” *J. Eng. Des. Technol.*, vol. 13, no. 3, pp. 419–434, 2015, doi: 10.1108/JEDT-02-2013-0008.
 - [36] M. F. Antwi-Afari, H. Li, J. O. Seo, and A. Y. L. Wong, “Automated detection and classification of construction workers’ loss of balance events using wearable insole pressure sensors,” *Autom. Constr.*, vol. 96, no. July, pp. 189–199, 2018, doi: 10.1016/j.autcon.2018.09.010.
 - [37] S. Sarkar, R. Raj, S. Vinay, J. Maiti, and D. K. Pratihari, “An optimization-based decision tree approach for predicting slip-trip-fall accidents at work,” *Saf. Sci.*, vol. 118, no. March, pp. 57–69, 2019, doi: 10.1016/j.ssci.2019.05.009.
 - [38] A. Darko, A. P. C. Chan, M. A. Adabre, D. J. Edwards, M. R. Hosseini, and E. E. Ameyaw, “Artificial intelligence in the AEC industry: Scientometric analysis and visualization of research activities,” *Autom. Constr.*, vol. 112, no. May 2019, p.

103081, 2020, doi: 10.1016/j.autcon.2020.103081.

- [39] M. F. Antwi-Afari, H. Li, Y. Yu, and L. Kong, "Wearable insole pressure system for automated detection and classification of awkward working postures in construction workers," *Autom. Constr.*, vol. 96, no. October, pp. 433–441, Dec. 2018, doi: 10.1016/j.autcon.2018.10.004.
- [40] Ö. Ekmekcioğlu, E. E. Başakın, and M. Özger, "Tree-based nonlinear ensemble technique to predict energy dissipation in stepped spillways," *Eur. J. Environ. Civ. Eng.*, vol. 1, no. 1, pp. 1–19, 2020, doi: 10.1080/19648189.2020.1805024.
- [41] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," *J. Anim. Ecol.*, vol. 77, no. 4, pp. 802–813, Jul. 2008, doi: 10.1111/j.1365-2656.2008.01390.x.
- [42] M. Sherafatian, "Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping," *Gene*, vol. 677, pp. 111–118, Nov. 2018, doi: 10.1016/j.gene.2018.07.057.
- [43] M. Omid, "Design of an expert system for sorting pistachio nuts through decision tree and fuzzy logic classifier," *Expert Syst. Appl.*, vol. 38, no. 4, pp. 4339–4347, Apr. 2011, doi: 10.1016/j.eswa.2010.09.103.
- [44] T. Lan, H. Hu, C. Jiang, G. Yang, and Z. Zhao, "A comparative study of decision tree, random forest, and convolutional neural network for spread-F identification," *Adv. Sp. Res.*, vol. 65, no. 8, pp. 2052–2061, Apr. 2020, doi: 10.1016/j.asr.2020.01.036.
- [45] M.-C. Kang, D.-Y. Yoo, and R. Gupta, "Machine learning-based prediction for compressive and flexural strengths of steel fiber-reinforced concrete," *Constr. Build. Mater.*, vol. 266, p. 121117, Jan. 2021, doi: 10.1016/j.conbuildmat.2020.121117.
- [46] P. Zhang, R.-P. Chen, and H.-N. Wu, "Real-time analysis and regulation of EPB shield steering using Random Forest," *Autom. Constr.*, vol. 106, p. 102860, Oct. 2019, doi: 10.1016/j.autcon.2019.102860.
- [47] X. Zhao, R. Lovreglio, and D. Nilsson, "Modelling and interpreting pre-evacuation decision-making using machine learning," *Autom. Constr.*, vol. 113, p. 103140, May 2020, doi: 10.1016/j.autcon.2020.103140.
- [48] S. Bai, M. Li, R. Kong, S. Han, H. Li, and L. Qin, "Data mining approach to construction productivity prediction for cutter suction dredgers," *Autom. Constr.*, vol. 105, p. 102833, Sep. 2019, doi: 10.1016/j.autcon.2019.102833.
- [49] M. M. Ghiasi and S. Zendehboudi, "Application of decision tree-based ensemble learning in the classification of breast cancer," *Comput. Biol. Med.*, vol. 128, p. 104089, Jan. 2021, doi: 10.1016/j.compbiomed.2020.104089.
- [50] U. Saeed, S. U. Jan, Y.-D. Lee, and I. Koo, "Fault diagnosis based on extremely randomized trees in wireless sensor networks," *Reliab. Eng. Syst. Saf.*, vol. 205, p. 107284, Jan. 2021, doi: 10.1016/j.ress.2020.107284.
- [51] M. W. Ahmad, J. Reynolds, and Y. Rezgui, "Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees," *J. Clean. Prod.*, vol. 203, pp. 810–821, Dec. 2018, doi: 10.1016/j.jclepro.2018.08.207.

- [52] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.
- [53] F. Lv, J. Wang, B. Cui, J. Yu, J. Sun, and J. Zhang, "An improved extreme gradient boosting approach to vehicle speed prediction for construction simulation of earthwork," *Autom. Constr.*, vol. 119, no. June, p. 103351, 2020, doi: 10.1016/j.autcon.2020.103351.
- [54] W. Dong, Y. Huang, B. Lehane, and G. Ma, "XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring," *Autom. Constr.*, vol. 114, p. 103155, Jun. 2020, doi: 10.1016/j.autcon.2020.103155.
- [55] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.*, vol. 374, no. 2065, p. 20150202, Apr. 2016, doi: 10.1098/rsta.2015.0202.
- [56] A. J.-P. Tixier, M. R. Hallowell, A. Albert, L. van Boven, and B. M. Kleiner, "Psychological Antecedents of Risk-Taking Behavior in Construction," *J. Constr. Eng. Manag.*, vol. 140, no. 11, p. 04014052, Nov. 2014, doi: 10.1061/(ASCE)CO.1943-7862.0000894.
- [57] X. Zhang, S. Huang, S. Yang, R. Tu, and L. Jin, "Safety Assessment in Road Construction Work System Based on Group AHP-PCA," *Math. Probl. Eng.*, vol. 2020, pp. 1–12, Mar. 2020, doi: 10.1155/2020/6210569.
- [58] W. T. Chen *et al.*, "Construction Safety Success Factors: A Taiwanese Case Study," *Sustainability*, vol. 12, no. 16, p. 6326, Aug. 2020, doi: 10.3390/su12166326.
- [59] W. Li, H. Zhu, L. Ding, and H. Luo, "Structural damage recognition by grouped data based on Principal Component Analysis theory," *Autom. Constr.*, vol. 22, pp. 258–270, Mar. 2012, doi: 10.1016/j.autcon.2011.09.001.
- [60] M. Özger, E. E. Başakın, Ö. Ekmekcioğlu, and V. Hacısüleyman, "Comparison of wavelet and empirical mode decomposition hybrid models in drought prediction," *Comput. Electron. Agric.*, vol. 179, no. May, p. 105851, Dec. 2020, doi: 10.1016/j.compag.2020.105851.
- [61] H. Son, C. Kim, and C. Kim, "Hybrid principal component analysis and support vector machine model for predicting the cost performance of commercial building projects using pre-project planning variables," *Autom. Constr.*, vol. 27, pp. 60–66, Nov. 2012, doi: 10.1016/j.autcon.2012.05.013.
- [62] F. ul Hassan and T. Le, "Computer-assisted separation of design-build contract requirements to support subcontract drafting," *Autom. Constr.*, vol. 122, no. November 2020, p. 103479, 2021, doi: 10.1016/j.autcon.2020.103479.
- [63] B. U. Ayhan and O. B. Tokdemir, "Predicting the outcome of construction incidents," *Saf. Sci.*, vol. 113, no. November 2018, pp. 91–104, 2019, doi: 10.1016/j.ssci.2018.11.001.
- [64] Y. Jang, I. B. Jeong, Y. K. Cho, and Y. Ahn, "Predicting Business Failure of Construction Contractors Using Long Short-Term Memory Recurrent Neural Network," *J. Constr. Eng. Manag.*, vol. 145, no. 11, pp. 1–9, 2019, doi: 10.1061/(ASCE)CO.1943-7862.0001709.
- [65] S. Sarkar, S. Vinay, R. Raj, J. Maiti, and P. Mitra, "Application of optimized

- machine learning techniques for prediction of occupational accidents,” *Comput. Oper. Res.*, vol. 106, pp. 210–224, Jun. 2019, doi: 10.1016/j.cor.2018.02.021.
- [66] H. Baker, M. R. Hallowell, and A. J. Tixier, “AI-based prediction of independent construction safety outcomes from universal attributes,” *Autom. Constr.*, vol. 118, no. June, p. 103146, 2020, doi: 10.1016/j.autcon.2020.103146.
 - [67] S. Yuanyuan, W. Yongming, G. Lili, M. Zhongsong, and J. Shan, “The comparison of optimizing SVM by GA and grid search,” in *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, Oct. 2017, pp. 354–360, doi: 10.1109/ICEMI.2017.8265815.
 - [68] F.-S. Wang and L.-H. Chen, “Heuristic Optimization,” in *Encyclopedia of Systems Biology*, New York, NY: Springer New York, 2013, pp. 885–885.
 - [69] M. Nazari-Heris, B. Mohammadi-Ivatloo, and G. B. Gharehpetian, “A comprehensive review of heuristic optimization algorithms for optimal combined heat and power dispatch from economic and environmental perspectives,” *Renew. Sustain. Energy Rev.*, vol. 81, pp. 2128–2143, Jan. 2018, doi: 10.1016/j.rser.2017.06.024.
 - [70] A. E. Ezugwu, O. J. Adeleke, A. A. Akinyelu, and S. Viriri, *A conceptual comparison of several metaheuristic algorithms on continuous optimisation problems*, vol. 32, no. 10. Springer London, 2020.
 - [71] R. Khosravanian, V. Mansouri, D. A. Wood, and M. R. Alipour, “A comparative study of several metaheuristic algorithms for optimizing complex 3-D well-path designs,” *J. Pet. Explor. Prod. Technol.*, vol. 8, no. 4, pp. 1487–1503, Dec. 2018, doi: 10.1007/s13202-018-0447-2.
 - [72] Z. Tong, “A genetic algorithm approach to optimizing the distribution of buildings in urban green space,” *Autom. Constr.*, vol. 72, pp. 46–51, Dec. 2016, doi: 10.1016/j.autcon.2016.10.001.
 - [73] H. Song, J. Ghaboussi, and T.-H. Kwon, “Architectural design of apartment buildings using the Implicit Redundant Representation Genetic Algorithm,” *Autom. Constr.*, vol. 72, pp. 166–173, Dec. 2016, doi: 10.1016/j.autcon.2016.09.001.
 - [74] M. H. Nili, H. Taghaddos, and B. Zahraie, “Integrating discrete event simulation and genetic algorithm optimization for bridge maintenance planning,” *Autom. Constr.*, vol. 122, p. 103513, Feb. 2021, doi: 10.1016/j.autcon.2020.103513.
 - [75] Y. Wang, G. Cui, and J. Xu, “Semi-automatic detection of buried rebar in GPR data using a genetic algorithm,” *Autom. Constr.*, vol. 114, p. 103186, Jun. 2020, doi: 10.1016/j.autcon.2020.103186.
 - [76] S. Taфраout, N. Bourahla, Y. Bourahla, and A. Mebarki, “Automatic structural design of RC wall-slab buildings using a genetic algorithm with application in BIM environment,” *Autom. Constr.*, vol. 106, p. 102901, Oct. 2019, doi: 10.1016/j.autcon.2019.102901.
 - [77] M. Mangal and J. C. P. Cheng, “Automated optimization of steel reinforcement in RC building frames using building information modeling and hybrid genetic algorithm,” *Autom. Constr.*, vol. 90, pp. 39–57, Jun. 2018, doi: 10.1016/j.autcon.2018.01.013.
 - [78] F. Yilmaz, “Monitoring and analysis of construction site accidents by using

- accidents analysis management system in Turkey,” *J. Sustain. Dev.*, vol. 8, no. 2, pp. 57–65, 2015, doi: 10.5539/jsd.v8n2p57.
- [79] J. A. Hess, L. D. Kincl, and K. Davis, “The impact of drywall handling tools on the low back,” *Appl. Ergon.*, vol. 41, no. 2, pp. 305–312, 2010, doi: 10.1016/j.apergo.2009.08.001.
 - [80] I. F. Mohd Kamar, A. C. Ahmad, M. Mohd Derus, M. Y. Kasiron, and M. A. A. Bakar, “Exploring the relationship between the safety and health cost dimensions and the accident costs incurred by the employers of urban rail infrastructure projects,” in *MATEC Web of Conferences*, 2019, vol. 266, p. 03014, doi: 10.1051/matecconf/201926626603014.
 - [81] P. Xue, Y. Jiang, Z. Zhou, X. Chen, X. Fang, and J. Liu, “Multi-step ahead forecasting of heat load in district heating systems using machine learning algorithms,” *Energy*, vol. 188, p. 116085, 2019, doi: 10.1016/j.energy.2019.116085.
 - [82] I. H. Sarker, H. Alqahtani, F. Alsolami, A. I. Khan, Y. B. Abushark, and M. K. Siddiqui, “Context pre-modeling: an empirical analysis for classification based user-centric context-aware predictive modeling,” *J. Big Data*, vol. 7, no. 1, pp. 1–23, 2020, doi: 10.1186/s40537-020-00328-3.
 - [83] I. H. Sarker, Y. B. Abushark, and A. I. Khan, “ContextPCA: Predicting context-aware smartphone apps usage based on machine learning techniques,” *Symmetry (Basel)*, vol. 12, no. 4, pp. 1–14, 2020, doi: 10.3390/SYM12040499.
 - [84] R. M. Swarna Priya *et al.*, “An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture,” *Comput. Commun.*, vol. 160, no. May, pp. 139–149, 2020, doi: 10.1016/j.comcom.2020.05.048.
 - [85] G. Kaur, “A comparison of two hybrid ensemble techniques for network anomaly detection in spark distributed environment,” *J. Inf. Secur. Appl.*, vol. 55, no. September, p. 102601, 2020, doi: 10.1016/j.jisa.2020.102601.
 - [86] M. Luo *et al.*, “Comparing machine learning algorithms in predicting thermal sensation using ASHRAE Comfort Database II,” *Energy Build.*, vol. 210, p. 109776, 2020, doi: 10.1016/j.enbuild.2020.109776.
 - [87] S. Bhattacharya *et al.*, “A Novel PCA-Firefly Based XGBoost Classification Model for Intrusion Detection in Networks Using GPU,” *Electronics*, vol. 9, no. 2, p. 219, Jan. 2020, doi: 10.3390/electronics9020219.
 - [88] G. Thippa Reddy *et al.*, “A deep neural networks based model for uninterrupted marine environment monitoring,” *Comput. Commun.*, vol. 157, no. January, pp. 64–75, 2020, doi: 10.1016/j.comcom.2020.04.004.
 - [89] T. R. Gadekallu *et al.*, “Early detection of diabetic retinopathy using pca-firefly based deep learning model,” *Electron.*, vol. 9, no. 2, pp. 1–16, 2020, doi: 10.3390/electronics9020274.
 - [90] H. F. Kaiser, “The application of electronic computers to factor analysis,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 141–151, 1960.
 - [91] S. R. Narang, M. K. Jindal, S. Ahuja, and M. Kumar, “On the recognition of Devanagari ancient handwritten characters using SIFT and Gabor features,” *Soft Comput.*, vol. 24, no. 22, pp. 17279–17289, 2020, doi: 10.1007/s00500-020-05018-z.

- [92] Y. Peng, C. Li, K. Wang, Z. Gao, and R. Yu, "Examining imbalanced classification algorithms in predicting real-time traffic crash risk," *Accid. Anal. Prev.*, vol. 144, no. March, p. 105610, Sep. 2020, doi: 10.1016/j.aap.2020.105610.
- [93] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique Nitesh," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002, doi: 10.1002/eap.2043.
- [94] X. W. Liang, A. P. Jiang, T. Li, Y. Y. Xue, and G. T. Wang, "LR-SMOTE — An improved unbalanced data set oversampling based on K-means and SVM," *Knowledge-Based Syst.*, vol. 196, p. 105845, 2020, doi: 10.1016/j.knosys.2020.105845.
- [95] Q. Wang, Z. Luo, J. Huang, Y. Feng, and Z. Liu, "A Novel Ensemble Method for Imbalanced Data Learning: Bagging of Extrapolation-SMOTE SVM," *Comput. Intell. Neurosci.*, vol. 1, pp. 1–11, 2017, doi: 10.1155/2017/1827016.
- [96] Z. Allah Bukhsh, I. Stipanovic, A. Saeed, and A. G. Doree, "Maintenance intervention predictions using entity-embedding neural networks," *Autom. Constr.*, vol. 116, no. March, p. 103202, 2020, doi: 10.1016/j.autcon.2020.103202.
- [97] M. Y. Cheng, D. Kusoemo, and R. A. Gosno, "Text mining-based construction site accident classification using hybrid supervised machine learning," *Autom. Constr.*, vol. 118, no. November 2019, p. 103265, 2020, doi: 10.1016/j.autcon.2020.103265.
- [98] C. M. Yeşilkanat, "Spatio-temporal estimation of the daily cases of COVID-19 in worldwide using random forest machine learning algorithm," *Chaos, Solitons & Fractals*, vol. 140, p. 110210, Nov. 2020, doi: 10.1016/j.chaos.2020.110210.
- [99] P. Probst, M. N. Wright, and A. Boulesteix, "Hyperparameters and tuning strategies for random forest," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 3, May 2019, doi: 10.1002/widm.1301.
- [100] Z. Khan *et al.*, "Ensemble of optimal trees, random forest and random projection ensemble classification," *Adv. Data Anal. Classif.*, vol. 14, no. 1, pp. 97–116, Mar. 2020, doi: 10.1007/s11634-019-00364-9.
- [101] Y.-H. Cheng, C.-M. Lai, and J. Teh, "Genetic algorithm with small population size for search feasible control parameters for parallel hybrid electric vehicles," *AIMS Energy*, vol. 5, no. 6, pp. 930–943, 2017, doi: 10.3934/energy.2017.6.930.
- [102] A. Hassanat, K. Almohammadi, E. Alkafaween, E. Abunawas, A. Hammouri, and V. B. S. Prasath, "Choosing Mutation and Crossover Ratios for Genetic Algorithms—A Review with a New Dynamic Approach," *Information*, vol. 10, no. 12, p. 390, Dec. 2019, doi: 10.3390/info10120390.
- [103] M. Buckland and F. Gey, "The relationship between Recall and Precision," *J. Am. Soc. Inf. Sci.*, vol. 45, no. 1, pp. 12–19, 1994, doi: 10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L.
- [104] R. Prabowo and M. Thelwall, "Sentiment analysis: A combined approach," *J. Informetr.*, vol. 3, no. 2, pp. 143–157, 2009, doi: 10.1016/j.joi.2009.01.003.
- [105] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors (Switzerland)*, vol. 15, no. 12, pp. 31314–31338, 2015, doi: 10.3390/s151229858.

- [106] S. Sarkar and J. Maiti, "Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis," *Saf. Sci.*, vol. 131, no. June, p. 104900, 2020, doi: 10.1016/j.ssci.2020.104900.
- [107] Y. Dai, J. Tian, H. Rong, and T. Zhao, "Hybrid safety analysis method based on SVM and RST: An application to carrier landing of aircraft," *Saf. Sci.*, vol. 80, pp. 56–65, 2015, doi: 10.1016/j.ssci.2015.07.006.
- [108] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in *International Symposium on Intelligence Computation and Applications, ISICA 2009, Communications in Computer and Information Science*, 2009, pp. 461–471, doi: 10.1007/978-3-642-04962-0_53.
- [109] N. W. Chi, J. P. Wang, J. H. Liao, W. C. Cheng, and C. S. Chen, "Machine learning-based seismic capability evaluation for school buildings," *Autom. Constr.*, vol. 118, no. March, p. 103274, 2020, doi: 10.1016/j.autcon.2020.103274.
- [110] T. Hasanin, T. M. Khoshgoftaar, J. L. Leevy, and R. A. Bauder, "Severely imbalanced Big Data challenges: investigating data sampling approaches," *J. Big Data*, vol. 6, no. 1, pp. 1–25, 2019, doi: 10.1186/s40537-019-0274-4.
- [111] I. W. H. Fung, V. W. Y. Tam, T. Y. Lo, and L. L. H. Lu, "Developing a risk assessment model for construction safety," *Int. J. Proj. Manag.*, vol. 28, no. 6, pp. 593–600, 2010, doi: 10.1016/j.ijproman.2009.09.006.
- [112] L. Sun, O. Paez, D. Lee, S. Salem, and N. M. Daraiseh, "Estimating the uninsured costs of work-related accidents, part I: A systematic review," *Theor. Issues Ergon. Sci.*, vol. 7, no. 3, pp. 227–245, 2006, doi: 10.1080/14639220500090521.
- [113] M. Jahangiri, H. R. J. Solukloei, and M. Kamalinia, "A neuro-fuzzy risk prediction methodology for falling from scaffold," *Saf. Sci.*, vol. 117, pp. 88–99, 2019, doi: 10.1016/j.ssci.2019.04.009.
- [114] O. N. Aneziris *et al.*, "Occupational risk quantification owing to falling objects," *Saf. Sci.*, vol. 69, pp. 57–70, 2014, doi: 10.1016/j.ssci.2014.02.017.
- [115] X. Yan, H. Li, C. Wang, J. O. Seo, H. Zhang, and H. Wang, "Development of ergonomic posture recognition technique based on 2D ordinary camera for construction hazard prevention through view-invariant features in 2D skeleton motion," *Adv. Eng. Informatics*, vol. 34, no. October, pp. 152–163, 2017, doi: 10.1016/j.aei.2017.11.001.
- [116] G. T. Çelîk and C. D. Temel, "Reward-Penalty Applications in the Use of Personal Protective Equipment in Construction Sector," *Çukurova Univ. J. Fac. Eng. Archit.*, vol. 33, no. 4, pp. 197–202, 2018.
- [117] P. Kumar *et al.*, "Ergonomics Methods to Improve Safety in Construction Industry," *Int. Res. J. Eng. Technol.*, vol. 3, no. 8, pp. 680–683, 2016.
- [118] K. Robert, Q. Elisabeth, and B. Josef, "Analysis of occupational accidents with agricultural machinery in the period 2008–2010 in Austria," *Saf. Sci.*, vol. 72, pp. 319–328, Feb. 2015, doi: 10.1016/j.ssci.2014.10.004.
- [119] S. Chi and S. Han, "Analyses of systems theory for construction accident prevention with specific reference to OSHA accident reports," *Int. J. Proj. Manag.*, vol. 31, no. 7, pp. 1027–1041, 2013, doi: 10.1016/j.ijproman.2012.12.004.
- [120] N. Erginel and Ş. Toptancı, "İş Kazası Verilerinin Olasılık Dağılımları İle

- Modellenmesi,” *Mühendislik Bilim. ve Tasarım Derg.*, vol. 5, no. 1, pp. 201–212, 2017, doi: 10.21923/jesd.10116.
- [121] N. D. Nath, A. H. Behzadan, and S. G. Paal, “Deep learning for site safety: Real-time detection of personal protective equipment,” *Autom. Constr.*, vol. 112, no. July 2019, p. 103085, 2020, doi: 10.1016/j.autcon.2020.103085.
- [122] H. Zahoor, A. P. C. Chan, R. Masood, R. M. Choudhry, A. A. Javed, and W. P. Utama, “Occupational safety and health performance in the Pakistani construction industry: stakeholders’ perspective,” *Int. J. Constr. Manag.*, vol. 16, no. 3, pp. 209–219, 2016, doi: 10.1080/15623599.2015.1138027.
- [123] Working Power and Professional Gaining Power Loss Rate Determination Procedures Regulation, “Law no: 5510,” 2008. [Online]. Available: <https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=12511&MevzuatTur=7&MevzuatTertip=5>.
- [124] Social Insurance and General Health Insurance Law, “Law No: 5510,” 2006. [Online]. Available: <https://www.resmigazete.gov.tr/eskiler/2006/06/20060616-1.htm>.
- [125] R. Awwad, O. El Souki, and M. Jabbour, “Construction safety practices and challenges in a Middle Eastern developing country,” *Saf. Sci.*, vol. 83, pp. 1–11, 2016, doi: 10.1016/j.ssci.2015.10.016.
- [126] R. M. Choudhry, D. Fang, and S. Rowlinson, “Challenging and Enforcing Safety Management in Developing Countries: A Strategy,” *Int. J. Constr. Manag.*, vol. 8, no. 1, pp. 87–101, 2008, doi: 10.1080/15623599.2008.10773110.
- [127] C. Nnaji, J. Gambatese, A. Karakhan, and R. Osei-Kyei, “Development and Application of Safety Technology Adoption Decision-Making Tool,” *J. Constr. Eng. Manag.*, vol. 146, no. 4, p. 04020028, 2020, doi: 10.1061/(ASCE)CO.1943-7862.0001808.
- [128] Turkish Statistical Institute (TSI), “Employment statistics by sector,” 2021. <https://data.tuik.gov.tr/Bulten/Index?p=Isgucu-Istatistikleri-Ekim-2021-37491> (accessed Jan. 21, 2021).
- [129] Ministry of family labor and social services (MFLSS), “The number of safety professionals,” 2020. <https://www.ailevecalisma.gov.tr/tr-tr/haberler/bakan-selcuk-is-guvenligi-uzmani-sayimiz-165-bini-asti/#:~:text=Aile%2C> (accessed Jan. 21, 2021).
- [130] P. K. Marhavilas, D. E. Koulouriotis, and S. H. Spartalis, “Harmonic analysis of occupational-accident time-series as a part of the quantified risk evaluation in worksites: Application on electric power industry and construction sector,” *Reliab. Eng. Syst. Saf.*, vol. 112, pp. 8–25, 2013, doi: 10.1016/j.ress.2012.11.014.
- [131] R. G. Kavasseri and K. Seetharaman, “Day-ahead wind speed forecasting using f-ARIMA models,” *Renew. Energy*, vol. 34, no. 5, pp. 1388–1393, 2009, doi: 10.1016/j.renene.2008.09.006.
- [132] P. H. Kuo and C. J. Huang, “A high precision artificial neural networks model for short-Term energy load forecasting,” *Energies*, vol. 11, no. 1, pp. 1–13, 2018, doi: 10.3390/en11010213.
- [133] L. Liang, L. Li, C. Liu, and L. Cuo, “Climate change in the Tibetan plateau three rivers source region: 1960-2009,” *Int. J. Climatol.*, vol. 33, no. 13, pp. 2900–2916,

2013, doi: 10.1002/joc.3642.

- [134] H. Adeli and X. Jiang, “Dynamic Fuzzy Wavelet Neural Network for Structural System Identification,” *Intell. Infrastruct.*, vol. 132, no. January, pp. 271–296, 2020, doi: 10.1201/9781482281767-20.
- [135] W. Chen, Z. Shang, and Y. Chen, “A Novel Hybrid Network Traffic Prediction Approach Based on Support Vector Machines,” *J. Comput. Networks Commun.*, vol. 1, pp. 1–10, Feb. 2019, doi: 10.1155/2019/2182803.
- [136] L. Ding, L. Ma, H. Luo, M. Yu, and X. Wu, “Wavelet Analysis for tunneling-induced ground settlement based on a stochastic model,” *Tunn. Undergr. Sp. Technol.*, vol. 26, no. 5, pp. 619–628, Sep. 2011, doi: 10.1016/j.tust.2011.03.005.
- [137] X. He, J. Luo, P. Li, G. Zuo, and J. Xie, “A Hybrid Model Based on Variational Mode Decomposition and Gradient Boosting Regression Tree for Monthly Runoff Forecasting,” *Water Resour. Manag.*, vol. 34, no. 2, pp. 865–884, Jan. 2020, doi: 10.1007/s11269-020-02483-x.
- [138] J. Chen, X. Song, and Z. Lin, “Revealing the ‘Invisible Gorilla’ in construction: Estimating construction safety through mental workload assessment,” *Autom. Constr.*, vol. 63, pp. 173–183, Mar. 2016, doi: 10.1016/j.autcon.2015.12.018.
- [139] C. Zhou, L. Ding, M. J. Skibniewski, H. Luo, and S. Jiang, “Characterizing time series of near-miss accidents in metro construction via complex network theory,” *Saf. Sci.*, vol. 98, pp. 145–158, 2017, doi: 10.1016/j.ssci.2017.06.012.
- [140] M. Özger, A. K. Mishra, and V. P. Singh, “Low frequency drought variability associated with climate indices,” *J. Hydrol.*, vol. 364, no. 1–2, pp. 152–162, Jan. 2009, doi: 10.1016/j.jhydrol.2008.10.018.
- [141] N. K. Al-Qazzaz, S. H. Bin Mohd Ali, S. A. Ahmad, M. S. Islam, and J. Escudero, “Selection of mother wavelet functions for multi-channel EEG signal analysis during a working memory task,” *Sensors (Switzerland)*, vol. 15, no. 11, pp. 29015–29035, 2015, doi: 10.3390/s151129015.
- [142] S. Samadianfard *et al.*, “Wavelet neural networks and gene expression programming models to predict short-term soil temperature at different depths,” *Soil Tillage Res.*, vol. 175, pp. 37–50, Jan. 2018, doi: 10.1016/j.still.2017.08.012.
- [143] A. Altunkaynak and M. Ozger, “Comparison of Discrete and Continuous Wavelet–Multilayer Perceptron Methods for Daily Precipitation Prediction,” *J. Hydrol. Eng.*, vol. 21, no. 7, p. 04016014, Jul. 2016, doi: 10.1061/(ASCE)HE.1943-5584.0001376.
- [144] S. G. Mallat, “A theory for multiresolution signal decomposition: the wavelet representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989, doi: 10.1109/34.192463.
- [145] Y. Meng, S. N. Qasem, M. Shokri, and S. S., “Dimension Reduction of Machine Learning-Based Forecasting Models Employing Principal Component Analysis,” *Mathematics*, vol. 8, no. 8, p. 1233, Jul. 2020, doi: 10.3390/math8081233.
- [146] N. Arabekar, W. Derguech, E. Burke, and E. Curry, “Autonomic Source Selection for Real-time Predictive Analytics Using the Internet of Things and Open Data,” in *Real-time Linked Dataspaces*, Cham: Springer International Publishing, 2020, pp. 237–253.

- [147] C.-J. Lu, T.-S. Lee, and C.-M. Lian, "Sales forecasting for computer wholesalers: A comparison of multivariate adaptive regression splines and artificial neural networks," *Decis. Support Syst.*, vol. 54, no. 1, pp. 584–596, Dec. 2012, doi: 10.1016/j.dss.2012.08.006.
- [148] W. Zhang and A. T. C. Goh, "Multivariate adaptive regression splines and neural network models for prediction of pile drivability," *Geosci. Front.*, vol. 7, no. 1, pp. 45–52, Jan. 2016, doi: 10.1016/j.gsf.2014.10.003.
- [149] L. E. Juarez-Orozco, O. Martinez-Manzanera, S. V. Nesterov, S. Kajander, and J. Knuuti, "The machine learning horizon in cardiac hybrid imaging," *Eur. J. Hybrid Imaging*, vol. 2, no. 1, pp. 1–15, Dec. 2018, doi: 10.1186/s41824-018-0033-3.
- [150] F. Mirahadi and T. Zayed, "Simulation-based construction productivity forecast using Neural-Network-Driven Fuzzy Reasoning," *Autom. Constr.*, vol. 65, pp. 102–115, 2016, doi: 10.1016/j.autcon.2015.12.021.
- [151] J. Reifman and E. E. Feldman, "Multilayer perceptron for nonlinear programming," *Comput. Oper. Res.*, vol. 29, no. 9, pp. 1237–1250, Aug. 2002, doi: 10.1016/S0305-0548(01)00027-2.
- [152] J. S. Chou, C. W. Lin, A. D. Pham, and J. Y. Shao, "Optimized artificial intelligence models for predicting project award price," *Autom. Constr.*, vol. 54, pp. 106–115, 2015, doi: 10.1016/j.autcon.2015.02.006.
- [153] W. Yi, A. P. C. Chan, X. Wang, and J. Wang, "Development of an early-warning system for site work in hot and humid environments: A case study," *Autom. Constr.*, vol. 62, pp. 101–113, 2016, doi: 10.1016/j.autcon.2015.11.003.
- [154] D. Chakraborty, H. Elhegazy, H. Elzarka, and L. Gutierrez, "A novel construction cost prediction model using hybrid natural and light gradient boosting," *Adv. Eng. Informatics*, vol. 46, no. October, p. 101201, 2020, doi: 10.1016/j.aei.2020.101201.
- [155] K. W. Chau, "Application of a PSO-based neural network in analysis of outcomes of construction claims," *Autom. Constr.*, vol. 16, no. 5, pp. 642–646, 2007, doi: 10.1016/j.autcon.2006.11.008.
- [156] J. H. Friedman, "Multivariate Adaptive Regression Splines," *Ann. Stat.*, vol. 19, no. 1, pp. 1–67, 1991.
- [157] D. Wheeler and M. Tiefelsdorf, "Multicollinearity and correlation among local regression coefficients in geographically weighted regression," *J. Geogr. Syst.*, vol. 7, no. 2, pp. 161–187, Jun. 2005, doi: 10.1007/s10109-005-0155-6.
- [158] R. M. Adnan, Z. Liang, S. Heddami, M. Zounemat-Kermani, O. Kisi, and B. Li, "Least square support vector machine and multivariate adaptive regression splines for streamflow prediction in mountainous basin using hydro-meteorological data as inputs," *J. Hydrol.*, vol. 586, p. 124371, Jul. 2020, doi: 10.1016/j.jhydrol.2019.124371.
- [159] M.-T. Cao, M.-Y. Cheng, and Y.-W. Wu, "Hybrid Computational Model for Forecasting Taiwan Construction Cost Index," *J. Constr. Eng. Manag.*, vol. 141, no. 4, p. 04014089, 2015, doi: 10.1061/(asce)co.1943-7862.0000948.
- [160] X. Chen and H. Yang, "A multi-stage optimization of passively designed high-rise residential buildings in multiple building operation scenarios," *Appl. Energy*, vol. 206, no. June, pp. 541–557, 2017, doi: 10.1016/j.apenergy.2017.08.204.

- [161] D. Guo, H. Chen, L. Tang, Z. Chen, and P. Samui, "Assessment of rockburst risk using multivariate adaptive regression splines and deep forest model," *Acta Geotech.*, vol. 2, pp. 1–23, Jul. 2021, doi: 10.1007/s11440-021-01299-2.
- [162] A. Suárez Sánchez, P. Riesgo Fernández, F. Sánchez Lasheras, F. J. De Cos Juez, and P. J. García Nieto, "Prediction of work-related accidents according to working conditions using support vector machines," *Appl. Math. Comput.*, vol. 218, no. 7, pp. 3539–3552, 2011, doi: 10.1016/j.amc.2011.08.100.
- [163] C. Cortes and V. Vapnik, "Support Vector Networks," *Mach. Learn.*, vol. 20, pp. 293–297, 1995, doi: 10.1109/64.163674.
- [164] A. Mahmoodzadeh, M. Mohammadi, H. Hashim Ibrahim, K. M. Gharrib Noori, S. Nariman Abdulhamid, and H. Farid Hama Ali, "Forecasting sidewall displacement of underground caverns using machine learning techniques," *Autom. Constr.*, vol. 123, no. November 2020, p. 103530, 2021, doi: 10.1016/j.autcon.2020.103530.
- [165] Y. Pan and L. Zhang, "BIM log mining: Learning and predicting design commands," *Autom. Constr.*, vol. 112, no. January, p. 103107, Apr. 2020, doi: 10.1016/j.autcon.2020.103107.
- [166] Z. Ye and H. Li, "Based on Radial Basis Kernel function of Support Vector Machines for speaker recognition," *2012 5th Int. Congr. Image Signal Process. CISP 2012*, no. Cisp, pp. 1584–1587, 2012, doi: 10.1109/CISP.2012.6469807.
- [167] H. Yu, T. Xie, S. Paszczyński, and B. M. Wilamowski, "Advantages of radial basis function networks for dynamic system design," *IEEE Trans. Ind. Electron.*, vol. 58, no. 12, pp. 5438–5450, 2011, doi: 10.1109/TIE.2011.2164773.
- [168] F. Zhang, "A hybrid structured deep neural network with Word2Vec for construction accident causes classification," *Int. J. Constr. Manag.*, vol. 1, no. 0, pp. 1–21, Nov. 2019, doi: 10.1080/15623599.2019.1683692.
- [169] J. Li *et al.*, "Identification and classification of construction equipment operators' mental fatigue using wearable eye-tracking technology," *Autom. Constr.*, vol. 109, no. October 2019, p. 103000, 2020, doi: 10.1016/j.autcon.2019.103000.
- [170] F. Zhang, H. Fleyeh, X. Wang, and M. Lu, "Construction site accident analysis using text mining and natural language processing techniques," *Autom. Constr.*, vol. 99, no. December 2018, pp. 238–248, 2019, doi: 10.1016/j.autcon.2018.12.016.
- [171] H. Ghanadzadeh, M. Ganji, and S. Fallahi, "Mathematical model of liquid–liquid equilibrium for a ternary system using the GMDH-type neural network and genetic algorithm," *Appl. Math. Model.*, vol. 36, no. 9, pp. 4096–4105, Sep. 2012, doi: 10.1016/j.apm.2011.11.039.
- [172] S. Elkatatny, M. Mahmoud, Z. Tariq, and A. Abdulraheem, "New insights into the prediction of heterogeneous carbonate reservoir permeability from well logs using artificial intelligence network," *Neural Comput. Appl.*, vol. 30, no. 9, pp. 2673–2683, Nov. 2018, doi: 10.1007/s00521-017-2850-x.
- [173] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I — A discussion of principles," *J. Hydrol.*, vol. 10, no. 3, pp. 282–290, Apr. 1970, doi: 10.1016/0022-1694(70)90255-6.
- [174] D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith, "Model Evaluation Guidelines for Systematic Quantification of

- Accuracy in Watershed Simulations,” *Trans. ASABE*, vol. 50, no. 3, pp. 885–900, 2007, doi: 10.13031/2013.23153.
- [175] P. Zhang, N. Li, Z. Jiang, D. Fang, and C. J. Anumba, “An agent-based modeling approach for understanding the effect of worker-management interactions on construction workers’ safety-related behaviors,” *Autom. Constr.*, vol. 97, no. September 2018, pp. 29–43, 2019, doi: 10.1016/j.autcon.2018.10.015.
- [176] E. M. Wetzel and W. Y. Thabet, “The use of a BIM-based framework to support safe facility management processes,” *Autom. Constr.*, vol. 60, pp. 12–24, 2015, doi: 10.1016/j.autcon.2015.09.004.
- [177] I. Maslova, A. M. Ticlavilca, and M. Mckee, “Adjusting wavelet-based multiresolution analysis boundary conditions for long-term streamflow forecasting,” *Hydrol. Process.*, vol. 30, no. 1, pp. 57–74, 2016, doi: 10.1002/hyp.10564.
- [178] B. B. Hazarika and D. Gupta, “Modelling and forecasting of COVID-19 spread using wavelet-coupled random vector functional link networks,” *Appl. Soft Comput. J.*, vol. 96, p. 106626, 2020, doi: 10.1016/j.asoc.2020.106626.
- [179] Labor Law, “Law No: 4857,” 2003. <https://www.mevzuat.gov.tr/mevzuat?MevzuatNo=4857&MevzuatTur=1&MevzuatTertip=5> (accessed Jan. 21, 2021).
- [180] Labor Inspection Board Regulation, “Ministry of Labor and Social Security,” 2012. <https://www.mevzuat.gov.tr/File/GeneratePdf?mevzuatNo=16724&mevzuatTur=KurumVeKurulusYonetmeligi&mevzuatTertip=5> (accessed Jan. 21, 2021).
- [181] M. Loosemore and N. Andonakis, “Barriers to implementing OHS reforms - The experiences of small subcontractors in the Australian Construction Industry,” *Int. J. Proj. Manag.*, vol. 25, no. 6, pp. 579–588, 2007, doi: 10.1016/j.ijproman.2007.01.015.
- [182] A. Hasan and K. N. Jha, “Safety incentive and penalty provisions in Indian construction projects and their impact on safety performance,” *Int. J. Inj. Contr. Saf. Promot.*, vol. 20, no. 1, pp. 3–12, 2013, doi: 10.1080/17457300.2011.648676.
- [183] A. Karakhan and J. Gambatese, “Hazards and Risk in Construction and the Impact of Incentives and Rewards on Safety Outcomes,” *Pract. Period. Struct. Des. Constr.*, vol. 23, no. 2, p. 04018005, 2018, doi: 10.1061/(asce)sc.1943-5576.0000359.
- [184] S. Xu, M. Zhang, and L. Hou, “Formulating a learner model for evaluating construction workers’ learning ability during safety training,” *Saf. Sci.*, vol. 116, no. August 2018, pp. 97–107, 2019, doi: 10.1016/j.ssci.2019.03.002.
- [185] A. Kirschenbaum, L. Oigenblick, and A. I. Goldberg, “Well being, work environment and work accidents,” *Soc. Sci. Med.*, vol. 50, no. 5, pp. 631–639, Mar. 2000, doi: 10.1016/S0277-9536(99)00309-3.
- [186] A. S. Wagstaff and J. A. S. Lie, “Shift and night work and long working hours - a systematic review of safety implications,” *Scand. J. Work. Environ. Heal.*, vol. 37, no. 3, pp. 173–185, 2011, doi: 10.5271/sjweh.3146.
- [187] R. Irumba, “Spatial analysis of construction accidents in Kampala, Uganda,” *Saf. Sci.*, vol. 64, pp. 109–120, 2014, doi: 10.1016/j.ssci.2013.11.024.

- [188] N. Tang, H. Hu, F. Xu, and F. Zhu, "Personalized safety instruction system for construction site based on internet technology," *Saf. Sci.*, vol. 116, no. October 2018, pp. 161–169, 2019, doi: 10.1016/j.ssci.2019.03.001.
- [189] S. Sarkar, A. Pramanik, J. Maiti, and G. Reniers, "Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data," *Saf. Sci.*, vol. 125, no. April 2019, p. 104616, 2020, doi: 10.1016/j.ssci.2020.104616.
- [190] H. Laitinen and K. Päiväranta, "A new-generation safety contest in the construction industry - A long-term evaluation of a real-life intervention," *Saf. Sci.*, vol. 48, no. 5, pp. 680–686, 2010, doi: 10.1016/j.ssci.2010.01.018.
- [191] Z. Zhu, M. W. Park, C. Koch, M. Soltani, A. Hammad, and K. Davari, "Predicting movements of onsite workers and mobile equipment for enhancing construction site safety," *Autom. Constr.*, vol. 68, pp. 95–101, 2016, doi: 10.1016/j.autcon.2016.04.009.
- [192] A. Albert, B. Pandit, Y. Patil, and J. Louis, "Does the potential safety risk affect whether particular construction hazards are recognized or not?," *J. Safety Res.*, vol. 75, pp. 241–250, 2020, doi: 10.1016/j.jsr.2020.10.004.
- [193] K. Koc, Ö. Ekmekcioğlu, and A. P. Gurgun, "Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers," *Autom. Constr.*, vol. 131, p. 103896, Nov. 2021, doi: 10.1016/j.autcon.2021.103896.
- [194] S. H. Chung, H. L. Ma, M. Hansen, and T. M. Choi, "Data science and analytics in aviation," *Transp. Res. Part E Logist. Transp. Rev.*, vol. 134, no. January, p. 101837, 2020, doi: 10.1016/j.tre.2020.101837.
- [195] B. Dietrich, J. Walther, M. Weigold, and E. Abele, "Machine learning based very short term load forecasting of machine tools," *Appl. Energy*, vol. 276, no. February, p. 115440, 2020, doi: 10.1016/j.apenergy.2020.115440.
- [196] R. Houari, A. Bounceur, M. T. Kechadi, A. K. Tari, and R. Euler, "Dimensionality reduction in data mining: A Copula approach," *Expert Syst. Appl.*, vol. 64, pp. 247–260, 2016, doi: 10.1016/j.eswa.2016.07.041.
- [197] G. H. Erharder, S. Oberhollenzer, A. Fankhauser, R. Marte, and T. Marcher, "Learning decision boundaries for cone penetration test classification," *Comput. Civ. Infrastruct. Eng.*, vol. 36, no. 4, pp. 489–503, Apr. 2021, doi: 10.1111/mice.12662.
- [198] K. Kirchner, J. Zec, and B. Delibašić, "Facilitating data preprocessing by a generic framework: a proposal for clustering," *Artif. Intell. Rev.*, vol. 45, no. 3, pp. 271–297, 2016, doi: 10.1007/s10462-015-9446-6.
- [199] K. Koc, Ö. Ekmekcioğlu, and A. P. Gurgun, "Accident prediction in construction using hybrid wavelet-machine learning," *Autom. Constr.*, vol. 133, no. February 2021, p. 103987, Jan. 2022, doi: 10.1016/j.autcon.2021.103987.
- [200] A. Möller *et al.*, "Photometric classification of type Ia supernovae in the SuperNova Legacy Survey with supervised learning," *J. Cosmol. Astropart. Phys.*, vol. 12, no. 6, pp. 1–23, 2016, doi: 10.1088/1475-7516/2016/12/008.
- [201] Y. Liu, L. Liu, L. Yang, L. Hao, and Y. Bao, "Measuring distance using ultra-wideband radio technology enhanced by extreme gradient boosting decision tree

- (XGBoost),” *Autom. Constr.*, vol. 126, no. March, p. 103678, 2021, doi: 10.1016/j.autcon.2021.103678.
- [202] M. Lombardi, M. Fagnoli, and G. Parise, “Risk profiling from the European statistics on accidents at work (ESAW) accidents’ databases: A case study in construction sites,” *Int. J. Environ. Res. Public Health*, vol. 16, no. 23, pp. 1–22, 2019, doi: 10.3390/ijerph16234748.
- [203] P. E. D. Love, P. Teo, J. Smith, F. Ackermann, and Y. Zhou, “The nature and severity of workplace injuries in construction: engendering operational benchmarking,” *Ergonomics*, vol. 62, no. 10, pp. 1273–1288, 2019, doi: 10.1080/00140139.2019.1644379.
- [204] B. Esmaeili, M. R. Hallowell, and B. Rajagopalan, “Attribute-Based Safety Risk Assessment. II: Predicting Safety Outcomes Using Generalized Linear Models,” *J. Constr. Eng. Manag.*, vol. 141, no. 8, p. 04015022, 2015, doi: 10.1061/(ASCE)CO.1943-7862.0000981.
- [205] J. Walther, D. Spanier, N. Panten, and E. Abele, “Very short-term load forecasting on factory level - A machine learning approach,” in *26th CIRP Life Cycle Engineering (LCE) Conference*, 2019, vol. 80, pp. 705–710, doi: 10.1016/j.procir.2019.01.060.
- [206] Y. Huang and S. Meng, “Automobile insurance classification ratemaking based on telematics driving data,” *Decis. Support Syst.*, vol. 127, no. August 2019, p. 113156, 2019, doi: 10.1016/j.dss.2019.113156.
- [207] F. Jalaei, F. Jalaei, and S. Mohammadi, “An integrated BIM-LEED application to automate sustainable design assessment framework at the conceptual stage of building projects,” *Sustain. Cities Soc.*, vol. 53, p. 101979, 2019, doi: 10.1016/j.scs.2019.101979.
- [208] A. I. Ozdagli and X. Koutsoukos, “Machine learning based novelty detection using modal analysis,” *Comput. Civ. Infrastruct. Eng.*, vol. 34, no. 12, pp. 1119–1140, 2019, doi: 10.1111/mice.12511.
- [209] R. Wang, V. Asghari, S. C. Hsu, C. J. Lee, and J. H. Chen, “Detecting corporate misconduct through random forest in China’s construction industry,” *J. Clean. Prod.*, vol. 268, p. 122266, 2020, doi: 10.1016/j.jclepro.2020.122266.
- [210] L. C. Lee, C. Y. Liong, and A. A. Jemain, “A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum,” *Chemom. Intell. Lab. Syst.*, vol. 163, no. December 2016, pp. 64–75, 2017, doi: 10.1016/j.chemolab.2017.02.008.
- [211] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, “New data preprocessing trends based on ensemble of multiple preprocessing techniques,” *TrAC - Trends Anal. Chem.*, vol. 132, p. 116045, 2020, doi: 10.1016/j.trac.2020.116045.
- [212] R. Akbıyıklı and Ü. Dikmen, “Main Indicators of OHS Management in Construction Sites,” *Düzce Üniversitesi Bilim ve Teknol. Derg.*, vol. 6, pp. 1391–1409, 2018, doi: 10.29130/dubited.432933.
- [213] F. D. Kakhki, S. A. Freeman, and G. A. Mosher, “Use of neural networks to identify safety prevention priorities in agro-manufacturing operations within commercial grain elevators,” *Appl. Sci.*, vol. 9, no. 21, pp. 1–16, 2019, doi:

10.3390/app9214690.

- [214] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.
- [215] L. Munkhdalai, T. Munkhdalai, O. E. Namsrai, J. Y. Lee, and K. H. Ryu, “An empirical comparison of machine-learning methods on bank client credit assessments,” *Sustain.*, vol. 11, no. 3, pp. 1–23, 2019, doi: 10.3390/su11030699.
- [216] E. E. Başakın, Ö. Ekmekcioğlu, H. Çıtakoğlu, and M. Özger, “A new insight to the wind speed forecasting: robust multi-stage ensemble soft computing approach based on pre-processing uncertainty assessment,” *Neural Comput. Appl.*, vol. 34, no. 1, pp. 783–812, 2022, doi: 10.1007/s00521-021-06424-6.
- [217] D. Effrosynidis and A. Arampatzis, “An evaluation of feature selection methods for environmental data,” *Ecol. Inform.*, vol. 61, no. January, p. 101224, 2021, doi: 10.1016/j.ecoinf.2021.101224.
- [218] M. Loecher, “Unbiased variable importance for random forests,” *Commun. Stat. - Theory Methods*, vol. 51, no. 5, pp. 1413–1425, Mar. 2022, doi: 10.1080/03610926.2020.1764042.
- [219] M. Fargnoli, M. De Minicis, and G. Di Gravio, “Knowledge Management integration in Occupational Health and Safety systems in the construction industry,” *Int. J. Prod. Dev.*, vol. 14, no. 1/2/3/4, p. 165, 2011, doi: 10.1504/IJPD.2011.042298.
- [220] G. Raviv, A. Shapira, and B. Fishbain, “AHP-based analysis of the risk potential of safety incidents: Case study of cranes in the construction industry,” *Saf. Sci.*, vol. 91, pp. 298–309, 2017, doi: 10.1016/j.ssci.2016.08.027.
- [221] J. Zhu, Z. Ge, Z. Song, and F. Gao, “Review and big data perspectives on robust data mining approaches for industrial process modeling with outliers and missing data,” *Annu. Rev. Control*, vol. 46, pp. 107–133, 2018, doi: 10.1016/j.arcontrol.2018.09.003.
- [222] M. Ayhan, I. Dikmen, and M. T. Birgonul, “Predicting the Occurrence of Construction Disputes Using Machine Learning Techniques,” *J. Constr. Eng. Manag.*, vol. 147, no. 4, p. 04021022, 2021, doi: 10.1061/(ASCE)CO.1943-7862.0002027.
- [223] J. L. Lansford and D. G. Vlachos, “Infrared spectroscopy data- and physics-driven machine learning for characterizing surface microstructure of complex materials,” *Nat. Commun.*, vol. 11, no. 1, pp. 1–12, 2020, doi: 10.1038/s41467-020-15340-7.
- [224] S. Sarkar, A. Pramanik, J. Maiti, and G. Reniers, “Predicting and analyzing injury severity: A machine learning-based approach using class-imbalanced proactive and reactive data,” *Saf. Sci.*, vol. 125, no. December 2019, p. 104616, 2020, doi: 10.1016/j.ssci.2020.104616.
- [225] M. Amiri, A. Ardeshtir, M. H. Fazel Zarandi, and E. Soltanaghaei, “Pattern extraction for high-risk accidents in the construction industry: a data-mining approach,” *Int. J. Inj. Contr. Saf. Promot.*, vol. 23, no. 3, pp. 264–276, 2016, doi: 10.1080/17457300.2015.1032979.
- [226] M.-Y. Leung, I. Y. S. Chan, and J. Yu, “Preventing construction worker injury

- incidents through the management of personal stress and organizational stressors,” *Accid. Anal. Prev.*, vol. 48, pp. 156–166, 2012, doi: 10.1016/j.aap.2011.03.017.
- [227] S. Unsar and N. Sut, “General assessment of the occupational accidents that occurred in Turkey between the years 2000 and 2005,” *Saf. Sci.*, vol. 47, no. 5, pp. 614–619, 2009, doi: 10.1016/j.ssci.2008.08.001.
- [228] V. Villanueva and A. M. Garcia, “Individual and occupational factors related to fatal occupational injuries: A case-control study,” *Accid. Anal. Prev.*, vol. 43, no. 1, pp. 123–127, 2011, doi: 10.1016/j.aap.2010.08.001.
- [229] F. Vahdatikhaki and A. Hammad, “Risk-based look-ahead workspace generation for earthwork equipment using near real-time simulation,” *Autom. Constr.*, vol. 58, pp. 207–220, 2015, doi: 10.1016/j.autcon.2015.07.019.
- [230] C. W. Liao and Y. H. Perng, “Data mining for occupational injuries in the Taiwan construction industry,” *Saf. Sci.*, vol. 46, no. 7, pp. 1091–1102, 2008, doi: 10.1016/j.ssci.2007.04.007.
- [231] B. Tarik and H. A. Adil, “Occupational health and safety in the Moroccan construction sites: Preliminary diagnosis,” *Int. J. Metrol. Qual. Eng.*, vol. 9, no. 6, pp. 1–9, 2018, doi: 10.1051/ijmqe/2018005.
- [232] F. Yilmaz, “Analysis of the interaction between safety measures and workers’ safety awareness from the construction workers’ perspective,” *Eng. Constr. Archit. Manag.*, Oct. 2021, doi: 10.1108/ECAM-07-2021-0564.
- [233] Z. Zhou, Y. M. Goh, and Q. Li, “Overview and analysis of safety management studies in the construction industry,” *Saf. Sci.*, vol. 72, pp. 337–350, 2015, doi: 10.1016/j.ssci.2014.10.006.
- [234] H. Alsulami, S. H. Serbaya, A. Rizwan, M. Saleem, Y. Maleh, and Z. Alamgir, “Impact of emotional intelligence on the stress and safety of construction workers’ in Saudi Arabia,” *Eng. Constr. Archit. Manag.*, Dec. 2021, doi: 10.1108/ECAM-06-2021-0481.
- [235] K. Versteeg, P. Bigelow, A. M. Dale, and A. Chaurasia, “Utilizing construction safety leading and lagging indicators to measure project safety performance: A case study,” *Saf. Sci.*, vol. 120, pp. 411–421, 2019, doi: 10.1016/j.ssci.2019.06.035.
- [236] Japan Industrial Safety and Health Association, “OSH statistics in Japan, industrial accident in 2019,” 2019. <https://www.jisha.or.jp/english/statistics/> (accessed Feb. 08, 2020).
- [237] Health and Safety Executive, “Workplace fatal injuries in Great Britain 2018,” *Heal. Saf. Exec.*, no. March, pp. 1–16, 2020, [Online]. Available: <http://www.hse.gov.uk/statistics/pdf/fatalinjuries.pdf>.
- [238] Y. Niu *et al.*, “Towards the ‘third wave’: An SCO-enabled occupational health and safety management system for construction,” *Saf. Sci.*, vol. 111, pp. 213–223, 2019, doi: 10.1016/j.ssci.2018.07.013.
- [239] M. F. Antwi-Afari *et al.*, “Sensing and warning-based technology applications to improve occupational health and safety in the construction industry: A literature review,” *Eng. Constr. Archit. Manag.*, vol. 26, no. 8, pp. 1534–1552, 2019, doi: 10.1108/ECAM-05-2018-0188.

- [240] A. Ajayi *et al.*, “Deep Learning Models for Health and Safety Risk Prediction in Power Infrastructure Projects,” *Risk Anal.*, vol. 40, no. 10, pp. 2019–2039, 2020, doi: 10.1111/risa.13425.
- [241] D. Alexander, M. Hallowell, and J. Gambatese, “Precursors of construction fatalities. II: Predictive modeling and empirical validation,” *J. Constr. Eng. Manag.*, vol. 143, no. 7, p. 04017024, 2017, doi: 10.1061/(ASCE)CO.1943-7862.0001297.
- [242] H. Baker, M. R. Hallowell, and A. J.-P. Tixier, “Automatically learning construction injury precursors from text,” *Autom. Constr.*, vol. 118, p. 103145, 2020, doi: 10.1016/j.autcon.2020.103145.
- [243] Ö. Ekmekcioğlu, K. Koc, and M. Özger, “District based flood risk assessment in Istanbul using fuzzy analytical hierarchy process,” *Stoch. Environ. Res. Risk Assess.*, vol. 35, no. 3, pp. 617–637, Mar. 2021, doi: 10.1007/s00477-020-01924-8.
- [244] T. Sarı, “Review On High-Rise Housing Projects in Istanbul: Toward A Sustainable Architecture,” *J. Sustain. Archit. Civ. Eng.*, vol. 1, no. 18, pp. 39–49, 2017, doi: 10.5755/j01.sace.18.1.16598.
- [245] Ç. Göksel, R. M. David, and A. O. Dogru, “Environmental Monitoring of Spatio-Temporal Changes in Northern Istanbul using Remote Sensing and GIS,” *Int. J. Environ. Geoinformatics*, vol. 5, no. 1, pp. 94–103, 2018, doi: 10.30897/ijgeo.410943.
- [246] A. Iqbal, S. Aftab, and F. Matloob, “Performance Analysis of Resampling Techniques on Class Imbalance Issue in Software Defect Prediction,” *Int. J. Inf. Technol. Comput. Sci.*, vol. 11, no. 11, pp. 44–53, Nov. 2019, doi: 10.5815/ijitcs.2019.11.05.
- [247] R. Ghorbani and R. Ghousi, “Comparing Different Resampling Methods in Predicting Students’ Performance Using Machine Learning Techniques,” *IEEE Access*, vol. 8, pp. 67899–67911, 2020, doi: 10.1109/ACCESS.2020.2986809.
- [248] TUIK, “Geographic Statistics Portal,” 2021. .
- [249] Ministry of Labor and Social Security, “Occupational Health and Safety Registration, Tracking and Monitoring Program,” 2021. .
- [250] E. Dogan and A. Stupar, “The limits of growth: A case study of three mega-projects in Istanbul,” *Cities*, vol. 60, pp. 281–288, Feb. 2017, doi: 10.1016/j.cities.2016.09.013.
- [251] K. Koc and A. P. Gurgun, “Stakeholder-Associated Life Cycle Risks in Construction Supply Chain,” *J. Manag. Eng.*, vol. 37, no. 1, p. 04020107, 2021, doi: 10.1061/(ASCE)ME.1943-5479.0000881.
- [252] S. Ulubeyli, V. Arslan, and S. Kivrak, “A semiotic analysis of cartoons about occupational health and safety issues in the construction workplace,” *Constr. Manag. Econ.*, vol. 33, no. 5–6, pp. 467–483, 2015, doi: 10.1080/01446193.2015.1024270.
- [253] Y. M. Goh, C. U. Ubeynarayana, K. L. X. Wong, and B. H. W. Guo, “Factors influencing unsafe behaviors: A supervised learning approach,” *Accid. Anal. Prev.*, vol. 118, no. June, pp. 77–85, 2018, doi: 10.1016/j.aap.2018.06.002.

- [254] S. R. Mohandes *et al.*, “Causal analysis of accidents on construction sites: A hybrid fuzzy Delphi and DEMATEL approach,” *Saf. Sci.*, vol. 151, no. March, p. 105730, Jul. 2022, doi: 10.1016/j.ssci.2022.105730.
- [255] C. W. Cheng, S. Sen Leu, C. C. Lin, and C. Fan, “Characteristic analysis of occupational accidents at small construction enterprises,” *Saf. Sci.*, vol. 48, no. 6, pp. 698–707, 2010, doi: 10.1016/j.ssci.2010.02.001.
- [256] S. Winge and E. Albrechtsen, “Accident types and barrier failures in the construction industry,” *Saf. Sci.*, vol. 105, no. February, pp. 158–166, 2018, doi: 10.1016/j.ssci.2018.02.006.
- [257] S. Baradan, S. U. Dikmen, and O. Akboga Kale, “Impact of human development on safety consciousness in construction,” *Int. J. Occup. Saf. Ergon.*, vol. 25, no. 1, pp. 40–50, 2019, doi: 10.1080/10803548.2018.1445069.
- [258] S. J. Ray and J. Teizer, “Dynamic blindspots measurement for construction equipment operators,” *Saf. Sci.*, vol. 85, pp. 139–151, 2016, doi: 10.1016/j.ssci.2016.01.011.
- [259] A. J. P. Tixier, M. R. Hallowell, B. Rajagopalan, and D. Bowman, “Automated content analysis for construction safety: A natural language processing system to extract precursors and outcomes from unstructured injury reports,” *Autom. Constr.*, vol. 62, no. 2016, pp. 45–56, 2016, doi: 10.1016/j.autcon.2015.11.001.
- [260] N. Ghodrati, T. W. Yiu, S. Wilkinson, and M. Shahbazzpour, “A new approach to predict safety outcomes in the construction industry,” *Saf. Sci.*, vol. 109, pp. 86–94, 2018, doi: 10.1016/j.ssci.2018.05.016.
- [261] S. Chen, X. Wang, H. Zhang, and J. Wang, “Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine,” *Expert Syst. Appl.*, vol. 173, no. January, p. 114756, 2021, doi: 10.1016/j.eswa.2021.114756.
- [262] L. Zhou, “Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods,” *Knowledge-Based Syst.*, vol. 41, pp. 16–25, Mar. 2013, doi: 10.1016/j.knosys.2012.12.007.
- [263] W. Wu, A. G. F. Gibb, and Q. Li, “Accident precursors and near misses on construction sites: An investigative tool to derive information from accident databases,” *Saf. Sci.*, vol. 48, no. 7, pp. 845–858, Aug. 2010, doi: 10.1016/j.ssci.2010.04.009.
- [264] S. Wang and X. Chen, “Adaptive Binary Coding for Scene Classification Based on Convolutional Networks,” *Comput. Mater. Contin.*, vol. 65, no. 3, pp. 2065–2077, 2020, doi: 10.32604/cmc.2020.09857.
- [265] S. Pradeep and J. S. Kallimani, “Machine Learning Based Predictive Action on Categorical Non-Sequential Data,” *Recent Adv. Comput. Sci. Commun.*, vol. 13, no. 5, pp. 1020–1030, Nov. 2020, doi: 10.2174/2213275912666190417150421.
- [266] F. Provost, “Machine learning from imbalanced data sets 101,” in *Proceedings of the AAAI’2000 Workshop on ...*, 2000, pp. 1–3, [Online]. Available: <https://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-001.pdf%5Cnpapers://1c40c143-2a6e-4e94-8c17-c5bc9ae73d7e/Paper/p11435>.
- [267] P. Wolff, S. A. Ríos, and M. Graña, “Setting up standards: A methodological proposal for pediatric Triage machine learning model construction based on

- clinical outcomes,” *Expert Syst. Appl.*, vol. 138, p. 112788, 2019, doi: 10.1016/j.eswa.2019.07.005.
- [268] G. Du, J. Zhang, Z. Luo, F. Ma, L. Ma, and S. Li, “Joint imbalanced classification and feature selection for hospital readmissions,” *Knowledge-Based Syst.*, vol. 200, p. 106020, 2020, doi: 10.1016/j.knosys.2020.106020.
- [269] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.
- [270] Ö. F. Arar and K. Ayan, “A feature dependent Naive Bayes approach and its application to the software defect prediction problem,” *Appl. Soft Comput.*, vol. 59, pp. 197–209, Oct. 2017, doi: 10.1016/j.asoc.2017.05.043.
- [271] F. Davoudi Kakhki, S. A. Freeman, and G. A. Mosher, “Evaluating machine learning performance in predicting injury severity in agribusiness industries,” *Saf. Sci.*, vol. 117, pp. 257–262, Aug. 2019, doi: 10.1016/j.ssci.2019.04.026.
- [272] J. R. Rico-Juan and P. Taltavull de La Paz, “Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain,” *Expert Syst. Appl.*, vol. 171, no. January, p. 114590, Jun. 2021, doi: 10.1016/j.eswa.2021.114590.
- [273] A. Di Noia, A. Martino, P. Montanari, and A. Rizzi, “Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction,” *Soft Comput.*, vol. 24, no. 6, pp. 4393–4406, Mar. 2020, doi: 10.1007/s00500-019-04200-2.
- [274] A. Azadeh, M. Rouzbahman, M. Saberi, F. Valianpour, and A. Keramati, “Improved prediction of mental workload versus HSE and ergonomics factors by an adaptive intelligent algorithm,” *Saf. Sci.*, vol. 58, pp. 59–75, Oct. 2013, doi: 10.1016/j.ssci.2013.03.004.
- [275] J. Xu, J. Zhao, Z. Song, and M. Liu, “Prediction of ultimate bearing capacity of Tubular T-joint under fire using artificial neural networks,” *Saf. Sci.*, vol. 50, no. 7, pp. 1495–1501, Aug. 2012, doi: 10.1016/j.ssci.2012.02.004.
- [276] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006, doi: 10.1016/j.patrec.2005.10.010.
- [277] B. Soni, D. K. Patel, and M. Lopez-Benitez, “Long Short-Term Memory Based Spectrum Sensing Scheme for Cognitive Radio Using Primary Activity Statistics,” *IEEE Access*, vol. 8, pp. 97437–97451, 2020, doi: 10.1109/ACCESS.2020.2995633.
- [278] Y. Wu, J. Yao, S. Chang, and B. Liu, “LIMCR: Less-informative majorities cleaning rule based on naïve bayes for imbalance learning in software defect prediction,” *Appl. Sci.*, vol. 10, no. 23, pp. 1–24, 2020, doi: 10.3390/app10238324.
- [279] A. D. Bykov, V. I. Voronov, and L. I. Voronova, “Machine Learning Methods Applying for Hydraulic System States Classification,” in *Systems of Signals Generating and Processing in the Field of on Board Communications*, 2019, pp. 1–4.
- [280] J. Zhang, Y. Wang, Y. Sun, and G. Li, “Strength of ensemble learning in multiclass classification of rockburst intensity,” *Int. J. Numer. Anal. Methods Geomech.*, vol. 44, no. 13, pp. 1833–1853, 2020, doi: 10.1002/nag.3111.

- [281] H. Son and C. Kim, "Early prediction of the performance of green building projects using pre-project planning variables: Data mining approaches," *J. Clean. Prod.*, vol. 109, pp. 144–151, 2015, doi: 10.1016/j.jclepro.2014.08.071.
- [282] B. H. Menze *et al.*, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," *BMC Bioinformatics*, vol. 10, no. 213, pp. 1–16, 2009, doi: 10.1186/1471-2105-10-213.
- [283] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: A corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, 2010, doi: 10.1093/bioinformatics/btq134.
- [284] C. F. Chi, T. C. Chang, and H. I. Ting, "Accident patterns and prevention measures for fatal occupational falls in the construction industry," *Appl. Ergon.*, vol. 36, pp. 391–400, 2005, doi: 10.1016/j.apergo.2004.09.011.
- [285] C. Melchior, R. R. Zanini, R. R. Guerra, and D. A. Rockenbach, "Forecasting Brazilian mortality rates due to occupational accidents using autoregressive moving average approaches," *Int. J. Forecast.*, vol. 37, no. 2, pp. 825–837, Apr. 2021, doi: 10.1016/j.ijforecast.2020.09.010.
- [286] D. Arditi, D.-E. Lee, and G. Polat, "Fatal accidents in nighttime vs. daytime highway construction work zones," *J. Safety Res.*, vol. 38, no. 4, pp. 399–405, 2007, doi: 10.1016/j.jsr.2007.04.001.
- [287] S. D. Choi, L. Guo, J. Kim, and S. Xiong, "Comparison of fatal occupational injuries in construction industry in the United States, South Korea, and China," *Int. J. Ind. Ergon.*, vol. 71, pp. 64–74, May 2019, doi: 10.1016/j.ergon.2019.02.011.
- [288] B. Yong Jeong, "Comparisons of variables between fatal and nonfatal accidents in manufacturing industry," *Int. J. Ind. Ergon.*, vol. 23, no. 5–6, pp. 565–572, Mar. 1999, doi: 10.1016/S0169-8141(98)00020-1.

PUBLICATIONS FROM THE THESIS

Conference Papers

1. K. Koc, and A. P. Gurgun “Machine Learning Applications in Construction Safety Literature” in The Eleventh International Structural Engineering and Construction Conference, ISEC 11- Interdisciplinary Civil and Construction Engineering Projects, 26-31 July 2021, 8(1), 1–6, Cairo, Egypt, doi: 10.14455/ISEC.2021.8(1).CSA-05.

Papers

1. K. Koc, Ö. Ekmekcioğlu, and A. P. Gurgun “Integrating feature engineering, genetic algorithm and tree-based machine learning methods to predict the post-accident disability status of construction workers”. *Automation in Construction*, 131, 103896, November 2021, doi: 10.1016/j.autcon.2021.103896.

2. K. Koc, Ö. Ekmekcioğlu, and A. P. Gurgun “Accident Prediction in Construction using Hybrid Wavelet-Machine Learning”. *Automation in Construction*, 133, 103987, January 2022, doi: 10.1016/j.autcon.2021.103987.

3. K. Koc, and A. P. Gurgun “Scenario-Based Automated Data Preprocessing to Predict Severity of Construction Accidents”. *Automation in Construction*, 140, 104351, August 2022, doi: 10.1016/j.autcon.2022.104351.

4. K. Koc, Ö. Ekmekcioğlu, and A. P. Gurgun “Prediction of construction accident outcomes based on imbalanced dataset through integrated resampling techniques and machine learning methods”. *Engineering, Construction and Architectural Management*, June 2022, doi: 10.1108/ECAM-04-2022-0305.