Prediction of Occupational Accidents Using Decision Tree Approach

Conference Paper · December 2016

DOI: 10.1109/INDICON.2016.7838969

Sobhan Sarkar

Indian Institute of Technology Kharagpur

20 PUBLICATIONS **26** CITATIONS

SEE PROFILE

Sarthak Madaan

Indian Institute of Technology Kharagpur

1 PUBLICATION **4** CITATIONS

SEE PROFILE

Prediction of Occupational Accidents Using Decision Tree Approach

Sobhan Sarkar
Research Scholar
Department of Industrial and Systems Engineering,
Indian Institute of Technology Kharagpur,
sobhan.sarkar@gmail.com

Atul Patel
B.Tech.
Department of Electrical Engineering,
Indian Institute of Technology, Kharagpur,
atulp.iitkgp@gmail.com

Sarthak Madaan
B.Tech.
Department of Industrial and Systems Engineering,
Indian Institute of Technology, Kharagpur,
sarthakmadaan5121995@gmail.com

Jhareswar Maiti
Professor
Department of Industrial and Systems Engineering,
Indian Institute of Technology, Kharagpur
jhareswar.maiti@gmail.com

Abstract: ***The focus of the present study is to build a predictive model which not only could predict the occupational incidents but also provide rules for explaining accident scenarios like near-miss, property damage, or injury cases. Classification and regression tree (CART) is used for prediction purpose. Furthermore, the parameters of CART have been tuned by grid based tuning and genetic algorithm (GA). The experimental results show that the GA optimized CART provides better accuracy than others. Additionally, the best rules extracted from GA optimized CART are discussed in order to adopt better safety precautionary measures at work.***

Keywords*: *Occupational accidents; Steel plant; Prediction; CART; Grid search; Genetic algorithm.

I. INTRODUCTION

Occupational accident is a major problem in every industry. As per the statistics provided by International Labour Organization (ILO), there are more than 317 million accidents reported at work every year, together with occupational diseases resulting in more than 2.3 million deaths annually [1]. According to the study by Hämäläinen et al. [2], more than 48000 workers die yearly in India because of occupational accidents, and almost 37 million cases leads to at least 3 days' absence from work. Their study reported that the rate of fatal accidents was 11.4 per 1,00,000 workers and rate of accidents was 8,700 per 1,00,000 workers in India. Further, it is seen that the overall accident rate has remained almost constant for many years with an increase in serious injury and fatality rates in recent years. There have been many accidents in steel plants in India in recent years. According to statistics reported by Press Information Bureau, 112 fatal and 339 non-fatal accidents were reported in the years 2011-2013 [3]. According to World Steel Association, employee safety and health is fundamental to sustainability [4]. It also states that safety in steel plants is associated with superior performance, the most successful steel companies are also found to be the safest. Thus, it is a matter of concern today that even though we have enough experience of fatal, serious, and lost time accidents, but we are not learning from the past data. If we could predict the future occurrences of events using the past data, that would be beneficial for the industry concerned. So, prediction, in the domain of occupational accident research, has been found out to be an important approach that works on the basic assumption that accidents can be prevented, if they could be predicted. Thus, prediction of occupational accidents in 'safety analytics' practice at works has been proved to be essential in order to reduce the probability of occurrence of accidents that, in turn, lessen the economic burden of nations, as well.

Based on the need, our research focuses on the predictive modeling of occupational accidents in an integrated steel plant in India. The main aim of our work is to predict the incident category (injury, near-miss, and property damage) of accidents. To build the predictive model, Classification and Regression Tree (CART) has been used to the data obtained from the plant. Genetic algorithm (GA) and grid based tuning have been used to optimize the parameter of CART which could provide maximum prediction accuracy.

The rest of the paper is organized as follows. In Section II, brief review of related works on analysis of occupational accident prediction is given. Problem description is provided in Section III. Methodological flowchart and brief description of methods are described in Section IV. In Section V, results and discussion are provided. Finally, conclusions and future scopes of the study are mentioned in Section VI.

II. RELATED WORKS

Since the last decade, analysis of occupational accident using various data mining techniques has increased rapidly. In recent years, many researchers have worked on accident prediction or classification using large available data. Classification and Regression Tree (CART) is a decision tree based data mining method, first introduced by Breiman et al. [5], which has been used by many researchers for occupational accident analysis. Chang et al. proposed a CART model to establish the relationship between injury severity and driver/vehicle characteristics, highway/environmental variables and accident variables [6]. Cheng et al. utilized Taiwan construction industry database using CART to establish potential cause and effect relationships regarding serious occupational accidents in the industry [7]. This study provided a framework for improving the safety practices and training programs that are essential in construction industry. Lukáčová et al. proposed the use of decision tree to predict the level of incident's seriousness and also being able to identify possible risk situations based on various input parameters [8]. Pereira et al. proposed a methodology using topic modelling for extracting information from textual data to implement appropriate mitigation measures in incident prediction [9]. In line with this, text mining based prediction with Bayesian Network has also much potential for incidents' prediction [10]. Rivas et al. draws a comparison between classical statistical techniques such as logistic regression and other data mining methods such as decision rule, classification tree in prediction and identification of the factors underlying accidents [11]. Bevilacqua et al. used classification tree methods on refinery accidents database to identify important relationships between variables which can be considered in safety management [12]. Mistikoglu et al. proposed a methodology using decision tree on Occupational Safety and Health administration (OSHA) dataset to determine associations between input variables and output variable (i.e., degree of injury) [13].

Based on the literature review, though all other works are not mentioned in this paper due to space limitation, it is found that there are some of the research gaps in this domain which are required to be addressed. *First,* previous research works deal with continuous data, or categorical data, or both continuous and categorical, or textual data extensively, but this study uses continuous data, categorical data, and unstructured text data simultaneously for incident prediction. *Second,* though CART has been used frequently by many researchers for classification purpose, optimization of CART for obtaining better accuracy has hardly been reported in occupational accident literature. Thus, our proposed method has much potential towards the prediction of occupational accidents with higher accuracy.

III. PROBLEM DESCRIPTION

In this study, we attempt to fill the research gap by taking data from a steel manufacturing industry situated in India. The organization (under study) has been experiencing accidents in its workplace. The safety norms, precautionary measures, and safety data base which, although are available, have not been nevertheless found to be effective enough to mitigate the occurrence of accidents. Thus, the company under study has been seeking some predictive solutions having much potential to reduce the incidents at workplace.

A. Data set

To conduct the research, data has been retrieved from the electronic database of the integrated steel plant. It contains 4744 records of incidents which were already taken place in a span of years 2010 to 2013. The dataset comprises 15 attributes including three numerical, two textual and 10 categorical ones. The types and classes of attributes are shown in Table I. The total dataset is split into 70% and 30% for training and testing purpose for decision tree implementation.

TABLE I. DATA SET DESCRIPTION.

Attribute	**Type**	**Number of Classes**
Month	Categorical	12 (January,...December)
Division	Categorical	14 (Div 1… Div14)
Incident Category	Categorical	3 (Injury, Near Miss, Property Damage)
Primary Cause	Categorical	11 (EOT Cranes, Material Handling, Slip/Trip/Fall, Process, Electrical, Rail, Road Incident, Working at Height, Heavy vehicles and Machinery equipment, Civil and Mechanical Structure, Others)
Status	Categorical	6 (Close,Pending1,..Pending5)
Working Condition	Categorical	3 (Single Working, Group Working, Not Applicable)
Machine Condition	Categorical	3 (M/C Idle, M/C Working, Not Applicable).
Observation Type	Categorical	4 (Unsafe Act, Unsafe Condition, Unsafe Act & Unsafe Condition, Unsafe Act by Other)
Incident Type	Categorical	2 (Behavioral, Process)
SOP	Categorical	3 (Safe, Dangerous, Very Dangerous)
Serious Process Incident Score	Numeric	5 i.e., 0: not serious (0 - 50): low (51 - 100): medium (101 - 150): high (151 - 500): very high
Injury Potential Score	Numeric	5 i.e., (1 – 5): very low (6 – 10): low (11 – 15): medium (16 – 20): high (21 - 25): very high
Equipment Damage Score	Numeric	5 i.e., (1 - 5): very low (6 – 10): low (11 - 15): medium (16 – 20): high (21 – 25): very high
Description of Incident	textual	9 (desc 1, desc 2, …, desc 9)
Event leading to Incident	textual	5 (event 1, event 2, …, event 5)

IV. METHODOLOGY

The methods used for analysis are depicted in the flowchart in Fig 1.

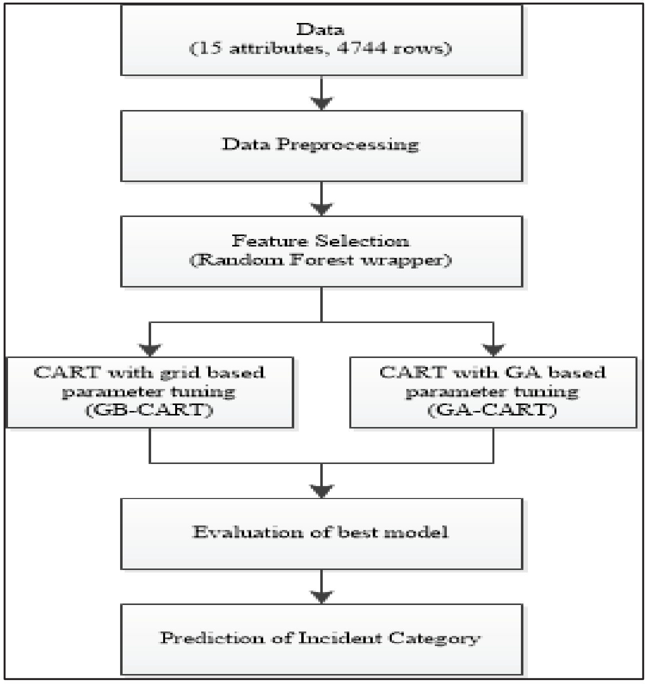
Data
(15 attributes, 4744 rows)
Data Preprocessing
Feature Selection
(Random Forest wrapper)
CART with grid based parameter tuning (GB-CART)
CART with GA based parameter tuning (GA-CART)
Evaluation of best model
Prediction of Incident Category

Fig. 1. Proposed methodological flowchart.

A. Data Preprocessing

Data preprocessing is an important part of data mining. Quality of results is directly proportional to quality of data. Missing values in the data were imputed using random forest method. Topic modeling (discussed later) was used to convert textual attributes into categorical and data transformation was applied to numeric attributes in order to convert them into categorical attributes by experts' judgement for improving the accuracy of the analysis.

B. Topic modeling

Topic modeling is a frequently used in text mining for discovery of hidden semantic structures in a body of text. It forms different clusters of similar words called topics from textual data. Documents can be classified as belonging to topics based on the statistics of words in the documents. A document belongs to different topics in different proportions depending on the frequencies of words from each topic and is assigned a topic on the basis of these proportions.

The Structural Topic Model (STM) was used for topic modeling of textual attributes in this study. It incorporates metadata into topic modeling framework and leverages covariate information. According to Roberts et al. [14], STM performs significantly better than other topic models like Latent Dirichlet Allocation (LDA). Using STM for topic modeling, each row of textual data in the text attributes was taken as a document, and all these documents were used to form the topic model. The number of topics was selected for the attributes on the basis of values of semantic coherence, held-out likelihood and residuals for different number of topics. After generating the topic model, topics were assigned to documents on the basis of topic proportions.

C. Feature Selection

Feature selection is the process of selecting a subset of relevant features or attributes as dependent attributes in a predictive model, thus leading to reduction of model overfitting and improved prediction accuracy [15]. A wrapper built around the random forest classification algorithm is used for feature selection [16]. It extends the information system by adding copies of variables and random forest classifier is applied on the extended information system. Finally, the Z-scores computed by dividing the average accuracy loss by its standard deviation is used as the measure of importance of corresponding features.

D. Classification and Regression Tree (CART)

Classification and Regression Tree (CART) is a widely used decision tree (DT) based classification technique introduced by Breiman et al. [5]. It is a recursive partitioning method that can be used for both classification and regression by using classification and regression trees for categorical and numeric dependent variables, respectively. This algorithm is basically a sequence of carefully crafted questions about the attributes of the data [17]. After an answer is given for a question, a subsequent question is asked until a conclusion about the class label can be derived. These questions can be framed into a form of hierarchical structure consisting of nodes and directed edges. In principle, there are many decision trees that can be constructed from a given set of attributes. A greedy strategy is employed by CART for partitioning the data. It uses Gini index as a measure for determining the best split which is based on the degree of impurity of the child nodes.

$$Gini(t) = 1 - \sum_{i=0}^{c-1} [p(i|t)]^2$$

Different control parameters can be used to control the results of a prediction model. CART has various control parameters like maximum depth of tree, minimum number of observations in a terminal node, and complexity parameter (cp).

E. Grid based parameter tuning

Different control parameters can be used to control the results of a prediction model. The control parameters can be selected to maximize accuracy of CART on unseen data by the use of optimization methods. One such method is grid based tuning of parameters that constructs a grid or set of combinations of control parameters. Then the model is applied using all combinations in the grid and finally the combination that gives maximum accuracy is used for the final model. Complexity parameter was tuned keeping other control parameters

constant to maximize accuracy in the CART model. The optimal parameter setting from grid based parameter tuning is as follows:
Complexity parameter (cp) – 0.0014
Minimum observation in the terminal nodes (minbucket) – 7.0
Maximum depth of the tree (maxdepth) - 30

F. Genetic Algorithms (GA)

Genetic algorithm, first introduced by Holland, belongs to the class of evolutionary algorithms, i.e., they are inspired by the process of natural evolution of living beings [18]. GA is generally used to solve optimization and search problems. A set of control parameters that gives maximum accuracy with previously unseen data is obtained from GA based optimization approach. Control parameters are allowed to vary between given lower and upper bounds and GA is used to form a population of solutions, i.e., a set of control parameters. CART model is trained on a portion of data with each set of control parameters and tested on previously unseen data. The control parameters used are maximum depth of tree, minimum number of observations in a terminal node, and cp. Accuracy of CART on test data is used as fitness measure for solutions, and solutions with better accuracy on test data are used to breed next generation of solutions. The best fitness value, i.e., maximum accuracy in a population increases over generations. Finally, one of the solutions that give maximum accuracy in the final iteration is used for the CART prediction model. The optimal parameter setting from GA-based CART is as follows:
Complexity parameter (cp) – 0.00137
Minimum observation in the terminal nodes (minbucket) – 27
Maximum depth of the tree (maxdepth) – 9

G. Pruning

Pruning is a machine learning technique used to reduce the complexity of a decision tree classifier in order to avoid model overfitting. Pruning can be categorized into two broad types, pre-pruning and post-pruning. Pre-pruning halts the tree growing algorithm before generating a fully grown tree by some stopping condition, whereas in post-pruning, tree is initially grown to its maximum size and then trimmed in a bottom-up fashion [17]. In our analysis we have post-pruned grid based CART model (GB-CART) as it was overfitting the training data. Post pruning was not considered on GA based CART (GA-CART) as no over fitting on training data was observed. In GB-CART, post pruning was performed by using cp equal to 0.005.

V. RESULTS AND DISCUSSION

The results are discussed below in section-wise.

A. Topic modeling

The STM was used to categorize "Description of Incident" and "Event leading to accident" into nine and five topics, respectively based on word frequencies. The new categorical variables were labelled as "descr" and "events", respectively. The number of factors for "descr" and "events" are shown in Table I.

B. Feature Selection

A wrapper built around random forest classification algorithm was used for feature selection. **Fig. 2** depicts a plot of attribute importance versus attributes using the wrapper algorithm.

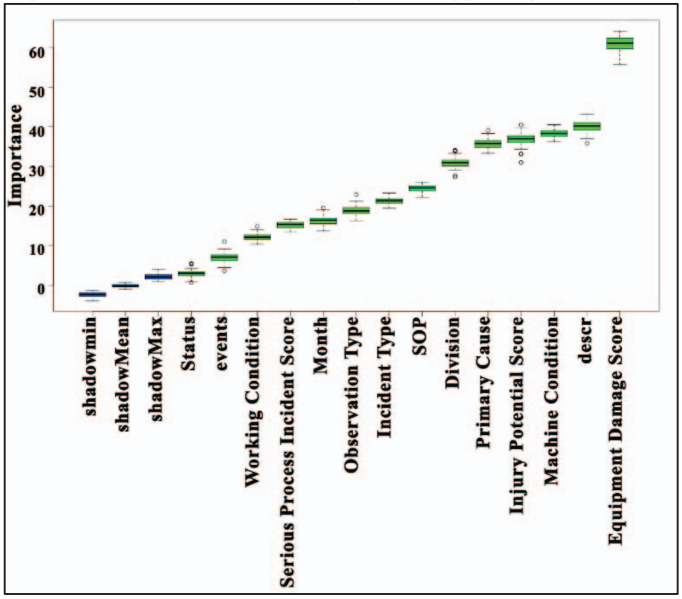

Fig. 2. Feature selection using Random Forest Wrapper.

Considering 30 % importance as threshold, six attributes i.e., 'descr', 'Equipment damage score', 'Machine condition', 'Primary cause', 'Injury potential score' and 'Division' were finally selected for prediction using decision trees **(Fig. 2)**. Rest of the attributes was discarded as they have very little effect in the prediction of the dependent variable.

TABLE II. PERFORMANCE OF DECISION TREES WITH DIFFERENT METRIC.

Method	Sensitivity (Recall)	Specificity	Precision	Class
Grid-based CART (GB-CART)	0.7923	0.8155	0.7005	Injury
	0.6122	0.7677	0.6411	Near Miss
	0.5653	0.8936	0.6357	Property Damage
Pruned GB-CART	0.7681	0.8220	0.6978	Injury
	0.6052	0.7700	0.6408	Near Miss
	0.6193	0.8917	0.6526	Property Damage
GA-CART	0.7802	0.8403	0.7234	Injury
	0.6487	0.7583	0.6453	Near Miss

	0.5881	0.9038	0.6677	Property Damage

TABLE III. ACCURACY OF THE DECISION TREE CLASSIFIERS.

Method	Accuracy (Training Data)	Accuracy (Test Data)
GB-CART	0.7209	0.6634
Pruned GB-CART	0.6612	0.6655
GA-CART	0.6998	0.6796

C. Classification and Regression Tree (CART)

First, CART was applied to the data with cp as obtained from tuning. Accuracies for different values of complexity parameter are used to select the value of complexity parameter that gives maximum accuracy for specific values of other control parameters. Table III shows the accuracies for both training and test data and there is a considerable difference between the two for CART model. As it was highly over fitting the training data, the CART model was pruned to decrease generalization error. Then GA based optimization was used to obtain the optimal parameter settings of the control parameters that give maximum accuracy. CART was applied using these control parameters. It is observed that the model generated meaningful results. It is shown from the Table III, that by using GA based optimization, accuracy of the classifier can be increased. An increase in accuracy is also observed after tree-pruning. With the help of this analysis prediction can be made with improved accuracy and based on the rules generated by the analysis, precautions and safety measures can be implemented in steel industry to prevent such accidents. GA based tuning of control parameters with population size 500 and 100 generations (i.e. iterations) results in the best fitness value 0.6796 which is the accuracy of the GA-CART algorithm. **Fig. 3** shows a plot of best, mean and median fitness values over the range of generations (0 - 100).

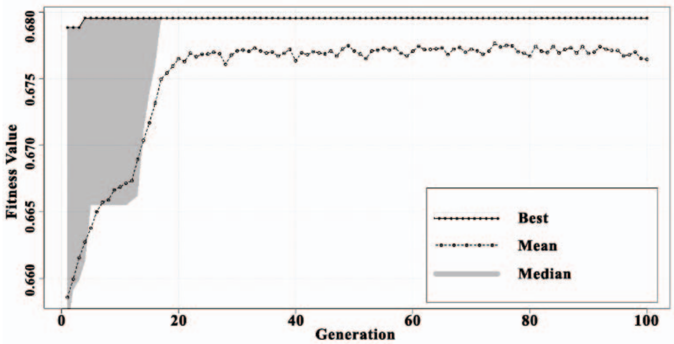

Fig. 3. Fitness values over generations in GA-CART.

From the Tables II and III, it is observed that GA-CART is performing best out of three models i.e., GB-CART, Pruned GB-CART, and GA-CART. For predicting injury, near-miss, and property damage, GA-CART has the highest precision which are equal to 0.7234, 0.6453, and 0.6677, respectively (Table II). Moreover, it is also observed that GA-CART outperforms others in terms of other performance parameters. The highest accuracy is observed in GA-CART which is equal to 69.98% and 67.96% in training and test data set, respectively (refer to Table III). From the result of better accuracy obtained from GA-CART, top six rules are extracted leading to incident categories. As per rule 1, if the description of incident corresponds to one of the topics desc2, desc4, desc5, desc6; Equipment damage score is very high; incident occurs in any division except Div 2, Div 11; and Injury potential score is high, medium or very low; then there is 84.30% chance of injury. Similarly, rule 6 says that if description of incident corresponds to one of desc1, desc3, desc7, desc8, desc9; and Equipment damage score is low or medium, there is 74.79% chance of Property damage. These rules provide the industry some useful insights of the situations leading to incidents and help decision makers to take the necessary steps to decrease the likely occurrences of incidents in future.

TABLE IV. RULES GENERATED BY GA-CART.

Rule no.	Splitting Conditions	Incident Category
1	1.) descr : desc 2,desc 4,desc 5,desc 6 2.) Equipment.Damage.Score: very high 3.) Division : any division except Div 2, Div 11 4.) Injury Potential Score: high, medium, very low	Injury (84.30 %)
2	1.) descr: desc 2,desc 4,desc 5,desc 6 2.) Equipment.Damage.Score : very high 3.) Division : Div 2, Div 11	Near Miss (59.12 %)
3	1.) descr: desc 1,desc 3,desc 7,desc 8,desc 9 2.) Equipment.Damage.Score : very high,very low 3.) Machine.Condition : M/C Idle,Not Applicable	Near Miss (64.76 %)
4	1.) descr : desc 2,desc 4,desc 5,desc 6 : 2.) Equipment.Damage.Score : high,low,medium,very low 3.) Primary.Cause : Material Handling, Process, Road Incident, Slip/Trip/Fall,Working at Height 4.) Injury. Potential.Score : very high	Property Damage (72.5 %)
–5	1.) descr : desc 1,desc 3,desc 7,desc 8,desc 9 2.) Equipment.Damage.Score : very high,very low 3.) Machine.Condition : M/C Working 4.) Primary.Cause : Electrical, Process, Road Incident, Slip/Trip/Fall, Working at Height	Near Miss (68.12 %)

6	1.) descr : desc 1,desc 3,desc 7,desc 8,desc 9 2.) Equipment.Damage.Score : low, medium	Property Damage (74.79 %)

VI. CONCLUSION

Prediction of occupational accidents is important goal in industry as it could help in deploying preventive measures for reduction of accidents. In this study, to build a predictive model, we have used GB-CART, pruned GB-CART, and GA-CART to search a better predictive model for occupational accidents. Results show that GA-CART outperforms others in terms of accuracy, precision, recall values. However, this study has some limitations. Experimentation with other techniques that could handle the problem of missing value is not performed. Data cleaning could be further done for obtaining better accuracy. Furthermore, the study used very limited data points for prediction. As future aspects, other predictive models like support vector machines, k-nearest neighbor, or other ensemble techniques could be used, or more data points could be collected and analysed for better obtaining prediction power as well as extraction of better quality rules that, in turn, could ensure the reduction of accidents.

ACKNOWLEDGEMENT

The authors are thankful to the safety management system of the steel industry for their support in this research.

REFERENCES

[1] "Safety and health at work", *Ilo.org*, 2016. [Online]. Available: http://www.ilo.org/global/topics/safety-and-health-at-work/lang--en/index.htm. [Accessed: 30- Jul- 2016].

[2] P. Hämäläinen, J. Takala and K. Saarela, "Global estimates of occupational accidents", *Safety Science*, vol. 44, no. 2, pp. 137-156, 2006.

[3] *pib.nic.in*, 2016. [Online]. Available: http://pib.nic.in /archieve/others/2014/feb/d2014022003.pdf. [Accessed: 30- Jul- 2016].

[4] "World Steel Association - Safety and health", *Worldsteel.org*, 2016. [Online]. Available: http://www. worldsteel.org/steel-by-topic/safety-and-health.html. [Accessed: 30- Jul- 2016].

[5] L. Breiman, H. Freidman, A. Olshen and J. Stone, "Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software", 1984.

[6] L. Chang and H. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques", *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 1019-1027, 2006.

[7] C. Cheng, S. Leu, Y. Cheng, T. Wu and C. Lin, "Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry", *Accident Analysis & Prevention*, vol. 48, pp. 214-222, 2012.

[8] A. Lukacova, F. Babic and J. Paralic, "Building the prediction model from the aviation incident data. In *Applied Machine Intelligence and Informatics (SAMI), IEEE 12th International Symposium on*", pp. 365-369, 2014.

[9] F. Pereira, F. Rodrigues and M. Ben-Akiva, "Text analysis in incident duration prediction", *Transportation Research Part C: Emerging Technologies*, vol. 37, pp. 177-192, 2013.

[10] S. Sarkar, V. Sammangi, & J. Maiti. (2016, March). "Text mining based safety risk assessment and prediction of occupational accidents in a steel plant". In *International Conference on Computational Techniques in Information and Communication Technologies (ICCTICT), IEEE*. pp. 439-444, 2016.

[11] T. Rivas, M. Paz, J. Martín, J. Matías, J. García and J. Taboada, "Explaining and predicting workplace accidents using data-mining techniques", *Reliability Engineering & System Safety*, vol. 96, no. 7, pp. 739-747, 2011.

[12] M. Bevilacqua, F. Ciarapica and G. Giacchetta, "Industrial and occupational ergonomics in the petrochemical process industry: A regression trees approach", *Accident Analysis & Prevention*, vol. 40, no. 4, pp. 1468-1479, 2008.

[13] G. Mistikoglu, I. Gerek, E. Erdis, P. Mumtaz Usmen, H. Cakan and E. Kazan, "Decision tree analysis of construction fall accidents involving roofers", *Expert Systems with Applications*, vol. 42, no. 4, pp. 2256-2263, 2015.

[14] M. E. Roberts, B. M. Stewart., and E. M. Airoldi, "A model of text for experimentation in the social sciences", *Journal of the American Statistical Association*, pp. 1-49, 2016.

[15] L. Guyon and A. Elisseeff, "An introduction to variable and feature selection. *Journal of machine learning research*", pp. 1157-1182, 2003.

[16] P. Tan, M. Steinbach and V. Kumar, *Introduction to data mining*. Boston: Pearson Addison Wesley, 2005.

[17] J. Holland, " *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*." U Michigan Press, 1975.

[18] Y. Park and I. Kweon, "Ambiguous Surface Defect Image Classification of AMOLED Displays in Smartphones", *IEEE Transactions on Industrial Informatics*, vol. 12, no. 2, pp.597-607,2016.