Computers and Operations Research 106 (2019) 210–224

Contents lists available at ScienceDirect

Computers and Operations Research

journal homepage: www.elsevier.com/locate/cor

Application of optimized machine learning techniques for prediction of occupational accidents

Sobhan Sarkar[a], Sammangi Vinay[b], Rahul Raj[c], J. Maiti[a,*], Pabitra Mitra[d]

[a] Department of Industrial & Systems Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
[b] Department of Mechanical Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
[c] Department of Electrical Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India
[d] Department of Computer Science & Engineering, Indian Institute of Technology Kharagpur, Kharagpur 721302, India

ARTICLE INFO

Article history:
Received 15 June 2017
Revised 11 February 2018
Accepted 28 February 2018
Available online 7 March 2018

Keywords:
Occupational accidents
Support vector machine
Artificial neural network
Genetic algorithm
Particle swarm optimization
Rule extraction

ABSTRACT

Although, the usefulness of the machine learning (ML) technique in predicting future outcomes has been established in different domains of applications (e.g., heath care), its exploration in predicting accidents in occupational safety domain is almost new. This necessitates the investigation of ML techniques in predicting accidents. But, ML-based algorithms cannot produce the best performance until its parameters are properly tuned or optimized. Moreover, only the selection of efficient optimized classifier may not fulfil the overall decision-making purposes as it cannot explain the inter-relationships among the factors behind the occurrence of accidents. Hence, in addition to prediction, decision-making rules are required to be extracted from the accident data. Considering the above-mentioned issues, in this research, optimized machine learning algorithms have been applied to predict the accident outcomes such as injury, near miss, and property damage using occupational accident data. Two popular machine learning algorithms, namely support vector machine (SVM) and artificial neural network (ANN) have been used whose parameters are optimized by two powerful optimization algorithms, namely genetic algorithm (GA) and particle swarm optimization (PSO) in order to achieve higher degree of accuracy and robustness. PSO-based SVM outperforms the other algorithms with the highest level of accuracy and robustness. Furthermore, rules are extracted by incorporating decision tree C5.0 algorithm with PSO-based SVM model. Finally, a set of nine useful rules are extracted to identify the root causes of the injury, near miss and property damage cases. A case study from a steel plant is presented in support of the proposed methodology.

1. Introduction

According to the International Labour Organization (ILO) estimation, globally about 2.3 million workers succumb to death annually due to occupational accidents and diseases which include approximately 3.6 lakh fatal accidents (Sánchez et al., 2011). Overall, nearly 337 million occupational accidents are reported per year. From ILO report, it is revealed that approximately 4% of the annual gross domestic product (GDP), which is equivalent to US $1.25 trillion, is drained off due to occupational accidents (ILO 2008). From EUROSTAT, it is reported that each year, 3.2% of workers in the European Union, i.e., EU-27 meet an accident at their working places (EUROSTAT 2009). In relation with this, ILO also makes the following comments: “Fatalities are not fated; accident do not just happen; illness is not random; they are caused” (ILO 2003). The basic causes of accidents are unsafe conditions or unsafe acts or both. There are multiple factors contributing towards an accident. There are many theories available in the literature that explain the causation of accidents. Khanzode et al. (2012) explained the various theories in their study behind the accidents such as accident proneness theory (Kunce, 1967), Domino theory (Heinrich et al., 1980), injury epidemiology (Haddon et al., 1964), system theory (Hale and Hale, 1970), sociotechnical system theory (Robinson, 1982), and macro-ergonomic theory (Hendrick, 1986). An injury event is occurred due to the presence of a chain of events or causal factors. If the causes are known, the outcomes (i.e., accidents) can be predicted. In addition, the predictive models will quantify the contribution of the various causal factors towards an accident to happen.

Predictive models for occupational accidents can be statistical learning-based or machine learning (ML)-based. Owing to the large amount of data available, ML supersedes traditional statistical counterpart in predicting future events, which has been used in

* Corresponding author.
E-mail addresses: jhareswar.maiti@hotmail.com, jmaiti@iem.iitkgp.ernet.in (J. Maiti), pabitra@cse.iitkgp.ernet.in (P. Mitra).

https://doi.org/10.1016/j.cor.2018.02.021
0305-0548/

various fields such as engineering, medical science, finance, and it renders very useful results (Witten et al., 2016). However, a review of literature shows that the ML techniques have been used in occupational accident analysis on a limited basis (Bevilacqua et al., 2008). So far, studies made on occupational analysis show the use of ML techniques in terms of their predictive power (Matías et al., 2008) and explanatory capacity (Martín et al., 2009). These methods, based on historical data from incident reports, or interview with employees, ensure their advantages over conventional statistics in terms of predictive functions and importance of predictors with a bearing on incident outcomes. The potential benefits of ML can not only be realized from the capability of processing large quantity of data but also from: (i) their capability to deal with large dimensional problems, (ii) their flexibility in reproducing the data generation structure irrespective of complexity, and (iii) their predictive and interpretative potential through the extraction of rules. Due to the capability of ML techniques, it has been used successfully in several domains including occupational accident analyses. However, the ML techniques do not produce good results if their parameters are not tuned. Optimization of parameters can provide better results. Usually, the concept behind the optimization is to search the optimal solution of the key parameter values that helps classifiers perform best on given dataset. Several studies in literature are available which show the utility of parameter optimization of ML techniques in different domains using genetic algorithm (GA), particle swarm optimization (PSO) and so on (Zheng et al., 2015).

Therefore, the primary objective of the present study is to develop a prediction model using machine learning techniques, namely support vector machine (SVM), and artificial neural network (ANN) for the prediction of occupational incident outcomes. In order to achieve the better accuracy, optimization techniques i.e., GA and PSO have been employed on the classifiers. In addition, rule extraction for the occurrence of injuries has been performed by the PSO-SVM-based classifier combined with decision tree (C5.0). The secondary objective includes the identification of the relevant variables attributable to incident outcomes using chi-square feature selection technique. The results of the analysis show the utility of the SVM classifier in terms of both prediction as well as rule extraction purposes.

2. Review of literature

In the domain of occupational accident prediction, there are many ML algorithms used such as SVM, ANN, extreme learning machine (ELM), and decision tree (DT). In the application of DT in accident analysis, some algorithms like C4.5, C5.0, classification and regression tree analysis (CART), Chi-square Automatic Interaction Detector (CHAID) etc. are usually used for prediction of occupational accident. The main aim to use DT is to predict and interpret qualitative and quantitative patterns lying in data, which leads to exploration of hidden information. Due to its relaxation on assumptions on distribution of attributes or independence of attributes, DTs have been successfully used in different fields like medicine (Oztekin et al., 2011), social sciences (Olson et al., 2012), business management (Aviad and Roy, 2011), construction engineering and management (Leu and Chang, 2013), process industry (Bevilacqua et al., 2008).

Other than DTs, algorithms like ANN, Bayesian classifier, adaptive neuro-fuzzy inference system (ANFIS), Bayesian network (BN), SVM, extreme learning machine (ELM) have been used in different domains like construction industry (Rivas et al., 2011), mining industry (Rivas et al., 2011), shipbuilding industry (Fragiadakis et al., 2014), service industry (Matías et al., 2008) etc. In 2008, Matias et al. used SVM, ELM (i.e., feed forward neural network), BN techniques for the analysis of causes and types of accidents like floor-level falls (Matías et al., 2008). They used 148 records obtained from different companies during 2003–2006 in Spain. As results, BN is found to be higher predictive capacity than others. Sánchez et al. carried out one study using SVM to classify those workers suffering work-related accidents for a year (Sánchez et al., 2011). They analysed the data consisting of 11,054 responses of the workers employed in all economic activities in Spain. Their findings show that SVM performs better than back-propagation neural network (BPNN) without over-fitting problems. In 2011, Rivas et al. modelled the accidents and incidents in two companies in construction and mining to identify the most important causes of accidents and developed predictive models using BN, SVM, and other ML techniques (Rivas et al., 2011). Bayesian network (BN)-based prediction model has also been used by many researchers in different sectors like mining (Matías et al., 2008), construction (Matías et al., 2008). For example, Sanmiquel et al. carried out one study in Spanish mining sector and analysed the 69,869 instances of occupational accidents during 2003–2012 using BN (Sanmiquel et al., 2015).

Another important prediction model used in this domain is ANN. Due to its important characteristics like the ability to learn from data, distributed memory, parallel operation and fault tolerance, it has been widely used in diverse fields of study along with in occupational accident domain. For instances, He et al. attempted to solve the problem of coal and gas outburst by classification technique using backward algorithm of ANN (BA-ANN) and exponent evaluation method (EEM) (He et al., 2010). Using BA-ANN, the weights of factors are calculated towards the response variables (i.e., coal and gas outburst). Yi et al. developed an early warning system for the workers in hot and humid environments using ANN (Yi et al., 2016). They have collected 550 data related to work, environment, and individuals which are analysed by ANN to predict the rating of perceived exertion (RPE) of the workers in the construction sites. Apart from the application of ANN in occupational accidents, there are plenty of literature available for other accidents of which research on road accidents is found to have attained more focus (Alikhani et al., 2013). More interestingly, artificial intelligence (AI) approaches like ANN are found to have greater performance in terms of prediction than regression analysis. Reviewing the literature on accident analysis domain, techniques like SVM and ANN are found to be popular and useful as they have a robust theoretical grounding that enables the successful learning from the data, capability of handling any level of complexity except computational complexity of the problem and flexibility with non-parametric philosophy.

However, all these machine learning algorithms do not provide optimal results like classification accuracy and understandability if the parameters of them are not properly tuned. To tune the parameters of the classification algorithms, optimization methods are found to be most useful than other techniques like manual tuning or grid search. In occupational accident research, hardly any study or no study has been reported that uses optimization techniques on classifiers (like SVM, or ANN) in order to obtain better classification accuracy. From the research of the other domain, it is observed that in order to get enhanced accuracy in SVM model, penalty factor (c), and kernel parameter (γ) are considered to be optimized (Chou et al., 2014). There are many optimization techniques used for this purpose like GA, PSO, gradient descent method, etc. (Pham and Triantaphyllou, 2011). Of them, GA, and PSO are found to be the most popular methods used for optimizing the parameters of classifiers (e.g. SVM) to achieve higher accuracy (Cervantes et al., 2017). Therefore, in this paper, GA, and PSO have been selected for parameter optimization of SVM. Similarly, for ANN, there exist several parameters which can be optimized like the number of layers, input and hidden neurons, type of transfer functions, topology of ANN, weights, thresholds

etc. Li et al. used initial parameters, network topology, weights, and thresholds of back-propagation neural network (BPNN) based on memetic algorithm with GA (Li et al., 2015). Xue and Liu used only initial weights and threshold values for BP model for predicting liquefaction susceptibility of soil (Xue and Liu, 2017). Das et al. focused on optimizing the weights, transfer functions, and topology of ANN for channel equalization (Das et al., 2014). Most of the studies showing the importance of optimization techniques on classifiers have been carried out to solve the problems in different domains other than the occupational accident. Hence, it is required to implement such useful optimization techniques on classifiers to improve their performance like predictive accuracy.

However, the performance of the classifiers not only depends on optimized parameter values, but also the types of data used. As it is a known fact that numerical attributes hold more information than categorical attributes or free-text attributes, thus dealing with different data types also impacts classifier's performance. Hence, it is really a challenging task for researchers to extract the pattern from different types of data like categorical or more specifically, free-text data. Most of the literature in accident domain used either numerical data or categorical data for the analyses of accident scenarios. However, analysis of free-text data remains underutilized in most of the cases as it is really hard task to extract the pattern from the passage of free-text. Narrative text is one of the key resources for the prediction of accident. It provides the valuable additional information in the analysis along with other types of data. To investigate the importance of narratives in prediction of occupational accidents, Jones & Lyons showed increase of home injuries identified by 19%, rugby injuries by 137%, and assaults by 26% (Jones and Lyons, 2003). Li & Guo tried to analyse the aviation safety data with the help of topic modelling techniques (Li and Guo, 2015). Related to this, a noteworthy contribution made by Brown is to analyse rail accident data to explore the main contributors behind the accident using text mining associated with other techniques like Latent Dirichlet Allocation (LDA), and Random Forest (RF) (Brown, 2016). Thus, the main challenge lies in the analysis of the unstructured text. To tackle the issue, Tixier et al. tried to develop a system that could overcome the problem by decoding the unstructured reports from accident database (Tixier et al., 2016). The system developed by them could use the unstructured injury database with 101 attributes, and produced with 95% classification accuracy. Vallmuur, therefore, mentioned in his study that future research on injury analysis would direct a continued growth and advancement in the application of text mining for utilizing information within the text (Vallmuur, 2015). The primary difficulty in the analysis of free unstructured text is the sparsity and high-dimensionality of document-term matrix. Moreover, text mining-based approach cannot capture the order of words and semantic meaning of them. One recent study by Pavlinek & Podgorelec shows that topic modelling of free-text could help in text classification task and reduction of sparsity (Pavlinek and Podgorelec, 2017). The study by Niraula et al. revealed the superiority of topic modelling in a supervised setting (Niraula et al., 2013). Consequently, many classification algorithms have been implemented using topic modelling in different domains. In road accident analysis, one study by Pereira et al. used topic modelling of incident reports of traffic to extract information in real time to predict incident duration (Pereira et al., 2013). Therefore, topic modelling of under-utilized free unstructured text has full potential in the extraction of latent information within text field that facilitates the prediction of occurrence of accidents.

Prediction analysis, as a standalone tool, may not serve the entire purpose of the accident analysis until a prescriptive analysis is not made for the interpretation of the accident causation. Rule extraction and its interpretation from the accident dataset are often considered to be an effective approach. The rules can generally be obtained by using either decision tree (DT), or by association rule mining (ARM) approach. In several studies of occupational accident, DT has been used for rule extraction and interpretation more than ARM. DT is found to be useful when target function is discrete-valued, when it is describable by attribute-value pairs, or when the datasets are noisy. DT works well in rule extraction when the dataset used for DT analysis are more informative than others. Otherwise, other datasets having less information might lead to generation of low-quality rules. Therefore, selection of the set of data, which is informative, is required for better rule building and rule interpretation. Some of the previous studies showed that the rule extraction based on support vectors identified by SVM is useful as the rules are less in number, and interpretable (Han et al., 2015). DT algorithms with SVM have also been attempted to make the black box nature of SVM-based decisions into transparent and comprehensible rules which can be utilized as secondary opinion for any decision-making task.

Based on the above-mentioned literature, it is found that none of the previous literature in occupational accident domain has reported to use text data and categorical data together for the building of prediction model. Moreover, to the best of authors' knowledge, optimization techniques on classification algorithms to get optimal solutions have not been also addressed by any researchers previously in this domain. Another important point to be noted is that very few studies have been conducted so far for the prevention of accidents in steel industry, whereas previous research focused more on either construction, or mining industry. Therefore, there remains a strong need for research on prevention of occupational accident using machine learning techniques for steel industry.

2.1. Research issues and contribution of the study

Based on the review of literature presented above, the following issues have been identified in the domain of prediction of occupational accident analysis.

(i) None of the previous studies reported have shown the combined analysis using text and non-text attributes together for incident prediction.
(ii) None of them reported the parameter optimization of the classifiers for better prediction accuracy.
(iii) There are no studies reporting the SVM-based rule extraction for incident occurrences.

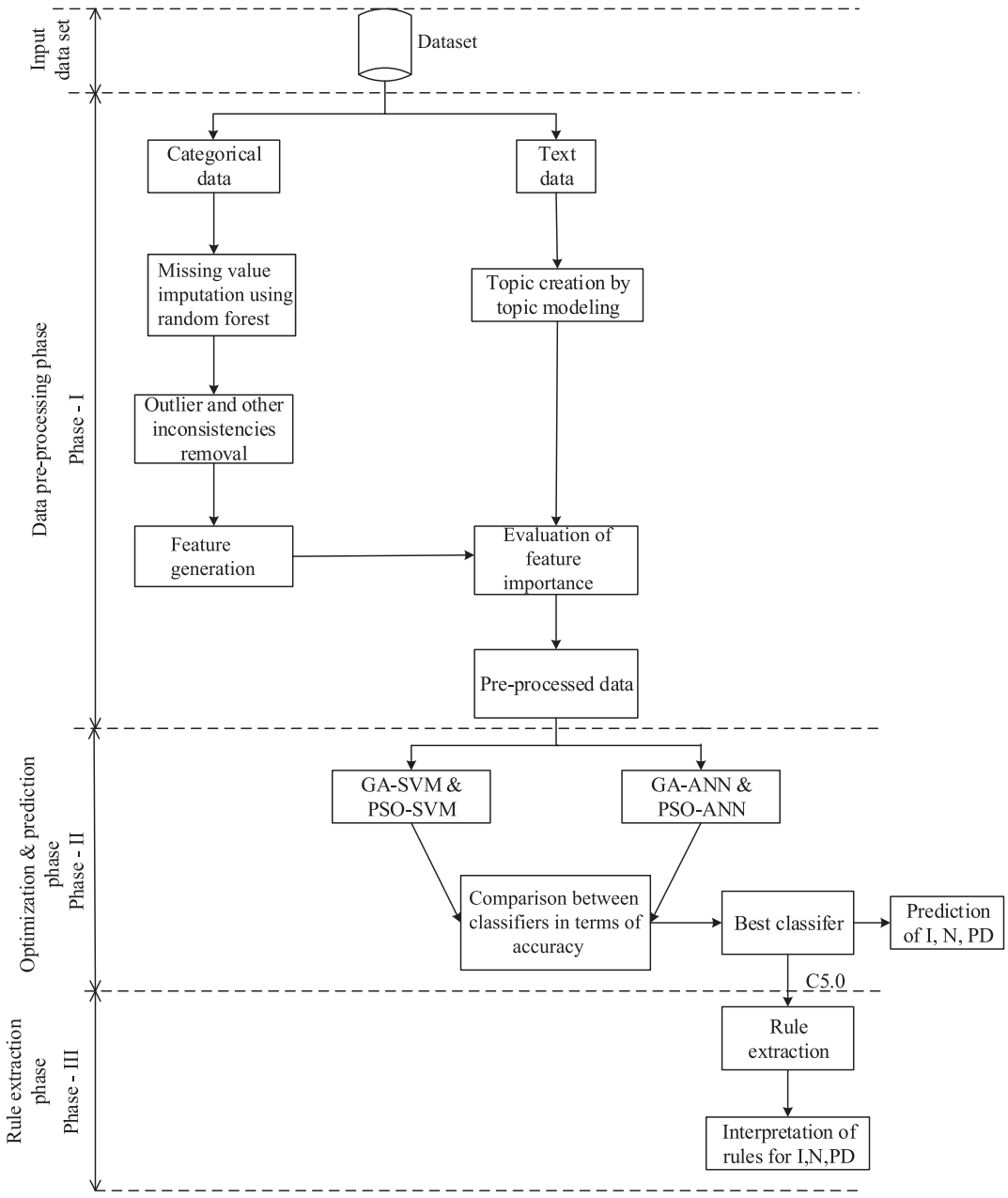
Realizing the issues in accident literature, our study, therefore, endeavours to contribute in the following ways:

(i) The study takes care of text and categorical attributes together for predicting the incident categories,
(ii) It uses optimization algorithms for parameter optimization of the classifiers for improved prediction accuracies,
(iii) It includes SVM-based rule extraction method for injury, near miss, and property damage cases
(iv) It identifies the importance of the predictors towards occurrence of incidents, and
(v) The developed methodology is validated with a case study in a steel plant.

The rest of the paper is organized as follows: Section 3 describes the methods used in this study; Section 4 presents the case study with data set and data pre-processing tasks; in Section 5, results and discussion are presented; and finally, conclusions with future scopes of the present study are discussed in Section 6.

3. Methods

In the methodological section, topic modelling, SVM, ANN, GA, PSO, and PSO-SVM combined DT-based rule extraction methods



Note: I: Injury; N: Near miss; PD: Property damage

Fig. 1. Proposed methodological flowchart of the study.

are discussed briefly. The total proposed methodological flowchart is depicted in Fig. 1. There are three important phases shown in the flowchart. They are: (i) *Data pre-processing phase:* In data pre-processing, three important tasks, namely feature addition, missing value handling, and evaluation of feature importance are performed on the dataset. The initial dataset has 3308 incident records and 16 attributes (15 categorical and one text) with a very low percent of missing values in three of them. Four categorical attributes, which are found to be interrelated in nature, are combined into one new attribute. In addition, a new attribute or feature is generated from text data using topic modelling technique which will be discussed in subsequent sections in details. Thereafter, missing value imputation has been done using random forest. Finally, feature importance is calculated. The final dataset generated after this phase has 1500 records and 13 attributes (all categorical) without any missing value; (ii) *Optimization & prediction phase:* In this phase, optimization techniques, namely GA and PSO have been implemented on two classifiers, namely SVM and ANN using 10-fold cross validation. Then, the classifier with the highest accuracy is considered as the best one; and finally (iii) *Rule extraction phase:* In this phase, useful rules from the best classifier, i.e., PSO-SVM combined with C5.0 decision tree are extracted. All the processes are illustrated in following sections.

3.1. Topic modelling

In machine learning and natural language processing (NLP), a topic model can be described as a type of statistical model to extract the "abstract", “topic” or “classes”. In topic modelling, latent Dirichlet allocation (LDA) is a very popular approach. In

order to use LDA, the number of topics is required to be fixed. There are several metrics developed by researchers to select the optimal number of topics for LDA model. We used four of those metrics in our study. 'Metric1', developed by Griffiths & Steyvers (Griffiths and Steyvers, 2004), shows that the number of topics for which log-likelihood of the data becomes maximum is considered to be optimal. Cao et al. (2009) developed 'Metric2' where they have used average cosine distance between every pair of topics to measure the stability of topic structure. It was observed that smaller the average distance, better the stability. Similarly, 'Metric3' has been developed by Arun et al. (2010). The measure is computed in terms of symmetric Kullback–Leibler (KL) divergence of salient distributions that are derived from these matrix factors. It was also observed that the divergence values become the lowest for the optimal number of topics. Recently, another metric 'Metric4' has been developed by Deveaud et al. (2014). They proposed a simple heuristic that estimates the number of latent concepts of a user query by maximizing the information divergence between all pairs of topics of LDA. So, when put together, in order to find optimal number of topics, Metric2 & Metric3 should be minimized and Metric1 & Metric4 should be maximized. The detailed description of basic principle of topic modelling and its applications are presented in Pereira et al. (2013).

3.2. Support vector machine (SVM)

SVM, developed by Vapnik (1995), is an emerging machine learning technique in statistical learning theory of multi-dimensional function which is used for classification and regression analysis. It holds an ability of being universal approximators of any multivariate functions to any desired level of accuracy. Initially, it was developed for regression tasks, but later was used as a powerful classifier. According to the previous studies (Wei et al., 2013), SVM has been used in different engineering fields with good accuracy. Theoretically, it has less overfitting problem and better generalization ability. However, the main problem encountered in constructing SVM model is to adequately select training parameter values as inappropriate parameter setting leads to poor prediction accuracy. The readers may refer to Kecman (2005) for basic understanding of the working principle of SVM.

3.3. Artificial neural network (ANN)

ANN is an artificial model of the human brain which can learn through adapting the present situations. It consists of an interconnected network of neurons and synapses. Usually, it has three layers (i.e., input, hidden and output) or more (when more than one hidden layer). Hidden layers are considered the root of all calculations in ANN. A network gets activated when a set of inputs are triggered that consequently produce desired results through output layers. Each input value is multiplied by its corresponding weight layers, then it is summed up and added to a scalar parameter called bias, which in turn generates output through final output layer. Modifying connection weights and biases using appropriate learning algorithm, training process can be accomplished. Many evolutionary algorithms or gradient descent methods have been used in this training process by updating weights and biases. At each iteration, they are modified until prediction error of the network gets minimized. Out of many learning algorithms, back propagation (BP), which is a gradient type of adjustment for the modification of weights, has been used in the paper of Bengio et al. (2017). Basically, the output of any node is determined by a mathematical operation on the input of the particular node. This operation is called the transfer function which facilitates the transformation of inputs into output either in linear or non-linear manner. There are three types of transfer functions used commonly in literature i.e., sigmoid, hyperbolic tangent, and linear. In this paper, sigmoid transfer function has been used.

3.4. Working principles of GA and PSO on classifiers

In this section, the two optimization algorithms i.e., GA and PSO have been described briefly on how they are used to optimize the parameter values of the classifiers, namely SVM and ANN. For detailed description of GA and PSO, interested readers are requested to go through (Holland, 1975; Kennedy and Eberhart, 1995). For GA, initial population is generated at random. Then the data is split into training and test sets. The fitness function is developed by which the fitness value i.e., accuracy of the classifier is computed for each chromosome for each iteration. Then, the criterion for termination of the algorithm is checked. Here, we have used the maximum number of iteration (i.e., 500) as termination criterion for both GA and PSO operations. If it is not satisfied, it goes for crossover, and then mutation process which ultimately creates a new population. Recursively, this process continues until the termination criterion gets satisfied. Once it is satisfied, optimal parameter values for both SVM and ANN are achieved which will be ultimately used for model building for prediction of incident outcomes (see in Fig. 2). Similarly, in Fig. 3, the process of optimization of SVM/ ANN parameters using PSO has been depicted.

3.5. C5.0

C5.0 is a decision tree algorithm developed from C4.5. In C4.5, during, all training samples are set as the root of the decision tree. Then, the gain information ratio of every feature is calculated based on the entropy of the feature, and the feature with the highest information gain is selected to split the data into multiple subsets. The algorithm repeats this procedure on each subset until all instances in the subset belong to the same class and a leaf node is created. For detailed understanding, interested readers may refer to Breiman (2001).

3.6. SVM and DT-based rule extraction

SVMs and ANN have shown better performance than other machine-learning algorithms in some application areas, such as speech recognition, computer vision, and medical diagnosis. Although SVM and ANN have an inherent inability to explain models and results, these algorithms construct black box models and learn patterns with no transparency and comprehensibility to humans. This drawback of these models impedes their application in some areas. Therefore, addressing this issue, in this study, we tried to extract meaningful rules from SVM to interpret the models. In literature, a proliferation of rule-extraction methods for trained SVMs has been proposed. Fu et al. have classified these motifs into three basic categories: "decomposition" (or transparent), "pedagogic" (or learning based), and "eclectic" (or hybrid) (Fu et al., 2004). The transparent approach focuses on extracting region-based rules by support vectors (SVs) and separating hyperplane. For instance, Núñez et al. proposed the SVM and prototype method and utilized the defined regions (ellipsoids and hyper-rectangles) to refine the rules (Núñez et al., 2002). Zhang et al. proposed the hyper-rectangle rule extraction (HRE) algorithm (Zhang et al., 2017), and Fung et al. suggested a linear programming formulation approach for rule extraction from linear SVMs (Fung et al., 2005). By contrast, the pedagogical approach treats SVMs model as a black box and uses the generated model to predict the label (class) for an extended data or unlabelled data. Barakat and Diederich used the resulting patterns to train a decision tree learning system and to extract the corresponding

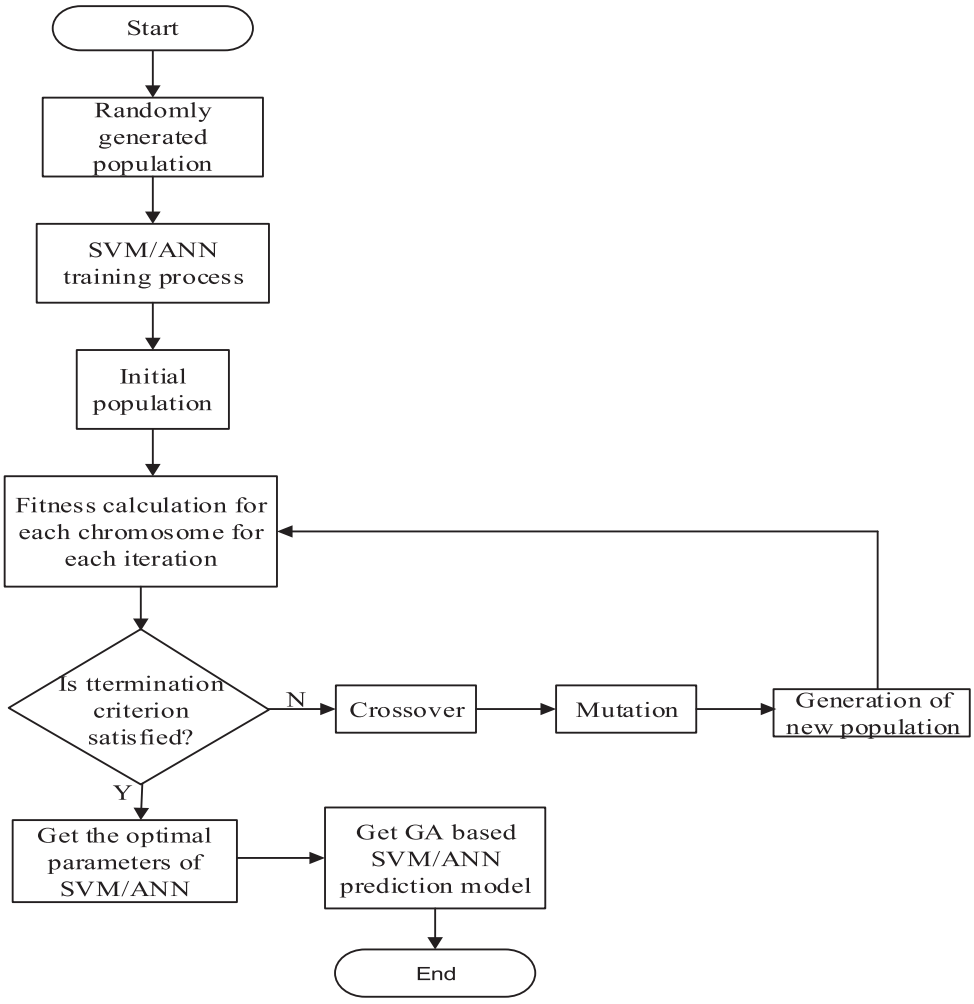


Fig. 2. Flowchart of GA optimized SVM/ANN.

rule sets (Barakat and Diederich, 2004). The eclectic approach incorporates both "decomposition" and "pedagogic" techniques; it only uses the SVs or applied rule-based model to train the artificial data based on SVs. Barakat & Bradley proposed a SQRex-SVM algorithm based on the sequential covering approach (Barakat and Bradley, 2007). The proposed method extracts rules directly from the SVs of a trained SVM using a modified sequential covering algorithm. It was observed that the proposed method exhibited both improved generalization performance and smaller as well as comprehensible rule sets compared to both other SVM rule extraction techniques and direct rule learning techniques. Therefore, the rule-extraction approach using PSO-SVM used in this study is performed in two basic steps illustrated below.

(i) Step 1: Generation of artificial data using support vectors

In this step, the training data are applied to build an SVM model with acceptable accuracy by finding the optimal parameters using PSO. In order to obtain a set of rules, the trained SVM model is used to provide class label y_i. Then, SVs are extracted and predicted by the obtained SVM model, and the predicted labels of SVs will replace the original labels of SVs to generate the artificial dataset (Han et al., 2015). The motivation of changing labels here is to ensure the future generated rules can mimic the predictions of SVMs as closely as possible. The idea behind this technique is the assumption that the trained SVM model can better represent the patterns than the artificial dataset. By changing the class labels of the data, some noises like class overlap are removed from the data.

(ii) Step 2: Rule generation by C5.0

C5.0 algorithm is applied on the artificial data generated in Step 1, and the best rule sets are generated. Thereafter, the rules are evaluated by performing 10-fold cross validation on artificial dataset. The proposed algorithmic flowchart is depicted in Fig. 4. Each rule generated by this process has two parts, i.e., n and (n/m) where n is the number of training instances covered by the rule and m is the number of instances in n that do not belong to the class predicted. Confidence is the estimated accuracy and can be calculated as $Confidence = \frac{(n-m+1)}{(n+2)}$; whereas lift can be calculated as the ratio of Confidence to the relative frequency of the class predicted in the entire dataset. The entire method of rule generation using SVM and DT is displayed in Fig. 4.

4. Case study

4.1. Data collection and data description

The accident data used in this study was collected from the electronic safety management system database of an integrated steel plant in India during the period from 2010 to 2013. A total of 1500 occupational accident records have been used in this study. The original dataset was pre-processed and restructured by deleting some of the attributes of no importance after discussing with experts. The dataset consists of sixteen attributes (15 categorical and one free-text) of which the attribute "incident outcomes" is considered as the response variable. A brief description of each of the attributes is given below.

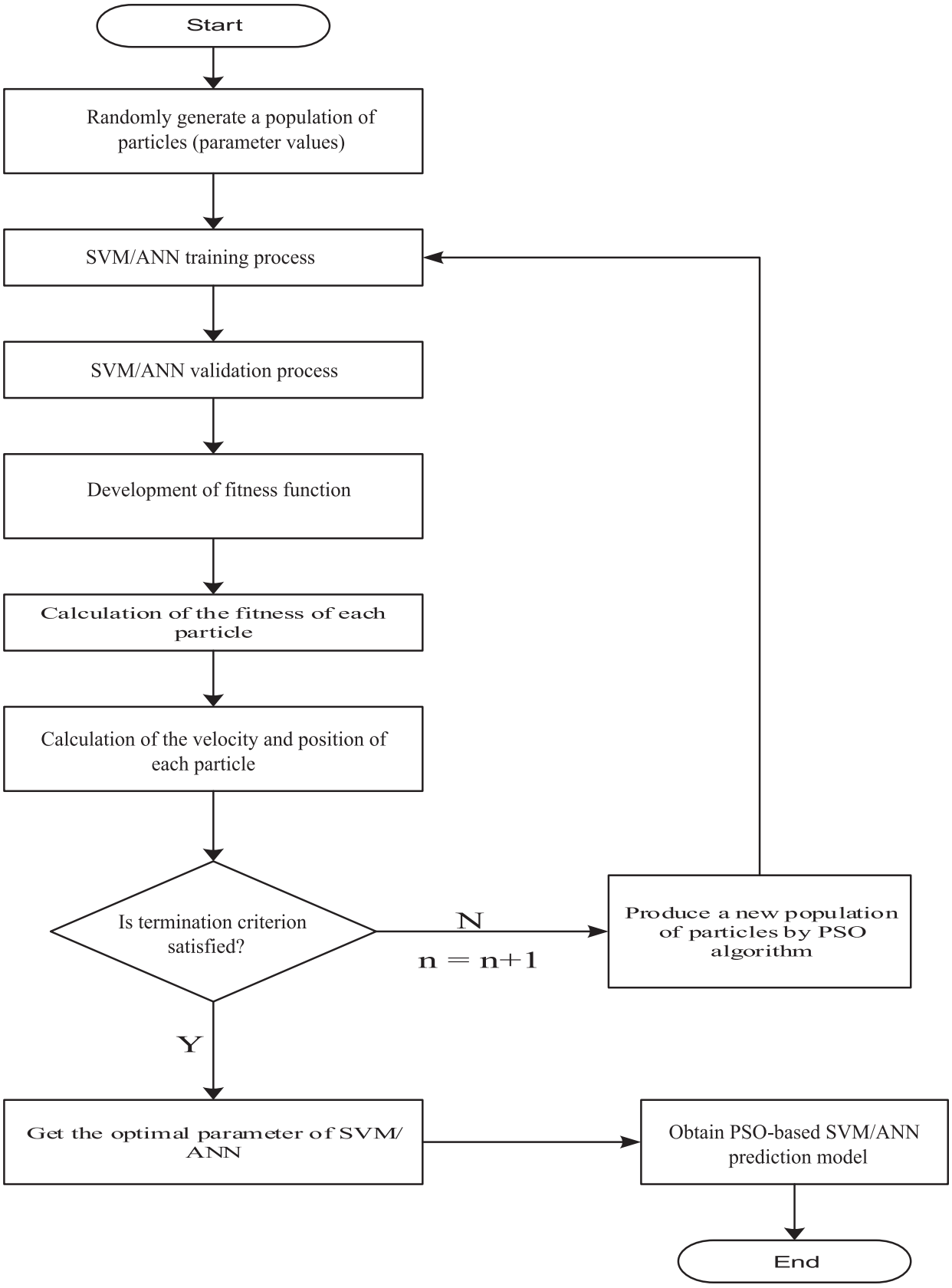


Fig. 3. Flowchart of PSO optimized SVM/ANN.

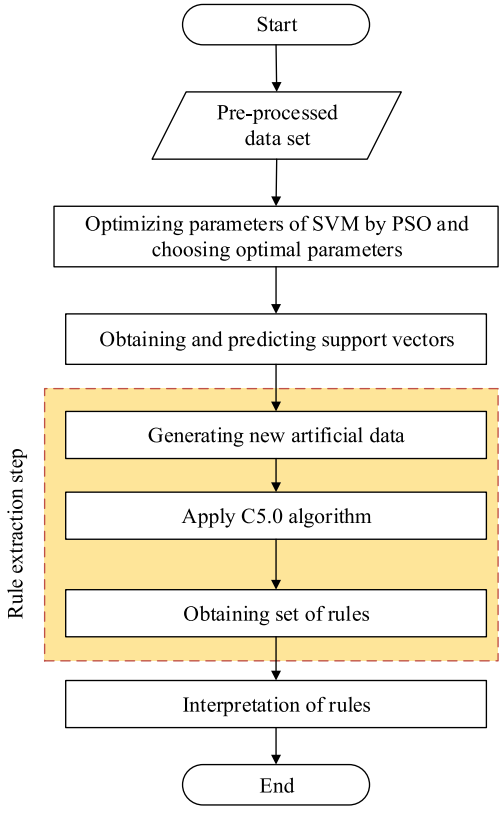


Fig. 4. Algorithmic flowchart of PSO-SVM and DT-based rule extraction process from accident data.

(i) *Day of incident (DOI):* This attribute implies the day on which the incident occurred. There are seven categories in it, namely Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, and Sunday.

(ii) *Month of incident (MOI):* This attribute implies the month in which the incident occurred. There are twelve classes in it, namely January, February, March, April, May, June, July, August, September, October, November, and December.

(iii) *Divisions (Div):* This represents the location where the incident was taken place. In total, thirteen divisions were considered, namely Div1, Div2, Div3, Div4, Div5, Div6, Div7, Div8, Div9, Div10, Div11, Div12, and Div13.

(iv) *Incident outcome (IO):* It is the outcome variable. It has three different classes: (i) injury (I)- when someone gets injured physically by an incident; (ii) near miss (N)- when someone is narrowly escaped from an incident having full potential to cause injury or damage; and (iii) property damage (PD)- when there is damage to private or public property, due to the incident

(v) *Incident event (IE):* This attribute refers as the top primary event. It has 23 classes. They are 'crane dashing (CD)', 'dashing/ collision (DC)', 'derailment (D)', 'electric flash (EF)', 'energy isolation (EI)', equipment/ machinery (EM), 'fire/ explosion (FE)', 'gas leakage (GL)', 'hot metals (HM)', 'hydraulic/pneumatic (HP)', 'lifting tools & tackles (LTT)', 'process incidents (PI)', 'rail (R)', 'road incidents (RI)', 'run over (RO)', 'skidding (S)', 'slip/trip/fall (STF)', 'structural integrity (SI)', 'toxic chemicals (TC)', and 'working at heights (WAH)'.

(vi) *Type of injury (IT):* This attribute represents the type of injury. It has 12 classes. They are 'claim injury on duty (IOD)', 'claim injury at work (IOW)', 'death', 'exgratia', 'fatal', 'first aid', 'foreign body', 'IOW', 'Injury type not applicable (ITNA)', 'normal', 'restricted work cases (RWI)', 'serious injury'. Note that for near miss and property damage cases, type of injury is termed as ITNA.

(vii) *Working Condition (WC):* This attribute represents the condition of work when the incident took place. It has three categories i.e., 'group working (W1)' representing the condition where people work in groups, 'single working (W2)' representing person working alone, and 'Others (W3)' representing situations when no workers were present.

(viii) *Machine Condition (MC):* It implies the condition of the machine when the accident took place; either machine is in 'idle condition (M1)' or in 'running condition (M2)', or 'others (M3)' i.e., not related to the machine.

(ix) *Observation type (OT):* This attribute represents the basic causes of incident and has four categories as; (i) 'unsafe act (OT1)' representing the person himself is responsible for the cause of incident, (ii) 'unsafe act and unsafe condition (OT2)' representing the incident occurred due to presence of both the factors, person's fault and hazardous condition, (iii) 'unsafe act by other (OT3)' representing the incident occurred due to the other's fault, and (iv) 'unsafe condition (OT4)' representing a situation which is likely to cause incidents.

(x) *Employee Type (ET):* This attribute has two classes in it namely, 'Employee' and 'Contractor'.

(xi) *Incident type (IT):* This attribute represents whether an accident happened is due to 'human behaviour (IT1)', or 'process type (IT2)' which is non-human fault.

(xii) *SOP Adequacy (SOPA):* Standard Operating Procedure (SOP) implies a procedure/ guideline to be followed while performing tasks by the workers/ operators. It has two categories: (i) 'SOP adequate (SOPA1)' – sufficient in quality and quantity; (ii) 'SOP inadequate (SOPA2)' – not sufficient in quality and quantity.

(xiii) *SOP compliance (SOPC):* This attribute indicates whether any SOP was 'followed (SOPC1)', or 'not followed (SOPC2)'.

(xiv) *SOP Availability (SOPAv):* This attribute implies that whether any SOP was 'available (SOPAv1)', or 'not available (SOPAv2)'.

(xv) *SOP Requirement (SOPR):* This attribute represents that whether any SOP was 'required (SOPR1)', or 'not (SOPR2)'.

(xvi) *Brief description of incident (BD):* This attribute consists of short description of how and why the incident occurred. The field contains free text logged by safety personnel after the incident was investigated.

4.2. Data pre-processing

Data pre-processing is an essential task of data mining. On an average, it consumes more than 60% of total effort in the entire modelling process (Houari et al., 2016). Prior to any analysis, dataset should be pre-processed or cleaned otherwise it leads to sub-standard, erroneous or misleading results of analysis due to the existence of outliers or redundant or missing values in the data, as it is a known fact that "no quality data, no quality results" (Houari et al., 2016). In the data-set used for our analysis, there were missing values, outliers, or other inconsistencies. First, proper missing data imputation technique has been applied on the data to overcome the problem. Then, new features are generated from both the categorical and text data. Finally, feature importance has been shown towards the prediction of incident outcomes. All the processes have been discussed in Section 3.

5. Results and discussion

5.1. Feature generation

In the dataset obtained from the electronic database, there are certain attributes present which can be merged into another

Table 1
Top eight terms across each topic and extracting a meaningful event from them.

Topic	Top eight terms	Meaningful event
1	0.051*person + 0.045*one + 0.044*hit + 0.019*remove + 0.019* piece + 0.016*injury + 0.014*take + 0.014*roof	Hitting by foreign body
2	0.054*operate + 0.043*crane + 0.022*roll + 0.019*coil + 0.019* place + 0.018*lift + 0.016*work + 0.016*move	Crane operation failure
3	0.039*fell + 0.029*work + 0.025*fall + 0.019*ground + 0.015* due + 0.015*level + 0.015*plate + 0.014*floor	Falling from heights
4	0.044*fire + 0.031*cable + 0.024*damage + 0.02*excavate + 0.017*power + 0.016*work + 0.015*site + 0.014*weld	Fire incidents
5	0.041*shift + 0.034*left + 0.031*first + 0.03*aid + 0.024*leg + 0.021*near + 0.02*duty + 0.019*plant	First aid incidents
6	0.046*area + 0.028*material + 0.019*belt + 0.018*end + 0.018* conveyer + 0.016*engage + 0.015*clean + 0.013*around	Incidents during cleaning
7	0.03*job + 0.029*due + 0.029*pipe + 0.024*gas + 0.023*line + 0.023*water + 0.019*open + 0.016*came	Pipe leakage
8	0.044*side + 0.028*load + 0.025*dumper + 0.023*driver + 0.021*road + 0.019*vehicle + 0.018*gate + 0.016*toward	Vehicle hitting/collision
9	0.068*got + 0.067*hand + 0.051*injury + 0.043*right + 0.034* cut + 0.033*finger + 0.027*slip + 0.027*left	Slipping

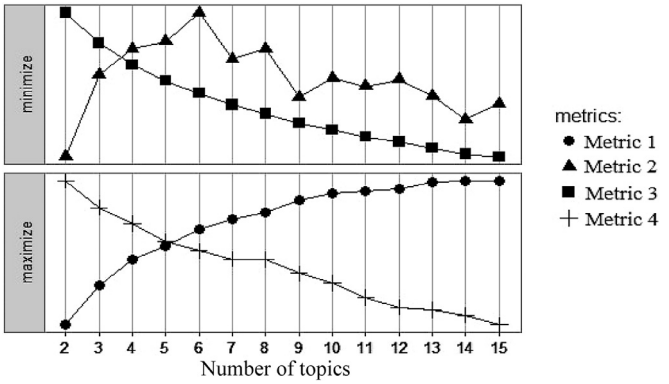


Fig. 5. Several metric distributions over the number of topics for the brief description of incident.

attribute with no information loss. In the categorical set, there are four attributes namely 'SOP Adequacy', 'SOP Compliance', 'SOP Availability', and 'SOP Requirement' which are merged into one attribute, namely 'SOP_combined'. In non-categorical set, there is a free-text attribute, namely "Brief description of incidents (BD)". From the field of free-text, topics are generated using topic modelling that helps us to utilize the information in text data. These two processes are described below.

(a) *From Categorical data:* Based on the discussion with experts, it was found that the attributes 'SOP Adequacy', 'SOP Compliance', 'SOP Availability', and 'SOP Requirement' are interrelated. So, these features are removed and a new feature called 'SOP_combined (SOPC)' with six inherent labels is added without any loss of information.
(b) *From Text data:* In the dataset, text attribute called 'BD' consists of accident narratives. To utilize the maximum information within the passage of unstructured text, LDA topic modelling has been used. Topic modelling is a frequently used text-mining tool for discovery of hidden semantic structures in a text body. The "topics" produced by topic modelling techniques are clusters of similar words. A topic model captures this intuition in a mathematical framework, which allows examining a set of documents and discovering, based on the statistics of the words in each, what the topics might be and what each document's balance of topics is.

In Fig. 5, four metrics used to find the optimal number of topics in the corpus of 'BD' in accident dataset are displayed. Since, to select the optimal number of topics, two metrics (i.e., Metric 2 and 3) should be minimum and the other two (i.e., Metric 1 and 4) should be maximum. It is found out that the optimal number of topics from LDA topic modelling is 9 for the best result. So, a new attribute 'Topic' having nine classes is added to the dataset. This is used in place of attribute 'BD'. Table 1 shows the extraction of a meaningful event from the set of terms under each of the topics. Here, the number of terms with a higher probability of occurrence for each topic is kept as eight. For example, in Topic1, the top eight terms were found to be 'person', 'one', 'hit', 'remove', 'piece', 'injury', 'take', and 'roof'. From these terms, it can be inferred that Topic1 can be described as 'Injury due to foreign body hitting the person'.

5.2. Missing value imputation

In the dataset, the attributes 'Machine Condition' and 'Employee Type' have 7.40% and 5.73% missing values, respectively. Also, the attributes 'SOP Adequacy', 'SOP Compliance', 'SOP Availability', and 'SOP Requirement' have 6.93% missing values in each of them. According to the studies done by Solaro et al. (2017) and Liao et al. (2014), RF is very useful for missing value imputation in the data with complex and non-linear relationships. Therefore, in this study also, RF has been used for missing data imputation.

5.3. Feature importance

In the stage of feature selection, chi-square, a statistical test of independence to determine the dependency of two categorical variables, has been used. Now, chi-square statistic can be calculated between each of the predictor variables and the target variable and the existence of a relationship between them can be observed. If the target variable is independent of the predictor variable, the predictor variable is discarded. If they are dependent, the predictor variable is important. Fig. 6 shows the importance of variables using chi-square test. The attributes 'injury type' and 'day' are found to be the most important and the least important predictors towards incident outcomes. It is also noteworthy that newly generated attribute 'Topic' is also revealed to be one of the important predictors. At the end of data pre-processing stage, we have thirteen attributes with no missing values. The pre-processed data can be used for the next step of analysis i.e., predictive analysis which results are discussed in the next section.

5.4. Predictive analysis

In this study, two classifiers namely SVM and ANN were used to predict the incident outcomes. Two optimization algorithms namely GA and PSO were used to optimize the parameters of the two classifiers. For SVM, the two parameters, 'cost (C)' and 'gamma (γ)' were considered. Initially, the suitable values of parameters of GA namely 'population size', 'number of generations', 'crossover probability', 'mutation probability', and 'elitism' were obtained using parametric study (refer to Table 2). Using these suitable values, the parameters 'cost' and 'gamma' of SVM were tuned for 500 iterations with 10-fold cross validation for each iteration, and the best accuracy value was recorded for every iteration (refer to Fig. 7). The ranges of 'cost' and 'gamma' were set as (0.25–128) and (0.0078–2), respectively. Over the 500 iterations, it is observed that the GA-SVM produces the best accuracy (i.e., 90.53%) at

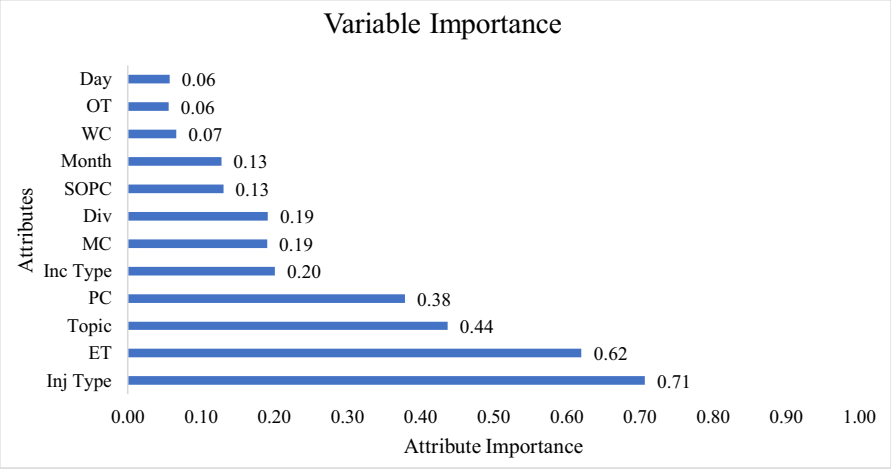


Fig. 6. Variable importance plot using chi-square technique.

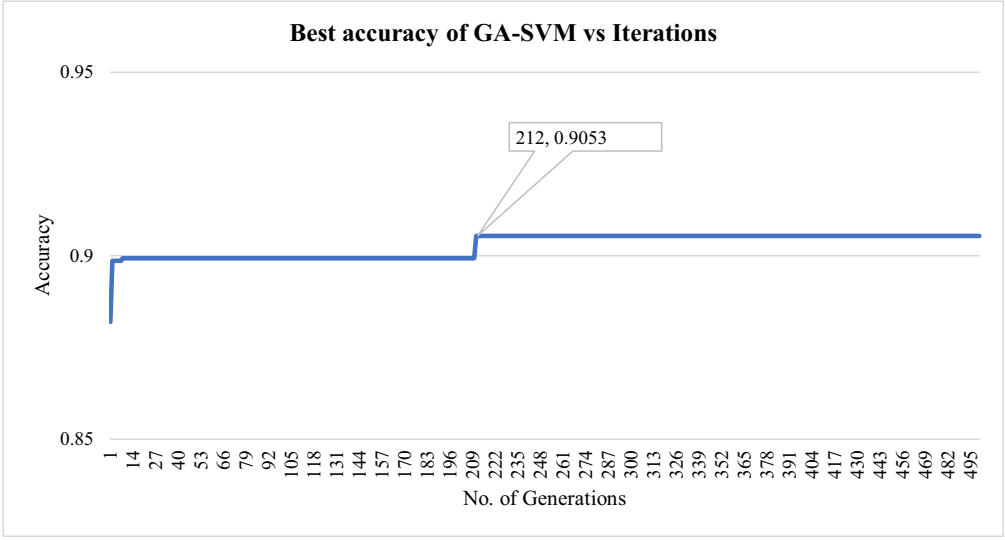


Fig. 7. Plot of best accuracy in each iteration of GA-SVM.

Table 2
The utilised GA parameters.

SL	GA parameter	Value
1	Population size	12
2	Number of generations	500
3	Crossover probability	0.8
4	Mutation probability	0.1
5	Elitism	0.05

iteration 212 with optimal values of 'cost' and 'gamma' as 1.1093, and 0.2474, respectively (refer to Table 3).

Similarly, the suitable values of parameters of PSO, namely 'the number of generations', 'swarm size', 'exponent for calculating number of informants', 'exploitation constant', 'local exploration constant', and 'global exploration constant' were obtained using parametric study (refer to Table 4). Over the 500 iterations, each with 10-fold cross-validation, using the suitable values of PSO, the PSO-optimized SVM produces best accuracies of 90.67% (refer to Fig. 8 and Table 3).

Similarly, for GA-optimized ANN, the three parameters of ANN, i.e., 'number of hidden layers', 'number of hidden nodes per hidden layer', and 'the learning rate' were optimized using the suitable values of GA (refer to Table 2). The range of these three parameters were set as (1–4), (5–30), and (0.01–1), respectively. It is observed that GA-ANN produces the best accuracy (i.e., 89.07%) at iteration 202 with optimal values of 'number of hidden layers', 'number of hidden nodes per hidden layers', and 'the learning rate' as 1, 15, and 0.0335, respectively (refer to Fig. 9 and Table 5). Likewise, over the 500 iterations, each with 10-fold cross-validation, using the suitable values of PSO (refer to Table 4), the PSO-optimized ANN produces the best accuracy of 89.33% (refer to Fig. 10 and Table 5).

It is interesting to note that the optimal values for some of the parameters differ when different optimization techniques (e.g., GA or PSO) are used. For example, for GA-SVM, the optimal values for the parameters 'iterations for convergence', 'cost' and 'gamma' are 212, 1.1093, and 0.2474, respectively, while for PSO-SVM, the respective values are 142, 1.3405 and 0.2257 (refer to Table 3). Similarly, for GA-ANN, the optimal values for the parameters 'iterations for convergence', 'number of nodes in hidden layers' and 'learning rate' are 202, 15, and 0.0335, respectively, while for PSO-ANN the respective values are 18, 30 and 0.0189 (refer to Fig. 10 and Table 5). Since the strategy for exploration and exploitation process of both GA and PSO are different in nature, the optimal values of parameters of the classifiers are found to be different. It is further to be noted that the best accuracy values are not practically different. For GA-SVM and PSO-SVM, the best accuracy values are 90.53 and 90.67 (very close), respectively

Table 3
Optimal parameter setting of SVM models.

Model	Iterations for convergence	Iterations	Cost for best solution	Gamma for best solution	Best accuracy (%)
GA-SVM	212	500	1.1093	0.2474	90.53
PSO-SVM	142	500	1.3405	0.2257	90.67

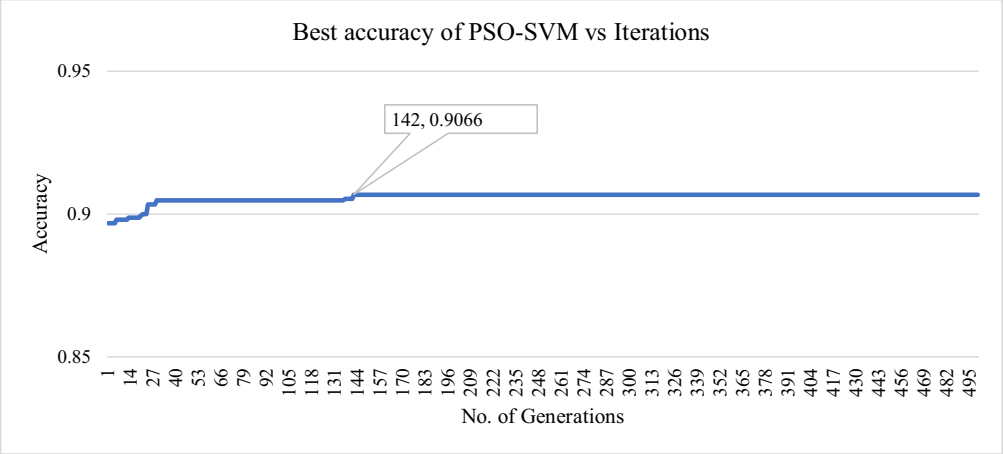


Fig. 8. Plot of best accuracy in each iteration of PSO-SVM.

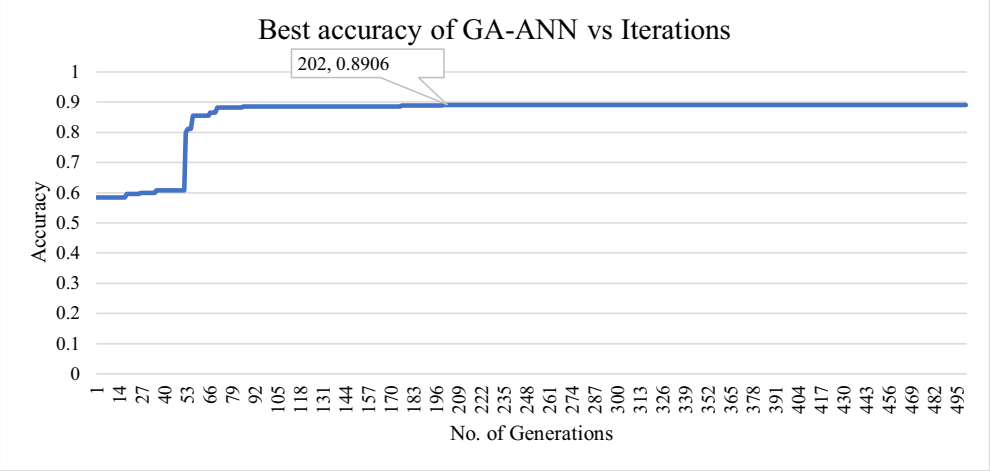


Fig. 9. Plot of best accuracy in each iteration of GA-ANN.

Table 4
The utilised PSO parameters.

SL	Parameter	Value
3	The number of generations	500
4	Swarm size	12
5	Exponent for calculating the number of informants	3
6	Exploitation constant	0.721
7	Local exploration constant	1.193
8	Global exploration constant	1.193

(refer to Table 3). Similarly, for GA-ANN and PSO-ANN, the best accuracy values are 89.07 and 89.33 (very close), respectively (refer to Table 5).

In order to check the robustness of the optimized classifiers, separate 5 runs were executed using 10-fold cross-validation. Following the strategy adopted by Oztekin et al. (2018), different random seeds were used for percent split options (for training and testing). The seeds were assigned to odd numbers, e.g., 123, 125, 127, 129 and 131 for 5 runs, which produced a set of cross-validation folds for each run. Consequently, for each model, 50 results (10-folds x 5 iterations) in terms of accuracies were obtained. Using these values, the box-plot analysis was performed (refer to Fig. 11). From Fig. 11, it reveals that the lowest range of the accuracies is observed in GA-ANN model with a wide degree of dispersion. The highest accuracies are yielded by PSO-SVM, with a very low degree of dispersion. Hence, PSO-SVM is a robust model. Considering both the factors, i.e., accuracy and robustness, the PSO-SVM algorithm has been considered to be the best model among the four classifiers and has been used for extraction of rules, which is explained in the following section.

5.5. Rule extraction

The PSO-SVM-based C5.0 algorithm, being the best classifier, was used to extract feasible assessment rules or combination of factors behind the occurrence of accidents. The reason behind the rule extraction using SVM is that though SVM has been found out as the better algorithm than ANN in terms of higher prediction,

Table 5
Optimal parameter setting of ANN models.

Model	Iterations	Iterations for convergence	Number of hidden layers for best solution	Number of nodes in hidden layers for best solution	Learning rate for best solution	Best accuracy (%)
GA-ANN	500	202	1	15	0.0335	89.07
PSO-ANN	500	18	1	30	0.0189	89.33

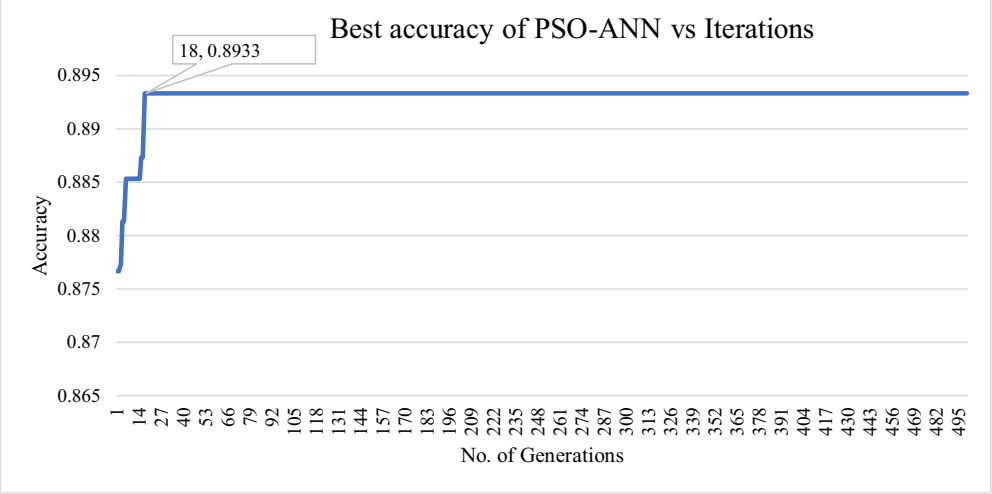


Fig. 10. Plot of best accuracy in each iteration of PSO-ANN.

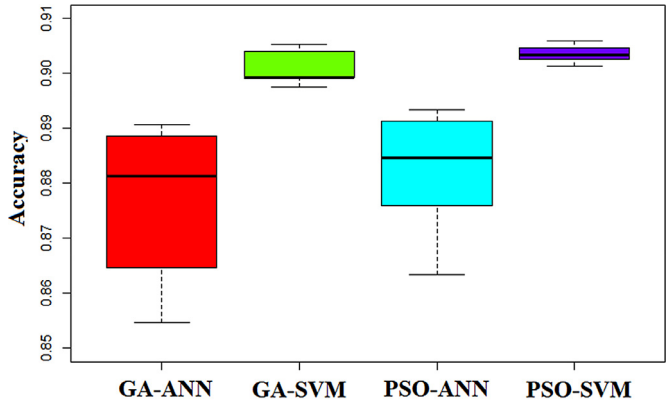


Fig. 11. Box-plot analysis of the accuracy measures for each of the four optimized classifiers.

it operates like a black box process. Thus, in order to make it interpretable, C5.0 algorithm has been applied on the data points represented as support vectors obtained by PSO-SVM to generate rules that can explain the factors behind the occurrence of accidents at work. The reason for hybridizing C5.0 algorithm with SVM is that the C5.0 algorithm can produce less number of rules with useful information than a large number of rules obtained by the normal application of decision tree algorithm on the whole dataset which makes the problems at hand incomprehensible (Han et al., 2015). Basically, the rule generation process consists of two steps; (i) during the first step, the SVM model, which is constructed by the best fold of 10-fold cross validation (CV), is applied to predict the labels of SVs and the original labels of SVs are discarded resulting in the generation of the artificial data; and (ii) during second step, the artificial data are used to train a C5.0 model, and hence, the rule set is built. Finally, the performance of the set of rule is evaluated using 10-fold cross validation. The rules obtained are shown in Table 6.

In Table 6, nine useful rules obtained from PSO-SVM based C5.0 algorithm are shown. Each of the rules explains the incident outcomes i.e., injury, near miss, and property damage with a definite lift and confidence values. The dataset considered for the analysis of rule extraction has 1062 observations which are represented as support vectors obtained from PSO-SVM operation. Of them, the numbers of injury, near miss and property damage cases are 510, 480, and 72, respectively. Hence, the relative frequencies of injury, near miss, and property damage are computed as 0.4802, 0.4520, and 0.0678, respectively. Using these values, Confidence and Lift are calculated for each rule. Following descending order of confidence value, the rules are organized for injury, nearmiss and property damage cases. For example, rule one (R1) explains that injury is occurred in divisions 2, 3, 4, 9, 10 and 11 (Div2, 3, 4, 9, 10, and 11) with some primary causes like dashing or collision, electric flash, equipment/ machinery damage, and road incidents. Three topics i.e., topic 5, 6 and 9 extracted from the text are also attributed to injury in those divisions specified. It explains the fact that injuries usually take place during cleaning operation at those divisions, and the common reason for these injuries is slipping which ultimately increases the number of first-aid cases.

Similarly, in some divisions like Div2, Div4, Div8, Div10, Div11, and Div12, near miss cases are happened more due to mainly slipping as identified by topic 9. Related to this, some days i.e., Monday, Tuesday, and Saturday are identified when near miss cases are found to occur. Investigating the causes behind the near miss cases revealed that derailment, energy isolation, electric flash, gas leakage, hot metals, working at height, toxic chemicals etc. are serving as primary causes of its occurrences. In particular, in some of the divisions like Div2, Div9, Div10, and Div12, some factors such as failure of crane operations, falling from height, fire incidents, pipe leakage and vehicle collision are found to be the main issues resulting in the occurrence of near miss cases.

Likewise, property damage case was also investigated through rule extraction procedure. It is found out from Rule 7 (R7) that in Div3 and Div8, the factors like collision and electrical flash

Table 6
Rules generated from optimized SVM and C5.0-based model.

Rule no.	Rules	Class	n or n/m	Lift	Confidence
R1	Day of Incident in {Friday, Sunday, Thursday, Wednesday} + Division in {Div10, Div11, Div12, Div2, Div3, Div4, Div9} + Incident event in {DC, EF, EMD, RI }+ Topic in {Topic 5, Topic 6, Topic 9}	Injury	35/1	2.0	0.946
R2	Division in {Div13, Div6, Div7} + Injury Type = ITNA	Near miss	114/2	2.2	0.974
R3	Injury Type = ITNA + Incident event in {D, EI, FE, GL, HM, HP, LTT, MA, MH, OI, PI, R, RO, S, SI, STF, TC, WH}	Near miss	386/16	2.1	0.956
R4	Day of Incident in {Monday, Saturday, Tuesday} + Division in {Div10, Div11, Div12, Div2, Div4, Div8} + Injury Type = ITNA + Topic in {Topic 5, Topic 6, Topic 9}	Near miss	23/1	2.0	0.920
R5	Division in {Div10, Div12, Div2, Div9} + Injury Type = ITNA + Topic in {Topic 2, Topic 3, Topic 4, Topic 7, Topic 8}	Near miss	155/15	2.0	0.898
R6	Injury Type = ITNA + Topic in {Topic 2, Topic 3, Topic 7, Topic 8}	Near miss	350/38	2.0	0.889
R7	Division in {Div3, Div8} + Injury Type = ITNA + Incident event in {DC, EF}	Property damage	5	12.6	0.857
R8	Day of Incident in {Friday, Monday, Tuesday} + Division = Div11 + Injury Type = ITNA + Incident event in {DC, EF, EMD}	Property damage	11/3	10.2	0.692
R9	Division = Div4 + Injury Type = ITNA + Incident event in {DC, EF, EMD} + Topic = Topic 4	Property damage	27/8	10.2	0.690

occurred from short circuit are the primary reasons for property damage. From R8, it is revealed that the property gets damaged in Div11 due to mainly collision, electric flash or short circuit which are mostly observed in three days in a week i.e., Monday, Tuesday, and Friday.

The rules, therefore, extracted in this study, are plant oriented. Some of them are more useful for the management to help them undertake initial proactive measures to minimize the number of occurrence of accidents in the plant. As consequences, some of the divisions are identified where incident outcomes i.e., injuries, near miss and property damage cases are identified separately. Incident events are also figured out for each of these outcomes. These measures, therefore, seem to be effective for the steel plant. However, similar studies on steel plant are very rare indeed and hence, it is very hard to validate the findings by previous research. Moreover, the rules extracted can be deemed as preliminary hypotheses for the future studies.

6. Conclusions and future scopes

In this present study, optimized machine learning-based prediction models have been developed to predict the incident outcomes at workplace. Two powerful and effective classifiers, namely SVM and ANN have been used for this task whose parameters are optimized by two popular optimizers, namely GA and PSO. The findings of this research work put forward some useful insights on data pre-processing tasks, parameter optimizations of classifiers, and rule extraction from the accident data. For examples, findings of the analysis reveal that PSO-based SVM outperforms other classifiers in terms of accuracy (i.e., accuracy of PSO-based SVM is 90.67%). In addition, using sensitivity analysis, PSO-SVM is found to be the most robust classifier as well. Furthermore, rules obtained from the PSO-SVM based C5.0 are also found to be effective as they can be used for the interpretation of the factors in terms of rules behind the incident occurrences. Some of the key findings from the analyses explore that slipping is the common cause for injury cases (as observed from Topic 9). Other than slipping, other causes including collision, electric flash, and road incidents are identified for the injuries in some of the divisions of the steel plant. Slipping issues remain also the primary cause for the near miss cases for some of the divisions. In addition, some days of the week like Saturday, Monday, and Tuesday are also identified where near miss incidents happen more frequently. In Div2, Div9, Div10, and Div12, crane operation, falling from height, fire incidents, pipe leakage and vehicle collision are responsible for the occurrence of near miss cases. Moreover, collision and electrical flash led to property damage in some of the divisions, Div3 and Div8.

Apart from the application of classifiers, this study also discussed a proper sequence of data pre-processing task using missing value imputation by RF, new feature addition like 'SOPC' from categorical attributes using expert judgement and 'topics' from unstructured incident narratives using topic modelling technique. The attributes like 'injury types' and 'day' are found to be the most important and the least important predictors using chi-square technique towards the prediction of incident outcomes. It is also noteworthy that from the analysis of chi-square, newly generated attribute 'Topic' is also revealed to be one of the important predictors. Therefore, the present study is expected to hold a good potential to contribute both in theoretical and practical aspects.

6.1. Contributions to theory

The higher predictive accuracy of the optimized classifiers reveals that accidents do not occur in a chaotic fashion, but rather than underlying patterns and trends do exist and hence, can be explored as well as captured with the utility of machine learning techniques. This finding suggests that occupational safety should be studied empirically in a systematic way rather than strictly following qualitative approach through subjective, expert-opinion-based data analysis. Higher predictive accuracy of the classifiers indicates that the topic modelling from the unstructured narratives/texts is viable and useful. Further, it generates structured data from unstructured accident reports. Moreover, it justifies the choice of algorithmic modelling over parametric counterpart. Another important point to be noted that optimization of the parameters improves the prediction power of a classifier which was not addressed in accident data analyses.

6.2. Contributions to practice

From the perspective of industry professionals, prediction of occupational accidents has been long aimed. Some recent studies on risk analysis, leading & lagging indicators, and precursor analysis are found to be very useful for this purpose. However, in most of the cases, these studies show the dependencies on experts or safety professionals while analysing the data. In fact, safety professionals have very limited personal history with injury cases and moreover, their judgment can be altered in the presence of a plethora of cognitive biases which lead to uncertainty. On the contrary, the use of ML algorithms can learn from the historical records or data effectively and efficiently, as well. The learning from data by ML techniques can be used to complement the experts' opinion which is potentially biased that eventually leads to efficient decision making as results. For instance, a user may

need to identify the factors attributed to accidents which can easily be done by the proposed model. If the user needs to analyse the text data and to use it towards the prediction of incident outcomes, the proposed model in the study can be used as a useful platform for such analysis. Thus, the prediction followed by identification of factors attributable to accidents can be used as actionable feedback to plan better in pre-job safety meetings.

However, like other studies, this study also has some limitations. In data pre-processing task, a lot of manual effort was required to clean the data suitable for analysis. Moreover, the dataset used in this study has limited number of accident records. It is recommended of using substantial amount data for the analysis for the better generality of the model. As the future scope, the work can be extended to build an automated decision support system which not only can predict the incident outcomes, but also can provide the smart decisions based on a set of rules derived. For obtaining better rules, other decision tree algorithms e.g., classification and regression tree, or other ensemble methods like random forest, boosting, bagging or stacking techniques can be used. In addition, to exploit the unstructured text data, text mining can be deployed to extract the body parts injured from the dataset, which could also be used in further analysis. Methodological advancements in the extraction of structured data from unstructured text through the use of the natural language processing (NLP) can also be done as future works. Another important direction for future research is that same type of analysis can be conducted in different industries like construction, mining, aviation etc. to validate the model used in this study.

Acknowledgement

We would like to thank the Ministry of Human Resource Development (MHRD), New Delhi, India, Ministry of Steel, New Delhi, India, and Tata Steel Limited, Jamshedpur, India for funding this research through Uchchatar Avishkar Yojana (UAY) for the project entitled **"Safety Analytics: Save People at Work from Accidents and Injuries (WAI)"**. We are thankful to the safety personnel of the steel industry for their support from the data collection phase to final analysis of the study.

References

Alikhani, M., Nedaie, A., Ahmadvand, A., 2013. Presentation of clustering-classification heuristic method for improvement accuracy in classification of severity of road accidents in Iran. Saf. Sci. 60, 142–150.

Arun, R., Suresh, V., Madhavan, C.E.V., Murty, M.N., 2010. On finding the natural number of topics with latent dirichlet allocation: some observations. Adv. Knowl. Discov. Data Min. 391–402.

Aviad, B., Roy, G., 2011. Classification by clustering decision tree-like classifier based on adjusted clusters. Expert Syst. Appl. 38, 8220–8228. doi:10.1016/j.eswa.2011.01.001.

Barakat, N.H., Bradley, A.P., 2007. Rule extraction from support vector machines : a sequential covering approach. IEEE Trans. Knowl. Data Eng. 19, 729–741.

Barakat, N., Diederich, J., 2004. Learning-based rule-extraction from support vector machines : performance on benchmark data sets. In: Proceedings of the 3rd Conference on Neuro-Computing Evol. Intell.

Bengio, Y., Mesnard, T., Fischer, A., Zhang, S., Wu, Y., 2017. STDP-compatible approximation of backpropagation in an energy-based model. Neural. Comput. 29, 555–577. doi:10.1162/NECO.

Bevilacqua, M., Ciarapica, F.E., Giacchetta, G., 2008. Industrial and occupational ergonomics in the petrochemical process industry : a regression trees approach. Accid. Anal. Prev. 40, 1468–1479. doi:10.1016/j.aap.2008.03.012.

Breiman, L., 2001. Decision tree forest. Mach. Learn. 45, 5–32.

Brown, D.E., 2016. Text mining the contributors to rail accidents. IEEE Trans. Intell. Transp. Syst. 17, 346–355.

Cao, J., Xia, T., Li, J., Zhang, Y., Tang, S., 2009. Neurocomputing A density-based method for adaptive LDA model selection. Neurocomputing 72, 1775–1781. doi:10.1016/j.neucom.2008.06.011.

Cervantes, J., Garcia-lamont, F., Rodriguez-mazahua, L., López, A., Ruiz-castilla, J., Trueba, A., 2017. PSO-based method for SVM classi fi cation on skewed data sets. Neurocomputing 228, 187–197.

Chou, J., Cheng, M., Wu, Y., Pham, A., 2014. Optimizing parameters of support vector machine using fast messy genetic algorithm for dispute classificatio. Expert Syst. Appl. J. 41, 3955–3964.

Das, G., Kumar, P., Kumari, S., 2014. Expert systems with applications artificial neural network trained by particle swarm optimization for non-linear channel equalization. Expert Syst. Appl. 41, 3491–3496.

Deveaud, R., SanJuan, E., Bellot, P., 2014. Accurate and effective latent concept modeling for ad hoc information retrieval. Document numérique 17 (1), 61–84.

EUROSTAT, 2009. Labour force survey 2007 ad hoc module on accidents at work and work-related health problems. In: Proceedings of the European Communities.

Fragiadakis, N.G., Tsoukalas, V.D., Papazoglou, V.J., 2014. An adaptive neuro-fuzzy inference system (anfis) model for assessing occupational risk in the shipbuilding industry. Saf. Sci. 63, 226–235. doi:10.1016/j.ssci.2013.11.013.

Fu, X., Ong, C., Keerthi, S., Hung, G.G., Goh, L., 2004. Extracting the knowledge embedded in support vector machines. In: Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, 2004, pp. 291–296.

Fung, G., Sandilya, S., Rao, R.B., 2005. Rule extraction from linear support vector machines. In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, pp. 32–40.

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. Proc. Natl. Acad. Sci. 101, 5228–5235.

Haddon, W., Suchman, E., Klein, D., 1964. Accident Research: Methods and Approaches. Harper & ROW, New York.

Hale, A., Hale, M., 1970. Accidents in perspective. Occup. Psychol. 44, 115–122.

Han, L., Luo, S., Yu, J., Pan, L., Chen, S., 2015. Rule extraction from support vector machines using ensemble learning approach : an application for diagnosis of diabetes. IEEE J. Biomed. Health Inform. 19, 728–734.

He, X., Chen, W., Nie, B., Zhang, M., 2010. Classification technique for danger classes of coal and gas outburst in deep coal mines. Saf. Sci. 48, 173–178. doi:10.1016/j.ssci.2009.07.007.

Heinrich, H., Petersen, D., Ross, N., 1980. Industrial Accident Prevention, fifth ed. McGraw-Hill, New York.

Hendrick, H., 1986. Macroergonomics as a Preventive Strategy in Occupational Health: An Organizational Level Approach, Vol. 124. Hendrick doi:10.1177/00333549091244S103.

Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. An Introductory Analysis with Application to Biology, Control, and Artificial Intelligence. Univ Michigan Press, Ann Arbor, MI.

Houari, R., Bounceur, A., Kechadi, M., Tari, A., Euler, R., 2016. Dimensionality reduction in data mining : a copula approach. Expert Syst. Appl. 64, 247–260.

ILO, 2003. Safety in Numbers, Global Safety Culture at Work. The International Labour Organisation, Geneva.

ILO, 2008. Promoting Safe and Healthy Jobs: The ILO Global Programme on Safety. Health and the Environment (Safework).

Jones, S.J., Lyons, R.A., 2003. Routine narrative analysis as a screening tool to improve data quality. Inj. Prev. 9, 184–186.

Kecman, V., 2005. Support vector machines – an introduction. Support Vector Mach. Theory Appl. 177, 1–47.

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In: Proceedings of the IEEE International Conference on Neural Networks. Perth, Australia, pp. 1942–1948.

Khanzode, V.V., Maiti, J., Ray, P.K., 2012. Occupational injury and accident research : a comprehensive review. Saf. Sci. 50, 1355–1367. doi:10.1016/j.ssci.2011.12.015.

Kunce, J.T., 1967. Vocational interests and accident proneness. J. Appl. Psychol. 51, 223–225.

Leu, S., Chang, C., 2013. Bayesian-network-based safety risk assessment for steel construction projects. Accid. Anal. Prev. 54, 122–133. doi:10.1016/j.aap.2013.02.019.

Li, J., Guo, X., 2015. Knowledge distribution and text mining of international aviation safety research. In: Proceedings of the 15th International Conference on Man–Machine–Environment System Engineering, pp. 151–159. doi:10.1007/978-3-662-48224-7.

Li, Q., Zhang, X., Rigat, A., Li, Y., 2015. Parameters optimization of back propagation neural network based on memetic algorithm coupled with genetic algorithm. In: Proceedings of the 2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing and 2015 IEEE 12th International Conference on Autonomic and Trusted Computing and 2015 IEEE 15th International Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom), pp. 1359–1364. doi:10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.245.

Liao, S.G., Lin, Y., Kang, D.D., Chandra, D., Bon, J., Kaminski, N., et al., 2014. Missing value imputation in high?dimensional phenomic data: imputable or not, and how? BMC Bioinform. 15, 346. doi:10.1186/s12859.

Martín, J.E., Rivas, T., Matías, J.M., Taboada, J., Argüelles, A., 2009. A Bayesian network analysis of workplace accidents caused by falls from a height. Saf. Sci. 47, 206–214. doi:10.1016/j.ssci.2008.03.004.

Matías, J.M., Rivas, T., Martín, J.E., Taboada, J., 2008. A machine learning methodology for the analysis of workplace accidents. Int. J. Comput. Math. 85, 559–578. doi:10.1080/00207160701297346.

Núñez, H., Angulo, C., Catala, A., 2002. Rule extraction from radial basis function networks by using support vectors. In: Proceedings of the Ibero-American Conference on Artificial Intelligence, pp. 440–449.

Niraula, N., Banjade, R., Ştefănescu, D., Rus, V., 2013. Experiments with semantic similarity measures based on LDA and LSA. In: International Conference on Statistical Language and Speech Processing, pp. 188–199.

Olson, D.L., Delen, D., Meng, Y., 2012. Comparative analysis of data mining methods for bankruptcy prediction. Decis. Support Syst. 52, 464–473. doi:10.1016/j.dss.2011.10.007.

Oztekin, A., Kong, Z.J., Delen, D., 2011. Development of a structural equation modeling-based decision tree methodology for the analysis of lung transplantations. Decis. Support Syst. 51, 155–166. doi:10.1016/j.dss.2010.12.004.
Oztekin, A., Al-Ebbini, L., Sevkli, Z., Delen, D., 2018. A decision analytic approach to predicting quality of life for lung transplant recipients: a hybrid genetic algorithms-based methodology. Eur. J. Oper. Res. 266, 639–651.
Pavlinek, M., Podgorelec, V., 2017. Text classification method based on self-training and LDA topic models. Expert Syst. Appl. 80, 83–93.
Pereira, F.C., Rodrigues, F., Ben-akiva, M., 2013. Text analysis in incident duration prediction. Transp. Res. Part C 37, 177–192.
Pham, H.N.A., Triantaphyllou, E., 2011. A meta-heuristic approach for improving the accuracy in some classification algorithms. Comput. Oper. Res. 38, 174–189. doi:10.1016/j.cor.2010.04.011.
Rivas, T., Paz, M., Martín, J.E., Matías, J.M., García, J.F., Taboada, J., 2011. Explaining and predicting workplace accidents using data-mining techniques. Reliab. Eng. Syst. Saf. 96, 739–747.
Robinson, G.H., 1982. Accidents and sociotechnical systems: principles for design. Accid. Anal. Prev. 14, 121–130.
Sánchez, A.S., Fernández, P.R., Lasheras, F.S., Juez, F.J.D.C., Nieto, P.J.G., 2011. Prediction of work-related accidents according to working conditions using support vector machines. Appl. Math. Comput. 218, 3539–3552. doi:10.1016/j.amc.2011.08.100.
Sanmiquel, L., Rossell, J.M., Vintro, C., 2015. Study of Spanish mining accidents using data mining techniques. Saf. Sci. 75, 49–55.
Solaro, N., Barbiero, A., Manzi, G., Ferrari, P.A., 2017. A sequential distance-based approach for imputing missing data : forward Imputation. Adv. Data Anal. Classif. 11, 395–414. doi:10.1007/s11634-016-0243-0.
Tixier, A.J., Hallowell, M.R., Rajagopalan, B., Bowman, D., 2016. Application of machine learning to construction injury prediction. Autom. Constr. 69, 102–114.
Vallmuur, K., 2015. Machine learning approaches to analysing textual injury surveillance data : a systematic review. Accid. Anal. Prev. 79, 41–49.
Vapnik, V., 1995. The Nature of Statistical Learning Theory. Springer-Verlag.
Wei, L.W., Wei, C.S., Wan, X.Q., 2013. Data classification using support vector. Adv. Mater. Res. 936–939. doi:10.4028/www.scientific.net/AMR.662.936.
Witten, I.H., Frank, E., Hall, M.A., Pal, C.J., 2016. Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.
Xue, X., Liu, E., 2017. Seismic liquefaction potential assessed by neural networks. Environ. Earth Sci. 76, 1–15. doi:10.1007/s12665-017-6523-y.
Yi, W., Chan, A.P.C., Wang, X., Wang, J., 2016. Automation in construction development of an early-warning system for site work in hot and humid environments : a case study. Autom. Constr. 62, 101–113. doi:10.1016/j.autcon.2015.11.003.
Zhang, H., Peng, Y., Tian, G., Wang, D., Xie, P., 2017. Green material selection for sustainability: a hybrid MCDM approach. PLoS One 12, e0177578.
Zheng, B., Zhang, J., Won, S., Lam, S.S., Khasawneh, M., 2015. Predictive modeling of hospital readmissions using metaheuristics and data mining. Expert Syst. Appl. 42, 7110–7120.