Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predicting Fatal Mine Accident Categories Using Text Mining and Machine Learning: A Comparative Model Analysis

Arra Kumar
arrakumar.rs.min19@itbhu.ac.in
Indian Institute of Technology BHU Varanasi https://orcid.org/0000-0001-7242-7618
Suprakash Gupta
Indian Institute of Technology BHU Varanasi

Research Article

Keywords: Classification model, Data Analysis, Fatal Accidents in Mines, Text Mining

Posted Date: July 11th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6634717/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License

Version of Record: A version of this preprint was published at International Journal of System Assurance Engineering and Management on October 6th, 2026. See the published version at https://doi.org/10.1007/s13198-026-03446-y.

Abstract

The threat of mine accidents continues to endanger workers worldwide, so proper classification of accidents becomes essential for developing specific prevention measures. The research focuses on advancing reactive safety practices because prediction models must identify upcoming accident types during planning. Our method combines machine learning with text mining techniques to examine historical fatal accident reports using approaches that have not yet been included in mining safety literature. A statistical model prediction pipeline has been established to study fatal accident reports using natural language processing (NLP) with vectorisation techniques and six machine learning (ML) algorithms: logistic regression (LR), support vector machine (SVM), random forest (RF), naïve Bayes (NB), decision tree (DT), and multilayer perceptron (MLP) algorithms with stratified 10-fold cross validation. We evaluated the performance using the confusion matrix, precision, recall, and weighted average F1 scores revealed that the MLP model achieved superior performance (0.84 F1 score), followed by LR (0.83), SVM (0.81), RF (0.70), NB (0.60), and DT (0.57). The innovative aspect of this study is its use of complete text mining techniques against unstructured accident reports, which enables detection capabilities that typical structured data evaluation methods cannot achieve. Research shows that Natural Language Processing (NLP) and Machine Learning (ML) integrated systems create exceptional improvements in accident predictions. Mining safety authorities and stakeholders now have evidence-based prevention tools that aid the development of focused safety initiatives to lower mining-related deaths.

1.0 Introduction

Coal is India's primary raw material for electricity generation, and the mining industry plays an important role in the Indian economy. In 2022, coal accounted for approximately 55% of India's total energy consumption and was used to generate over 70% of the electricity. India is the third-largest coal producer in the world, with an estimated production of over 890 million tonnes in 2023 (Ministry of Coal, 2023). Indian coal employs over 1 million people, accounting for 14% of the world's employment in the coal industry. Mining significantly contributes to India's GDP, but it offers one of the most dangerous workplaces globally. Occupational accidents account for over 2.78 million deaths annually, with mining accounting for one in seven fatal accidents. Retrospective analysis of accidents is crucial to identify the type of accident and devise suitable countermeasures to prevent similar future incidents.

Mining is characterised by one of the most hazardous workplaces, with a significant fatality rate in workplace accidents. Different countries conducted previous research on mining accidents to enhance safety and reduce accidents in various industries. In the United States, Kecojevic et al. (2006) studied mine accidents to establish the causes of equipment fatalities in surface mining industries. In China, where there is significant concern about coal mining accidents, several research studies have attempted to investigate the role of human and managerial factors. A study on variables predicting the safety climate of accident rates was conducted by Chen et al. (2011) in Chinese coal mines. Kausar Sultan Shah et al. (2020) have studied accidents in Pakistan's Cherat coal field and found that roof falls are the leading cause of accidents. Similarly, Ismail et al. (2021) have revealed that mechanical failures are the most common cause of accidents in mines. In his study, Yunfei Zhu et al. (2018) concluded that explosions and fires are the leading causes of mine accidents in China. Ghasemi et al. (2012) conducted a qualitative investigation of roof collapse incidents in Iranian underground coal mines. They went on to list a few important factors, such as poor ground conditions and a lack of infrastructure, that would account for these incidents. An investigation by Yin et al. (2016) into incidents of methane gas explosions in Chinese coal mines focused on patterns in the features of the explosions and countermeasures. In their study, Zhang et al. (2018) noted that road conditions, vehicle maintenance, and other factors are critical when analysing transportation-related accidents in Chinese coal mines. Amyotte et al. (2008) looked into previous dust explosion accidents in the mining and other sectors and how preventative and cautious measures have been implemented in the literature. The findings of statistical modelling of occupational accidents in the Spanish mining sector based on the employee's experience, the time of occurrence of accidents, and the kind of mines were reported by Sanmiquel et al. (2015). With an emphasis on heat stress, Xiang et al. (2013) studied the effect of humid environmental conditions and accidents at Australian mining sites. Studying 188 accidents, Zhang J et al. (2019) have concluded that gas explosions, mine water inrush, and coal dust explosions are the significant causes of accidents in underground coal mines in China.

During the past years, immense studies have been conducted on India's mining accidents, but they are still insufficient. Many types of research have been conducted on the types of mine accidents, such as falls of ground, gas, mechanical and transport-associated, electrical, dust, inundation, and multi-causal accidents. Sharma et al. (2019) reviewed fatalities in Indian coal mines and concluded that roof falls and transportation accidents were the most prevalent. In their paper, Tripathy and Ala (2018) discussed the research on safety regulations in Indian mines, acknowledging that they have not been fully complied with and enforced. Pramod Kumar et al. (2019) analysed Indian mining accident reports using fuzzy mathematical concepts and concluded that slips and KBMs (knowledge base mistakes) are the leading causes of accidents. Arra et al. (2023) developed a statistical model with the help of a questionnaire survey from mine employees and concluded that working environmental conditions, age, and experience influenced the occurrence of accidents for an employee working in underground coal mines.

Most research studies in the mining industry have used statistical analysis to examine the relationship between accidents and various factors. However, these studies have relied heavily on secondary data, which may suffer from human error bias and a lack of repeatability. In contrast to using unstructured text data for statistical analysis, analysing text reports with the proper text mining methods has shown that it can produce trustworthy results. Natural Language Processing (NLP) and machine learning (ML) models were used to analyse text documents and extract key information. Various industries have successfully used text mining analysis for safety improvement, but its use in the mining industry remains limited. In the construction industry, Ubeynarayana and Goh (2017) and Cheng et al. (2020) have used different machine learning (ML) models to classify construction accident narratives. Onan & Korukoğlu (2015) conducted sentiment analysis using web mining and text mining techniques, applying NLP and ML models for feature selection based on genetic rank aggregation. A study by Pilar Lopez Ubeda et al. (2020) has shown that convolutional neural networks (NN) and ML models are promising in identifying unexpected observations in radiology reports written in Spanish. Toraman et al. (2020) have used capsule networks with artificial neural networks to classify instances of COVID-19 infection based on X-ray images. Chen et al. (2012) and Debortoli et al. (2014) showed that text mining is a valuable way to look at

large amounts of text data that are hard to analyse with regular statistical methods, and they also explained how text mining can find information in text that traditional analysis might overlook.

These studies demonstrate the diverse nature of mining accidents and the importance of tailored approaches to accident prevention based on specific types and causes. They also highlight the value of detailed accident classification and analysis in developing effective safety strategies. To enhance safety within the mining sector, it is necessary to conduct a comprehensive analysis of past accident reports to comprehend the underlying reasons and subsequent reduction of such incidents by making error-free classification of accidents.

While text mining and machine learning techniques have been applied in various safety domains, their application to mining accident classification represents a novel approach that addresses critical gaps in current safety practices. The innovation in our study lies not in developing new algorithms but in the novel application of established techniques to an underexplored domain, unstructured mining accident reports. This study represents a significant shift from traditional statistical analysis of structured accident data to leveraging rich textual information that contains valuable contextual insights often lost in structured data collection. The practical contribution of this research is substantial, as it provides mining safety professionals with a validated methodology to automatically classify and identify accident patterns from narrative text, enabling more targeted prevention strategies in one of the world's most hazardous industries.

In summary, it is important to clarify that, while accident reports are documented records, we consider them primary data in the context of our research because they contain direct, firsthand accounts and official investigation findings rather than summarised or aggregated statistics. These reports provide rich, unprocessed narrative details about accidents that have not been previously analysed using text mining techniques, making them a primary source of information for our specific research objective. Unlike secondary analyses that work with preprocessed or summarised data, our approach extracts insights directly from the original narrative text, accessing information that would otherwise be lost in statistical summaries. The objectives of this study are as follows:

1. Studying popular supervised machine learning models and natural language processing techniques in the context of accident report categorisation.
2. Using the Stratified K-fold Cross-Validation (SKCV) approach to enhance the efficacy of machine learning models.
3. Determining the best classification model for reports on fatal mining accidents.

We structure the rest of the article as follows: Section 2 gives an overview of the materials and methodology used in the systems and describes the construction of the classification model. Section 3 presents the experimental results and discussion. Finally, the paper concludes in section 4 and presents future work direction.

2.0 Materials and Methodology

This study aims to classify the fatal mining accident reports based on different categories/types of accidents. Figure 1 depicts the proposed methodology for the classification of accident reports. This research extensively explores the capability of text mining to correctly identify the causes of fatal mining accidents. Text mining is an AI technique that uses Natural Language Processing (NLP) to transform unstructured text documents into normalised structured data suitable for ML analysis.

2.1 Natural Language Processing

Natural Language Processing (NLP) analyses, comprehends and interprets unstructured text data. We have developed NLP approaches to operate with natural language text, allowing us to extract data from texts, identify patterns, and make predictions. Accident reports contain a large volume of relevant information in text documents bearing unstructured characteristics, and it is difficult to derive insightful conclusions from them using conventional statistical approaches.

In this study, we aim to classify the fatal accident reports by the cause of the accident. Collected data (accident reports) are in text format and need to be converted into a machine-understanding language for further processing (model development) using machine learning models.

The categorisation of textual documents using NLP consists of four significant steps:

1. Data gathering
2. Preprocessing of collected information.
3. Vectorisation

2.1.1 Data gathering

This study begins by collecting fatal accident reports from the Directorate General of Mines Safety (DGMS) in Dhanbad, as documented in the Statistics of Mines in India, Volume-I (Coal), covering the years 1995 to 2015. DGMS has classified these fatal accidents based on eight types of accidents, namely, 'Ground Vibrations', 'Non-Winding Transportation Machinery', 'Other than transportation machinery', 'Explosives', 'Electrical', 'Fall of objects', 'Dust, Gas, and other combustible matters', and 'Other causes.

It is important to acknowledge that mining accidents often result from multiple contributing factors acting in concert. The DGMS classification system used in our dataset assigns each accident to a single primary causal category based on investigation findings, though secondary factors may have contributed. This single-label classification approach aligns with standard safety practice that identifies primary causal factors to prioritise intervention efforts. While multi-label classification would offer additional insights, the current DGMS reporting structure does not provide explicit multi-causal coding. Our

methodology respects this existing classification framework while establishing a foundation for future work that might incorporate multi-label approaches. Primary cause identification enables safety officers to prioritise the most critical intervention points in a hierarchical causation structure, where secondary factors often cascade from primary failure points.

Table 1
Category/Type-wise distribution of collected fatal mine accident reports

Sl. No	Reported type of accident	Number of mine fatal accident reports due to a particular cause
1	Non-Winding Transportation Machinery	**518**
2	Other than Transportation Machinery	**210**
3	Fall of Objects	**186**
4	Ground Vibrations	**182**
5	Electrical	**85**
6	Other Cause	**45**
7	Explosive	**44**
8	Dust, Gas, and other combustible matters	**38**
	Total	**1308**

For this study, we collected a total of 1320 fatal mine accident reports. Of the 1320 accident reports, 1308 were used in the analysis, while 12 were discarded due to some cases regarding accidents that were pending in court and detailed information not being available. These reports contain the category/type of accident, date, time of the accident, basic information regarding the mines, the deceased employee, and chronological details of the incident. Sl. No.1 in Table 2 presents a tailored accident report based on the collected data.

Table 2
Collected Accident Report and step-by-step output of the process

Sl. No.	Step Description	Collected Accident Report
1	Narrative of the incident as per the collected report	While five loaders were loading the blasted coal of rib at the goaf edge in a depillaring district, a mass of roof coal, attached with a thin shale band, measuring about 5m x 3m x 0.25m (thick), fell from a height of about 2.7m from the roof inflicting serious bodily injuries to two loaders and reportable injuries to three of them. One of the seriously injured succumbed to the injuries after nine days.
2	Narrative of the incident after removal of Stop words/Punctuation/Digits	loaders loading blasted coal rib goaf edge depillaring district mass roof coal attached thin shale band measuring fell height roof inflicting bodily injuries loaders reportable injuries seriously injured succumbed injuries days.
3	Output of tokenization	loaders, loading, blasted, coal, rib, goaf, edge, depilating, district, mass, roof, attached, thin, shale, band, measuring, fell, height, inflicting, bodily, injuries, reportable, seriously, injured, succumbed, days.
4	Output vectors of the document (vector of Size 300)	array ([-5.92593312e-01, 9.05643463e-01, -2.42457891e + 00, 1.08177686e + 00, 1.61426938e + 00, 2.70948172e-01, 1.31447434e + 00, 4.75955105e + 00, ……..……..…… . . ……5.98857462e-01, -4.41703349e-01, -6.12378120e-02, 9.53051686e-01, -2.19528794e + 00, 1.19853640e + 00, 5.08765727e-02, -1.85591292e + 00, -4.38102186e-01, 1.16300535e + 00]

2.1.2 Preprocessing of the data

The collected fatal accident reports contained much information that was not essential for model development. We must filter out this data to enhance the model's efficiency and decrease its run time. Data is screened out to increase the model's efficiency and decrease the model run time. Vajjala et al. (2020) achieved this by eliminating digits, punctuation (special characters), stop words, and tokenisation.

- Stop Word Removal: Test reports frequently utilise several English words such as a, an, the, of, in, he, was, be, have, and do. These words do not have any content that distinguishes them across various categories. These words are known as "stop words" and are typically (but not always) skipped over during data processing.
- Tokenisation: Tokenisation is the process by which the textual document is divided into smaller portions, which may be a word or a phrase in the context, to obtain valuable information from it.

We removed the stop words, punctuation marks, and numerical elements from the accident reports using the SpaCy library in Python. We then used the resulting clean text data (Sl. No. 2 in Table 2), which was utilised in the subsequent study phases. In tokenisation, we were able to extract insights and patterns that are embedded in the report. The frequency of specific words or phrases in the text document was analysed, and common themes and topics related to the fatal mine accidents were identified. The output of this step is given below in Sl. No. 3 in Table 2.

2.1.3 Vectorisation using Word Embeddings

This technique represents words as dense and continuous vectors of numerical values inside a high-dimensional space where each value captures different aspects of the word's meaning (Mikolov et al., 2013). The process of capturing the semantic and syntactic relationships between words facilitates the comprehension and analysis of natural language.

In conventional practices, computers process text as aggregations of discrete symbols without capturing the meaning and context of words. Word embeddings convert words into vectors in a continuous vector space, where the spatial relationships between the vectors capture relationships between the relevant words. Vector representations of words that have similar meanings or appear in comparable settings frequently resemble one another; using this characteristic of vectors, word embeddings record connections and analogies between words.

One can perform word embeddings in various ways, often involving training a neural network with a substantial corpus of textual data. A large volume of training data requirements is a hindrance to its applicability. Word2Vec is a popular word-embedding technique that trains a shallow neural network on a sizable corpus of text to learn word embeddings. This word embedding technique predicts the neighbouring terms of a given word using the context of the words. The word represents itself in vector form, with a vector size of 300. Sl. No. 4 in Table 2 presents the output vectors of the document. Interested readers may refer to the work of Vajjala et al. (2020) for more details on the word embedding technique.

The Word2Vec implementation in our study was particularly valuable for capturing mining-specific terminology relationships. Mining accident reports contain specialised vocabulary where technical terms carry significant semantic weight in understanding accident causality. For example, our word embedding model successfully captured the semantic proximity between terms like 'methane', 'ventilation', and 'explosion', allowing the classification algorithms to recognise these relationships even when specific terms varied across reports. Similarly, transportation-related accidents showed distinct vector clusters around terminology related to vehicle operations, maintenance states, and operator actions. By mapping these specialised word relationships in a 300-dimensional vector space, our method allowed the machine learning models to find subtle language patterns that suggest certain types of accidents, which simpler word-count methods might overlook.

2.2 Machine Learning Classification Models

After being classified into machine-readable sectoral form, it is difficult to choose a unique algorithm that learns effectively across all sets of data. Comparing and testing a group of algorithms is a valuable strategy for selecting the appropriate model. Here, we briefly review six classifier models developed for further data analysis.

2.2.1 Logistic Regression (LR):

An algorithm for statistical classification determines the relationship between input parameters and output factors. The logistic regression method helps predict how likely it is that a certain piece of data will belong to a specific category in classification problems (Zhang et al., 2020). LR seeks to establish a linear separator between classes to maximise the likelihood of data allocation. The logistic function assigns weight to features across all classes to enhance classification ability (Hosmer & Lemeshow, 2004).

Advantages of this model are interpretable, efficient training, work well with linear boundaries, and probabilistic output. Decision boundaries remain linear; the model provides inferior results for complex non-linear relationships and shows sensitivity to outliers.

2.2.2 Decision tree (DT):

This algorithm uses a tree-based structure and divides the input space into manageable chunks based on the input data characteristics. Each leaf node of the DT corresponds to a class name or a numeric value, while the inside nodes indicate decisions based on the input data characteristics. Recursive partitioning is a common practice used to train decision trees, helping to manage overfitting issues by pruning the trees (Ikonomakis et al., 2005).

Decision trees offer three main benefits: simple interpretation, the ability to detect non-linear patterns, and the need for very little preprocessing work. The main challenges of this method include overfitting the training data and being unstable, as small changes in the input can lead to significant changes in the structure, along with a limited ability to manage complex relationships.

2.2.3 Random Forest (RF):

It is an ensemble learning technique that integrates many decision trees to enhance the accuracy of predictions and control the issue of overfitting. The methodology uses various subsets of input features of training data and trains a set of decision trees for classifying data. The final prediction is made by computing the average of the forecasted values of individual decision trees. Random forests are well-recognised for their robustness and superior precision, making them a popular choice for addressing classification and regression tasks (Onan, A., 2016).

This classification model offers two key benefits: it reduces overfitting through ensemble learning and handles high-dimensional inputs effectively while maintaining robustness to outliers. Random Forest faces two main drawbacks: its lesser interpretability compared to single decision trees, its heavy computational demands, and its tendency to overfit data with high noise levels.

2.2.4 Naïve Bayes (NB):

This Bayes' theorem-based probabilistic classification technique estimates the likelihood of each class label using the input characteristics of data (John & Langley, 2013). It assumes that the input characteristics are conditionally independent for a given class label and simplifies the posterior probability computation. Naïve Bayes is frequently practised as the starting step for text classification and similar tasks when the input characteristics are discrete and

categorical. NB classifiers produce comparable outputs to artificial neural networks and decision tree classifiers. Interested readers may find details of Naïve Bayes as a text classifier in Vajjala et al. (2020).

The main benefits of Naïve Bayes include its rapid operation and ability to process both small and high-dimensional data efficiently. The feature independence assumption used by Naïve Bayes classifiers consistently fails to match real data patterns, yet its expressiveness remains limited.

2.2.5 Support Vector Machines (SVMs):

SVM is a supervised algorithm with high generalisation capability and adaptability to handle various applications. In 1995, Cortes and Vapnik created the fundamental SVM model to solve regression and classification problems involving numerous continuous and categorical variables. The algorithm aims to identify the hyperplane that achieves the highest possible margin to separate the data points into discrete classes effectively. The hyperplane that exhibits maximum generalisation performance on previously unseen data is called the optimal hyperplane. Support vectors refer to the data points near the hyperplane and play a crucial role in the classification process. SVMs can handle linear and non-linear classification issues using a kernel function that translates data into a higher-dimensional feature space (Vapnik, V. N., 1995). Text classification, picture classification, and bioinformatics are some areas where SVMs have been extensively employed.

Advantages of this model are effective in high-dimensional spaces, memory efficient, and versatile through kernel functions. This method necessitates a complex interpretation of the results, depends on proper parameter settings for optimal performance, and requires additional time to process large datasets.

2.2.6 Artificial Neural Networks (ANNs):

The potential applications for Artificial Neural Networks (ANNs) are vast and almost limitless. Artificial neural networks (ANNs) are often used in classification tasks, including a range of applications, such as speech and image recognition, identification of anomalous events, analysis of customer purchasing patterns, and several others. Regression analysis is a statistical technique for prediction, real-time optimisation, model predictive control, and several other applications (Alla et al., 2021).

A feedforward network (FFN) is a specific kind of artificial neural network (ANN) that exhibits a distinctive arrangement of nodes in its layers, mimicking a feedforward topology/structure. The Feedforward Neural Network (FFN) typically utilises a pre-defined activation function within its network. The Feedforward Neural Network (FNN) technology effectively adjusts inputs to get the desired output (Nath et al., 2019).

The term "Multilayer Perceptron (MLP)" refers to a kind of neural network that is characterised by its fully linked multilayer structure. It is a feedforward neural network characterised by a three-layer architecture, which includes an input layer, one or more hidden layers, and an output layer. The system's architecture exhibits a configuration consisting of interconnected nodes organised at many levels. Each layer within this arrangement is fully connected to the previous and subsequent layers (Del Frate et al., 2007). Multilayer perceptrons find their primary use in research applications, including domains such as mathematical modelling and engineering, among several others. To achieve the intended result, a transfer function inside a multilayer perceptron sends the summation of the inputs, which are assigned weights based on their values, to the activation layer (Padmavathi J, 2011).

The main advantages of MLPs include their ability to understand complex non-linear relationships, flexible architecture, and modelling capability for any function with ample neurones. The significant challenges for MLPs include large tuning needs and high computational demands without proper regularisation capabilities.

2.3 Model Development

During the model development stage in this study, we employed the stratified k-fold cross-validation technique for splitting the training data points. The approach described in the current section is a more refined version of the K-Fold method described by Allen et al. in their recent work (Allen et al., 2021). For the stratified k-fold cross-validation, we selected k = 10 based on Kohavi (1995) and Simon's (2007) recommendations, as this value produces nearly unbiased estimates of prediction error while maintaining computational efficiency. This approach ensured that each fold maintained the proportional representation of accident categories found in the overall dataset, addressing potential bias due to the unbalanced nature of mining accident distributions.

While our initial experiments with default hyperparameters yielded satisfactory results, we conducted limited hyperparameter exploration to assess potential performance improvements. For the MLP model, which demonstrated superior performance, we tested learning rates (0.001, 0.01), hidden layer configurations (single layer with 50, 100, and 150 neurones), and activation functions (ReLU, tanh). The final MLP architecture consisted of an input layer corresponding to our word embedding dimensions (300), a hidden layer with 100 neurones using ReLU activation, and an output layer with SoftMax activation for the eight accident categories. The Adam optimiser was used with a learning rate of 0.001 and categorical cross-entropy as the loss function. Early stopping was implemented with a patience of 5 epochs, monitoring validation loss to prevent overfitting.

For other algorithms, we used parameters from the Scikit-learn implementation but adjusted key settings: for SVM, we tested linear and RBF kernels; for Random Forest, we explored various numbers of estimators (100, 200, 500); and for Logistic Regression, we examined different regularisation strengths. More extensive hyperparameter optimisation through grid search or random search could potentially yield further improvements, but our primary focus was on comparing the relative performance of different algorithmic approaches rather than exhaustive optimisation.

Our dataset exhibits considerable class imbalance, with non-winding transportation machinery accidents accounting for approximately 50% of all cases. To address this imbalance and prevent model bias, we implemented class weights inversely proportional to class frequencies during model training. This approach penalises misclassification of minority classes more heavily, ensuring the model remains sensitive to less common but equally important accident categories.

To ensure robust evaluation while preventing data leakage, we implemented a two-stage validation approach. First, we divided the complete dataset into an 80% training set and a 20% held-out test set using stratified sampling to maintain class distributions. The training set was then subjected to stratified 10-fold cross-validation for model development and comparison, while the held-out test set remained untouched until final evaluation. This approach preserves the integrity of the test set while leveraging the benefits of cross-validation for model selection and hyperparameter tuning. All reported performance metrics represent results on the held-out test set that was not used during model development, ensuring unbiased evaluation of model generalisation capabilities. The weighted F1 score was selected as our primary evaluation metric specifically because it accounts for class imbalance when assessing model performance. The following section provides a concise further description of the developed model's results.

3.0 Results and Discussions

The efficacy of classification models is evaluated using a set of measures like accuracy, precision, recall, and F1 score (Vajja et al., 2020). The confusion matrices used to find these measures are presented below. In the confusion matrix (Fig. 2), horizontal rows represent the truth or actual data, and vertical columns represent the predicted data while evaluating a classification model with test data.

The accuracy of a model is expressed as a percentage of all correct predictions, while precision demonstrates how accurate or precise the model's predictions are, i.e., how many of the positive instances the model can accurately identify given all the positive examples. Accuracy and precision also illustrate how comprehensive the model is, and they are expressed as Eq. (1) and Eq. (2), respectively, based on the confusion matrix.

Accuracy(A)= (TP + TN)/(TP + TN + FP + FN) (1)

Precision(P) = TP/(TP + FP) (2)

Recall is a supplement to Precision and demonstrates how effectively the model can recall positive classes given in Eq. (3).

Recall(R) = TP/(TP + FN) (3)

Similarly, the F1-score, introduced by Buckland et al. (1994), combines recall and Precision to produce a single statistic that reflects model performance and is expressed as Eq. (4).

F1 Score(F1) = (2(P×R))/(P + R) (4)

We preferred to assess model performance using the weighted average F1 score, as it can effectively control the effect of an unbalanced dataset for a specific label. Eq. (5) presents the expression for the weighted average F1 score.

$$AvgF1_{weighted} = \sum_{i=1}^{N} \left(\frac{S_i}{T} * F1_i\right) \quad (5)$$

- T is the support of all labels,
- Si = number of cases supported by label i,
- F1i = F1 score of label i.
- N = total number of labels.

The confusion matrix and the performance results for each classifier model are displayed in Table 3, while Table 4 shows the accuracy, recall, F1 score, weighted average F1 score, and support for each label. Table 4 indicates each model's F1 score for each cause, and the overall model's weighted average F1 score is in bold.

Each machine learning algorithm employed in this study offers distinct advantages and limitations in the context of mining accident classification. The MLP model demonstrated superior overall performance (F1 score: 0.84) due to its ability to capture complex, non-linear relationships in textual data through its hierarchical structure. This architecture is particularly well-suited for identifying subtle linguistic patterns that distinguish different accident categories, especially when terminology varies but semantic meaning remains similar. The exceptional performance of MLP for 'Non-Winding Transportation Machinery', ‘Ground Vibration’ & ‘Explosives’ accidents, indicated by an F1 score of 0.91, 0.92 & 0.93 respectively, suggests that these reports contain complex language patterns which benefit from deep feature learning.

Logistic regression (F1 score: 0.83) offers the advantage of computational efficiency and interpretability, performing particularly well on categories with distinctive vocabulary. Its strong performance suggests that many accident narratives contain linearly separable features despite the complexity of natural language. SVM (F1 score: 0.81) also does well because it works effectively with the complex nature of text data, especially excelling in classifying 'Electrical cause' (F1 score: 0.89), where the use of technical terms helps make the distinctions more straightforward.

The relatively poorer performance of decision trees (F1 score: 0.57) reflects their tendency to overfit on text data and difficulty capturing the semantic nuances in accident narratives. Random Forest (F1 score: 0.70) mitigates some of these overfitting issues through ensemble learning, but still falls short of the more sophisticated models. Naïve Bayes (F1 score: 0.60) underperforms due to its strong independence assumptions, which are frequently violated in natural language, where word co-occurrences carry significant meaning.

Table 3
Confusion Matrix of Classification Models used in this study.

		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	Support by
		Logistic regression								SVM								
True values	1	**489**	15	10	2	0	1	0	1	**488**	15	11	1	0	2	0	1	**518**
	2	43	**149**	11	4	1	0	0	2	48	**138**	18	4	0	0	0	2	**210**
	3	18	22	**125**	12	5	2	0	2	17	22	**120**	21	3	2	0	1	**186**
	4	2	0	12	**166**	0	0	1	1	1	1	10	**169**	0	0	0	1	**182**
	5	3	4	6	0	**70**	1	0	1	2	6	7	0	**70**	0	0	0	**85**
	6	5	3	6	1	0	**29**	0	1	6	4	10	3	0	**22**	0	0	**45**
	7	1	4	0	0	0	1	**38**	0	1	5	2	1	0	0	**35**	0	**44**
	8	1	3	2	2	0	4	0	**26**	3	0	4	5	0	3	0	**23**	**38**
		Predicted values																1308
		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	Support by
		Decision Tree								Random Forest								
True values	1	**377**	74	34	4	9	8	3	8	**494**	13	8	2	0	1	0	0	**518**
	2	64	**80**	32	6	6	8	2	12	108	**87**	11	4	0	0	0	0	**210**
	3	40	33	**68**	15	13	6	7	4	41	19	**107**	16	2	1	0	0	**186**
	4	9	5	23	**139**	2	1	1	2	5	2	8	**167**	0	0	0	0	**182**
	5	9	17	11	3	**41**	1	3	0	7	9	8	0	**61**	0	0	0	**85**
	6	10	4	10	0	4	**13**	1	3	17	7	11	4	1	**5**	0	0	**45**
	7	5	5	5	1	1	3	**22**	2	12	5	4	0	1	0	**22**	0	**44**
	8	11	6	4	1	1	3	3	**9**	11	8	8	1	0	0	0	**10**	**38**
		Predicted values																1308
		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8	Support by
		Naïve Bayes								MLP								
True values	1	**490**	14	10	4	0	0	0	0	**481**	21	12	1	1	1	0	1	**518**
	2	156	**40**	10	3	1	0	0	0	38	**152**	12	4	1	1	0	2	**210**
	3	68	13	**83**	21	0	1	0	0	14	26	**124**	9	6	5	0	2	**186**
	4	8	2	10	**162**	0	0	0	0	0	3	9	**168**	0	0	1	1	**182**
	5	18	6	7	0	**52**	1	0	1	4	2	4	0	**73**	1	0	1	**85**
	6	23	6	8	4	0	**4**	0	0	3	2	6	1	0	**33**	0	0	**45**
	7	22	3	5	0	0	0	**14**	0	1	3	0	0	0	0	**40**	0	**44**
	8	24	1	7	4	0	0	0	**2**	2	1	2	1	0	4	1	**27**	**38**
		Predicted values																1308

Table 4
Precision, Recall and F1score of Classification Models used in this study.

	Logistic Regression			SVM			Naïve Bayes			Decision Tree			Random Forest			MLP		
Type of Accident	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Non-Winding Transportation Machinery	0.87	0.94	**0.91**	0.86	0.94	0.90	0.61	0.95	0.74	0.72	0.73	0.72	0.71	0.95	0.81	0.89	0.93	**0.91**
Other than Transportation Machinery	0.75	0.71	**0.73**	0.71	0.66	0.68	0.44	0.19	0.27	0.36	0.38	0.37	0.58	0.41	0.48	0.72	0.72	0.72
Fall of Objects	0.73	0.67	**0.70**	0.66	0.65	0.65	0.61	0.45	0.51	0.36	0.37	0.36	0.65	0.58	0.61	0.73	0.67	**0.70**
Ground Vibrations	0.88	0.91	0.90	0.85	0.93	0.89	0.83	0.89	0.86	0.82	0.76	0.79	0.86	0.92	0.89	0.91	0.92	**0.92**
Electrical	0.92	0.82	0.87	0.96	0.82	**0.89**	0.98	0.61	0.75	0.53	0.48	0.51	0.94	0.72	0.81	0.90	0.86	0.88
Other Cause	0.76	0.64	0.70	0.76	0.49	0.59	0.67	0.09	0.16	0.30	0.29	0.30	0.71	0.11	0.19	0.73	0.73	**0.73**
Explosive	0.97	0.86	0.92	1.00	0.80	0.89	1.00	0.32	0.48	0.52	0.50	0.51	1.00	0.50	0.67	0.95	0.91	**0.93**
Dust, Gas other combustible matters	0.76	0.68	0.72	0.82	0.61	0.70	0.67	0.05	0.10	0.23	0.24	0.23	1.00	0.26	0.42	0.79	0.71	**0.75**
Weighted Average Score	0.83	0.83	**0.83**	0.81	0.81	**0.81**	0.65	0.65	**0.60**	0.58	0.57	**0.57**	0.73	0.73	**0.70**	0.84	0.84	**0.84**

This comparative analysis demonstrates that neural network approaches are particularly well-suited to the classification of mining accident texts, likely due to their ability to model complex relationships in natural language without requiring explicit feature engineering.

4.0 Conclusions and future directions

The mining industry is dynamic and prone to incidents and accidents, making proper assessment of safety measures a critical concern. However, manually processing accident data is time-consuming and labour-intensive. As a result, a more successful strategy is required to raise assessment quality.

Our comparative analysis conclusively demonstrates that the MLP model outperforms traditional classification algorithms for mining accident categorisation, achieving a weighted F1 score of 0.84, followed closely by logistic regression (0.83) and SVM (0.81). These results establish that neural network approaches are particularly well-suited to classifying texts related to mining safety, likely due to their ability to capture complex linguistic patterns and contextual relationships present in accident narratives. The substantial performance gap between these top-performing models and the decision tree approach (0.57) highlights the importance of algorithm selection in safety-critical classification tasks. This hierarchy of model performance provides clear guidance for safety professionals seeking to implement automated classification systems for accident prevention.

The novelty of this research lies in its comprehensive comparison of multiple machine learning approaches for classifying texts about mining accidents—a domain-specific application not previously explored in the literature. Even though the algorithms used are well-known, comparing them in this specific area gives important information for safety experts. This study bridges the gap between theoretical text mining capabilities and practical mining safety applications, establishing a foundation for more advanced predictive systems that can potentially save lives in this high-risk industry.

Future work direction:

Although this approach's results are promising, several improvements are feasible in future work. Future research should expand the dataset scope to include non-fatal accidents and near-miss incidents, which could provide valuable insights into accident prevention. The integration of additional data sources, such as safety inspection reports, worker interviews, and international mining accident data, would enhance the generalisability of our findings and potentially reveal cross-cultural safety patterns. Exploring unsupervised techniques, such as topic modelling and clustering, our supervised approach reveals latent patterns in accident narratives without relying on predetermined categories. Additionally, implementing explainable AI techniques would improve the interpretability of model predictions, making them more actionable for safety professionals. Last but not least, creating multi-label classification techniques would maintain the beneficial aspects of our current approach while more precisely reflecting the multi-causal nature of mining accidents.

Declarations

Acknowledgment: We would like to express our profound gratitude to the Directorate General of Mines Safety for providing us with the necessary information.

Funding: The authors declare that no funding was granted for the study.

Declaration of interest statement: The authors declare no conflicts of interest.

References

1. Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. Machine Learning, 6(1), 37–66. https://doi.org/10.1023/a:1022689900470
2. Alla, H., Moumoun, L., & Balouki, Y. (2021). A Multilayer Perceptron Neural Network with Selective-Data Training for Flight Arrival Delay Prediction. *Scientific Programming*, *2021*, 1–12. https://doi.org/10.1155/2021/5558918
3. Allen, J., Liu, H., Iqbal, S., Zheng, D., & Stansby, G. (2021). Deep learning-based photoplethysmography classification for peripheral arterial disease detection: a proof-of-concept study. *Physiological Measurement*, *42*(5), 054002. https://doi.org/10.1088/1361-6579/abf9f3
4. Amyotte, P. R., Pegg, M. J., & Khan, F. I. (2008). Application of inherent safety principles to dust explosion prevention and mitigation. *Process Safety and Environmental Protection*, *87*(1), 35–39. https://doi.org/10.1016/j.psep.2008.06.007
5. Arra, K., Gunda, Y. R., & Gupta, S. (2023). Development of a predictive model for workers' involvement in workplace accidents in an underground coal mine. *Sadhana*, *48*(2). https://doi.org/10.1007/s12046-023-02121-3
6. Buckland, M., & Gey, F. (1994). The relationship between Recall and Precision. Journal of the American Society for Information Science, 45(1), 12–19. https://doi.org/10.1002/(sici)1097-4571(199401)45:1
7. Chen, H., Qi, H., Long, R., & Zhang, M. (2011). Research on 10-year tendency of China coal mine accidents and the characteristics of human factors. *Safety Science*, *50*(4), 745–750. https://doi.org/10.1016/j.ssci.2011.08.040
8. Chen, N., Chiang, N., & Storey, N. (2012). Business Intelligence and Analytics: From big data to Big impact. *MIS Quarterly*, *36*(4), 1165. https://doi.org/10.2307/41703503
9. Cheng, M., Kusoemo, D., & Gosno, R. A. (2020). Text mining-based construction site accident classification using hybrid supervised machine learning. *Automation in Construction*, *118*, 103265. https://doi.org/10.1016/j.autcon.2020.103265
10. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018
11. Debortoli, S., Müller, O., & Brocke, J. V. (2014). Comparing business intelligence and big data skills. *Business & Information Systems Engineering*, *6*(5), 289–300. https://doi.org/10.1007/s12599-014-0344-2
12. Del Frate, F., Pacifici, F., Schiavon, G., & Solimini, C. (2007). Use of neural networks for automatic classification from High-Resolution images. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(4), 800–809. https://doi.org/10.1109/tgrs.2007.892009
13. Ghasemi, E., Ataei, M., Shahriar, K., Sereshki, F., Jalali, S. E., & Ramazanzadeh, A. (2012). Assessment of roof fall risk during retreat mining in room and pillar coal mines. *International Journal of Rock Mechanics and Mining Sciences*, *54*, 80–89. https://doi.org/10.1016/j.ijrmms.2012.05.025
14. Hosmer, D. W., Jr, & Lemeshow, S. (2004). *Applied Logistic Regression*. John Wiley & Sons.
15. Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text Classification Using Machine Learning Techniques. *Research Gate*. https://www.researchgate.net/profile/V_Tampakas/publication/228084521_Text_classification_using_machine_learning_techniques/links/0c96051ee1d
16. Ismail, S. N., Ramli, A., & Aziz, H. A. (2021). Research trends in mining accidents study: A systematic literature review. *Safety Science*, *143*, 105438. https://doi.org/10.1016/j.ssci.2021.105438
17. John, G. H., & Langley, P. (2013). Estimating continuous distributions in Bayesian classifiers. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1302.4964
18. Jung, Y. (2017). Multiple predictingK-fold cross-validation for model selection. *Journal of Nonparametric Statistics*, *30*(1), 197–215. https://doi.org/10.1080/10485252.2017.1404598
19. Kecojevic, V., Komljenovic, D., Groves, W., & Radomsky, M. (2006). An analysis of equipment-related fatal accidents in U.S. mining operations: 1995–2005. *Safety Science*, *45*(8), 864–874. https://doi.org/10.1016/j.ssci.2006.08.024
20. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model 99 selection. *International Joint Conference on Artificial Intelligence*, *2*, 1137–1143. 100 Retrieved from http://ijcai.org/Proceedings/95-2/Papers/016.pdf
21. Kumar, P., Gupta, S., & Gunda, Y. R. (2019). Estimation of human error rate in underground coal mines through retrospective analysis of mining accident reports and some error reduction strategies. *Safety Science*, *123*, 104555. https://doi.org/10.1016/j.ssci.2019.104555
22. López-Úbeda, P., Díaz-Galiano, M. C., Martín-Noguerol, T., Ureña-López, A., Martín-Valdivia, M., & Luna, A. (2020). Detection of unexpected findings in radiology reports: A comparative study of machine learning approaches. *Expert Systems With Applications*, *160*, 113647. https://doi.org/10.1016/j.eswa.2020.113647
23. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, January 16). *Efficient estimation of word representations in vector space*. arXiv.org. http://arxiv.org/abs/1301.3781
24. Ministry of Coal. (2023). India Achieves 47 % growth in Coal Production during last Nine Years. https://coal.nic.in/sites/default/files/2023-05/PIB1925521.pdf Date of Access: 20.05.2023
25. Mujalli, R. O., López, G., & Garach, L. (2015). Bayes classifiers for imbalanced traffic accidents datasets. *Accident Analysis & Prevention*, *88*, 37–51. https://doi.org/10.1016/j.aap.2015.12.003
26. Nath, A. S., Pal, A., Mukhopadhyay, S., & Mondal, K. C. (2019). A survey on cancer prediction and detection with data analysis. *Innovations in Systems and Software Engineering*, *16*(3–4), 231–243. https://doi.org/10.1007/s11334-019-00350-6
27. Onan, A. (2016). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, *44*(1), 28–47. https://doi.org/10.1177/0165551516677911

28. Onan, A., & Korukoğlu, S. (2015). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25–38. https://doi.org/10.1177/0165551515613226
29. Padmavathi, J. (2011). A Comparative Study on Breast Cancer Prediction Using RBF and MLP. 135 International Journal of Scientific & Engineering Research, Volume 2, Issue 1
30. Sanmiquel, L., Rossell, J. M., & Vintró, C. (2015). Study of Spanish mining accidents using data mining techniques. *Safety Science*, *75*, 49–55. https://doi.org/10.1016/j.ssci.2015.01.016
31. Shah, K. S., Khan, M. A., Khan, S., Rahman, A., Khan, N. M., & Abbas, N. (2020). Analysis of underground mining accidents at Cherat Coalfield, Pakistan. *International Journal of Economic and Environmental Geology*, *11*(1), 113–117. https://doi.org/10.46660/ojs.v11i1.422
32. Sharma, D., Ghosh, M.K., & Parida, P.K. (2019). An analytical study on fatal accidents in Indian coal mines. Journal of Mines, Metals and Fuels, 67(5), 207-212. http://doi.org/10.1016/j.jmmf.2019.05.007.
33. Simon, R. (2007). Resampling Strategies for Model Assessment and Selection. In *Springer* 173 *eBooks* (pp. 173–186). https://doi.org/10.1007/978-0-387-47509-7_8
34. Statistics of Mines in India, Volume-I (Coal), 1995 – 2015. Delhi: Controller of Publication
35. Toraman, S., Alakus, T. B., & Turkoglu, I. (2020). Convolutional capsnet: A novel artificial neural network approach to detect COVID-19 disease from X-ray images using capsule networks. *Chaos Solitons & Fractals*, *140*, 110122. https://doi.org/10.1016/j.chaos.2020.110122
36. Tripathy, D. P., & Ala, C. K. (2018). Identification of safety hazards in Indian underground coal mines. *Journal of Sustainable Mining*, *17*(4), 175–183. https://doi.org/10.1016/j.jsm.2018.07.005
37. Ubeynarayana, C. U., & Goh, Y. M. (2017). An ensemble Approach for classification of accident narratives. *Computing in Civil Engineering*. https://doi.org/10.1061/9780784480847.051
38. Vajjala, S., Majumder, B., Gupta, A., & Surana, H. (2020). *Practical natural language processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly Media.
39. Vapnik, V. N. (1995). The nature of statistical learning theory. In *Springer eBooks*. https://doi.org/10.1007/978-1-4757-2440-0
40. Xiang, J., Bi, P., Pisaniello, D., & Hansen, A. (2013). Health Impacts of workplace heat exposure: An Epidemiological review. *Industrial Health*, *52*(2), 91–101. https://doi.org/10.2486/indhealth.2012-0145
41. Yin, W., Fu, G., Yang, C., Jiang, Z., Zhu, K., & Gao, Y. (2016). Fatal gas explosion accidents on Chinese coal mines and the characteristics of unsafe behaviors: 2000–2014. *Safety Science*, *92*, 173–179. https://doi.org/10.1016/j.ssci.2016.09.018
42. Zhang, J., Cliff, D., Xu, K., & You, G. (2018). Focusing on the patterns and characteristics of extraordinarily severe gas explosion accidents in Chinese coal mines. *Process Safety and Environmental Protection*, *117*, 390–398. https://doi.org/10.1016/j.psep.2018.05.002
43. Zhang, J., Xu, K., Reniers, G., & You, G. (2019). Statistical analysis of the characteristics of extraordinarily severe coal mine accidents (ESCMAs) in China from 1950 to 2018. *Process Safety and Environmental Protection*, *133*, 332–340. https://doi.org/10.1016/j.psep.2019.10.014
44. Zhu, Y., Wang, D., Shao, Z., Xu, C., Zhu, X., Qi, X., & Liu, F. (2018). A statistical analysis of coalmine fires and explosions in China. *Process Safety and Environmental Protection*, *121*, 357–366. https://doi.org/10.1016/j.psep.2018.11.013

Figures

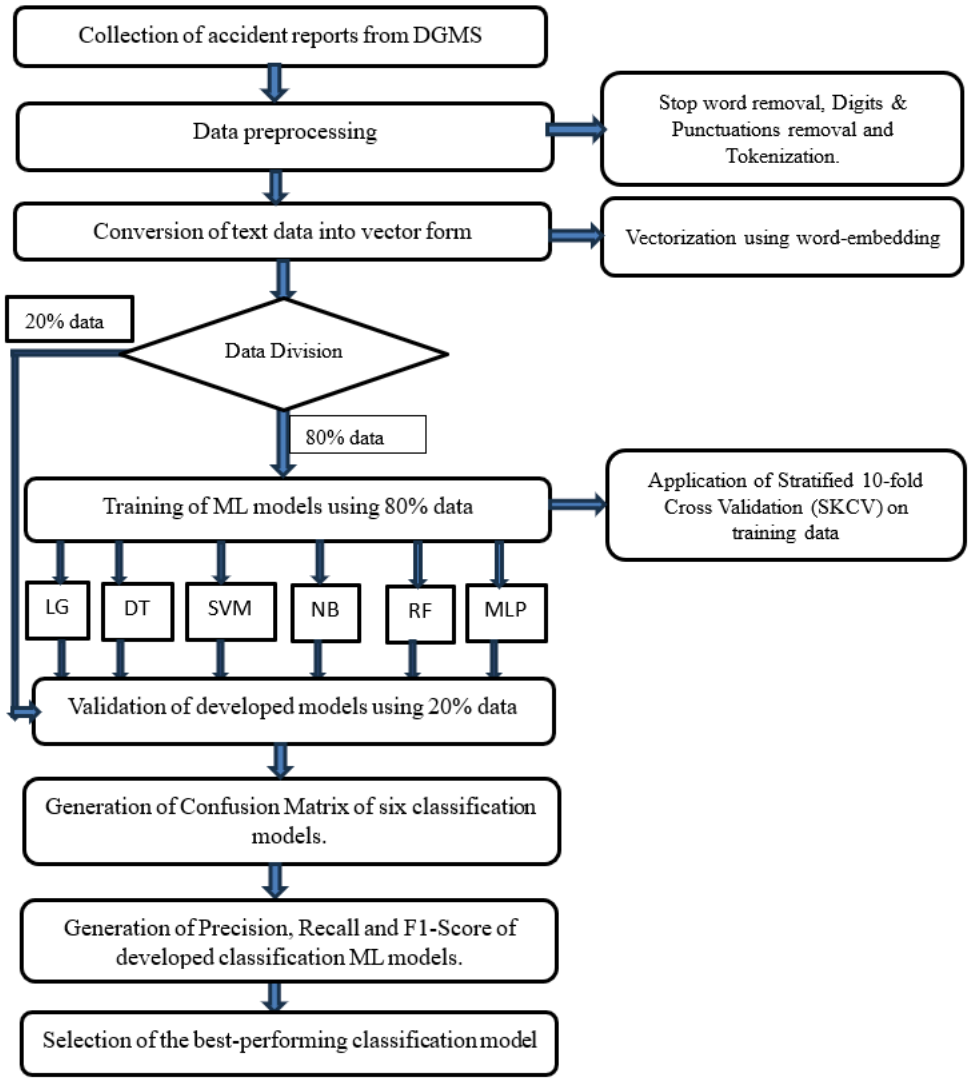


Figure 1

Methodology of classification of fatal accident reports based on the category/type.

True Values		Predicted Values: Yes	Predicted Values: No
	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

Figure 2

Confusion matrix representation