Workplace Incident and Injuries Prevention Using Machine Learning

Arti Deshpande[1](✉) and Arya Kumar[2]

[1] Thadomal Shahani Engineering College, Mumbai, India
arti.deshpande@thadomal.org
[2] University of Georgia, Athens, GA, USA

Abstract. Today in the area of field operations, there is no systematic way to assess and identify if any of the active or pipeline assignments are prone to mishaps, illness, and injuries. Actionable insights are missing from the leading indicators like concern reports, near-misses, or work-stop events reported by the employees. Illness and Injuries (I&I) incidents result in temporary or permanent loss of valuable human assets that are expensive and difficult to replace. Based on historic statistical analysis, injuries and Illnesses are extreme events in operations. Such a highly unbalanced distribution of data makes these events highly unpredictable. Besides unbalanced distribution, normal and incident cases within operations overlap in their characteristics. Incidents and normal cases share a high level of commonalities and therefore are difficult to be separated by any clear decision boundary.

Deep-learning & AutoML framework-based Machine Learning algorithms bring the required computational power to assess and minutely study the characteristics represented in I&I incident vs normal records and can help identify the root cause and segregate them. Anomaly detection is another Machine Learning technique that allows the identification of unusual patterns that are not expected (also referred to as outliers). Considering I&I incidents as anomalies, this paper has given anomaly detection algorithms to separate such incidents from normal events.

Keywords: Machine Learning · Health · and Safety · Injuries and Illnesses · Workplace Safety · Preventive Analytics

1 Introduction

Its duty of employer to confirm a safe and healthy workplace for their employees as well as for any person who is visiting the workplace. They should provide a safe workspace without risk to health. There must be some controlling measures for risk and accidents so that they can be eliminated or reduced at a certain level. If the reasons for injuries and accidents happening at the workplace can be identified then they can be eliminated or proper training and information can be provided to employees to avoid those. Those reasons can be identified and controlled if reporting of injuries, accidents, health issues of employees, diseases, etc. is done properly and records are maintained.

S. Tamane et al. (Eds.): ICAMIDA 2022, ACSR 105, pp. 499–512, 2023.
https://doi.org/10.2991/978-94-6463-136-4_43

The safety, health, and welfare of people patronizing any organization is the responsibility of Occupational Safety and Health (OSH). Ima Ilyani Ibrahim at. al. [1] suggested creating a safety culture in the office by examining how the office environment, management commitment, staff attitudes, and organizational policies are related to workplace safety. Correlation and multiple regression are used to analyze the data related to higher learning institutions. The findings of their study showed that policies and procedures are meaningfully related and should be introduced to the employees.

The research of Hood Bin Atan [2] is based on the wood-based related manufacturing industry and suggested that cost components should be identified, defined, and classified for occupational accidents happening at the workplace. He has proposed a risk prevention plan based on direct to indirect costs of occupational accidents and prevention. So, the author has given only an analysis of injuries and accidents that happened at the manufacturing company based on historical data and the risk plan is generated based on the cost factor. Manikandan Krishnamurthy at. el. [3] has given safety measures for workers who are working in steel industries where the temperature is hot. This hot and harsh temperature creates well-known risks of heat-related illnesses and output decreases. Occupational heat exposures and the risks associated with them have been studied in the past and present.

Health and safety must be taken care and the well-being of employees and employers should be promoted in all industries. Employee protection at the workplace during the job is the moral responsibility of the company [13]. An econometric model developed by Vani K. Borooah et al. [14] is based on variables reflecting worker characteristics and the institutional and legislative environment, one can calculate the number of workplace injuries in Queensland. This model shows the ratio of reported injuries to actual workplace injuries. They have used the Queensland Employee Injury Database. The findings of this research help to make policy on workplace health and safety.

The study was done by Indecon International Economic Consultants [15] to evaluate the costs incurred by Irish small businesses/employers due to work-related injuries in retail, hospitality, manufacturing, and other services sectors. The survey of 809 small businesses with a varying number of employees was considered to assess workplace injuries. It is concluded that the cost of workplace injuries was very significant and it affects the overall business cost. The research by Lu et al. [16] analytically evaluates safety climate and safety behavior in the passenger ferry context. Data collected using a survey from 155 respondents working for passenger ferry companies in Taiwan is used for experimental and analysis purposes. The author used hierarchical regression to the analysis of how safety climate affects self-reported safety behaviors. Zubaidah Ismail et al [17] study highlighted the influential safety factors which affect the safety management system for construction sites. Those safety factors are collected through interviews with related people at construction sites. The author suggested some designs of equipment to improve the efficiency and productivity of construction workers. Michael S. Christian et al. [18] imply that crucial aspects of workplace safety include both the individual and the situation. It is possible to recruit, teach, and support workers through a positive safety climate to enhance safety motivation and knowledge. This results in safe behaviors and fewer accidents and injuries.

In this study, machine learning algorithms are used to predict the risk to prevent injuries and illness at the workplace. A comparative study of various ML models is given as an experimental result. These proposed methods and techniques can be applied for any workplace provided systematic data is available. Methodical data preprocessing is done before the experiment.

2 Dataset Description

In order to provide a safer work environment, a deeper assessment of conditions and settings that lead to injuries and illnesses is required so those can be avoided. Dataset is prepared by collecting the key data elements that can help us to predict and prevent workplace injuries and illnesses. It is identified that Employee behavior, type of work, hours of work, equipment and activities, plant or place of execution, employee's business function, and any associated leading indicators like concern reports, inspections, or near-misses are key data attributes that can be used to train the machine. Figure 1 explains the various data attributes related to questions that are considered to collect and prepare the training dataset.

The hourly employee time card dataset has a total of 8.93 million records that links individual employee with project or assignment, type of work, place of work, hours of work, and other key attributes like the expertise level of the employees by date. This was further enriched by adding project or assignment descriptions, customer details, condition of the workplace, type of work facility, and the location of the assignment. A total of 8.93 million records for the period of August 2017 – May 2019 were provided. Out of those 8.93 million records, approx.15000 transactions are labeled as incidents and belong to "Incident Class". Feature 'Incident Happened' is the response variable and it takes value 1 in case of an incident and 0 otherwise. The subset data from September 2018 and February 2019 is used to validate the performance and accuracy of various models. The following data attributes are used to train various algorithms. Table 1. Provides the name and description of every input data attribute. Table 2. Provide engineered columns from the "work_date" that help to identify if there is any specific month, weekday, or weekend when most incidents are reported (a.k.a. seasonality).

Three additional calculated variables were used as described in Table 2.

Our goal is to correctly predict which of the timecards for Sept 18 and Feb 19 had an I&I incident associated with them. The performance and accuracy were measured based on the Precision and Recall matrix. Recall or accuracy is the number of incidents the model can correctly predict and precision or false alarm rate is the number of benign incidents the model incorrectly labels as having an incident [6].

The objective is to increase the model's validity so workplace stoppages are included in the training. Which gives the work on the premises temporarily halted and added to the dataset as an additional label improving the minority "incidents" class (change this sentence). Using a classification model to discern unusual occurrences along with balancing the data with sampling methods allows the machine to identify the labeled "incidents" as significant, giving these "incidents" more weight in the final accuracy of the model. Another objective is to identify if any of the reported incidents by the employees (concern-reports) may lead to a future injury or illness (Near-Misses, Stop-Work, or I&I) event.

Table 1. Input Data Attributes used for Training the models

Input Data Column	Description
Employee ID	Unique identifier for every employee
Work Org	Organization employee works for
Employee Title	Employee's job title
Department	Department within the Organization
Type of Work	Type of work being performed. Examples - Field Services, Repair, Maintenance, Installation, etc.
Expertise Level	Expertise level of employee. Examples -Engineer, Technician, Senior Technician, Supervisor, etc.
Project Number	Unique identifier for the assignment that the employee is assigned to
Company cd	Company code under which the organization operates
Description of Work	Description of the assignment that the employee is assigned to
Supervisor	Supervisor of the employee
Customer	Customer to whom the services are provided
Engagement Type	Terms of Services. Examples - Temporary, Long Term, Seasonal, etc.
Customer Desc	Description of customer's business function whom the services are provided
Customer Name	Name of the customer
Location Name	Name of the location where the task will be performed
Location Type	Type of the work location
Address Line 1	The physical address of the location where the task will be performed
Address Line 2	The physical address of the location where the task will be performed
City	The physical address of the location where the task will be performed
State	The physical address of the location where the task will be performed
Postal Code	The physical address of the location where the task will be performed
Country Code	Physical address of the location where the task will be performed
Region	Physical address of the location where the task will be performed
Work Condition	Condition of the location where the task will be performed
Straight Time Hours	Total routine hours reported by the employee rolled up from his/her daily time card
Overtime Hours	Total overtime hours reported by the employee rolled up from his/her daily time card
Other Hours	Any break, holidays, time-away hours reported by the employee rolled to from his/her daily time card
Work Date	Date for which the employee reports the hours in his/her time card.

Table 2. Feature Engineered columns in order to extract any seasonality associated with the I&I events

Work Month	Engineered column from Work Date
Weekday	Engineered column from Work Date
Work Year	Engineered column from Work Date

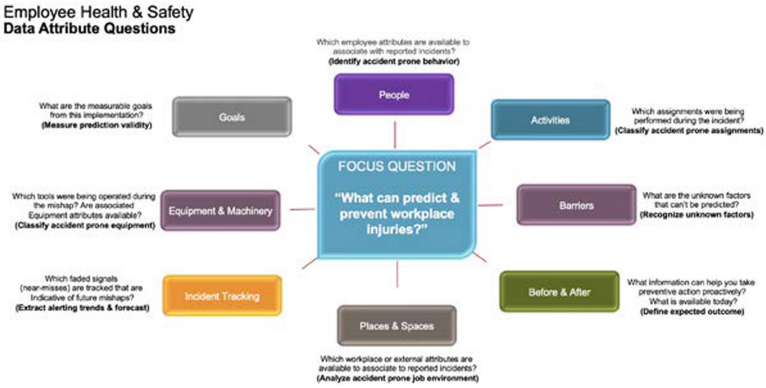

Fig. 1. Data attribute questionnaire for training data collection

The primary goal is to take into account the employee's concerns within their "Concern reports". These reports reported by the employees help to validate the occurrence of near misses, work stoppages, and I&I reports. A "Concern report" may begin with a complaint or grievance against a workplace condition or work task. The employee may feel like their safety has been compromised and therefore have filed a concern report. If this report is attended to and fixed immediately, then the issue is dissolved, however, if this concern persists, it may become a near miss, in which an alarming event has almost occurred. If still remains unchecked, this can further develop into a stoppage of work in which some small disturbance has occurred, causing work to be halted. If even these fall into neglect, there is a high chance of an illness & injury report. The concern that arises from the employee can result directly in an I&I or any other workplace disruption. These reportable events can be used to predict the possible occurrence of I&Is, which is the goal of our study. The hypothesis was that Characteristics represented within historic Recordables and First Aid (I&I) can be utilized to correlate and predict if any of the Reportable events (e.g. Near Misses) can turn into future illness or injury. A different dataset with a total of 43,738 records that contains 6,519 labeled near-misses, stop-work, and I&I cases is provided for experimental purposes.

3 Algorithms Considered

As stated earlier, the hourly employee dataset has a total of 8.93 million records. Out of the total of 8.93 million records, approximately 15000 transactions are labeled as incidents

and belong to "Incident Class". Because of this, the dataset is highly imbalanced (with only 0.16% of total transactional data belonging to "Incident Class"). Feature 'Incident Happened' is the response variable and it takes value 1 in case of an incident and 0 otherwise. It was important to handle this problem with several available techniques. Given the imbalance present, Anomaly detection algorithms [4–6] and classification algorithms [5] trained on the sampled dataset are most effective instead of the classical classification approach.

Anomaly detection is a method for finding odd patterns that deviate from expected behavior. It has several uses, including problem detection in operational environments and fraud detection in credit card transactions. In this study, an I&I incident as an unexpected behavior or an outlier is considered. Considering the minor class ("injury, illness, near miss, stop-work") as the outlier class to separate it from the normal ("no injury, routine day") class from the provided dataset.

Following specific ML libraries are used to conduct various experiments:

A. Isolation Forest using Sklearn [7]
B. Autoencoders using Deep Learning (H2O) [8–11]
C. Classification using H2O with sampled (balanced) dataset [10]
D. PU (Positive-unlabeled) Classification semi-supervised learning

PU (Positive-unlabeled) semi-supervised learning using Deep Learning Auto ML Classifier [9] is used to identify which of the leading indicators (Concern Reports) may lead to an I&I incident ("injury, illness, near miss, stop-work").

A. **Isolation Forest**

Isolation Forests are one of the newest methods for finding abnormalities. The technique is based on the idea that anomalies are made up of a small number of distinct data points. These characteristics make anomalies prone to a technique known as isolation. The steps of the algorithm are given in Fig. 2.

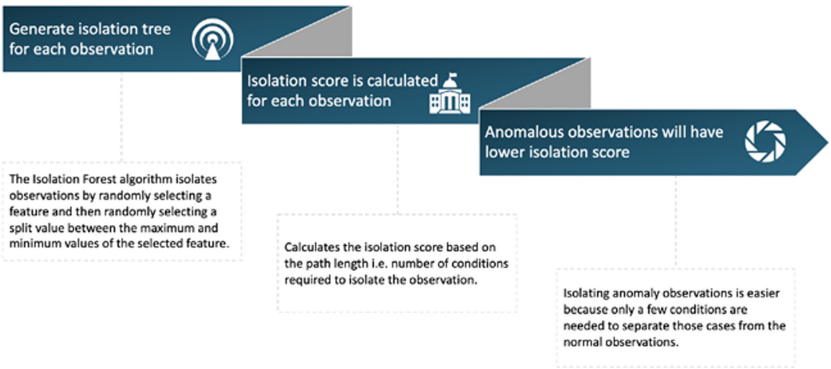

Fig. 2. Steps of Isolation Forest to detect anomalies

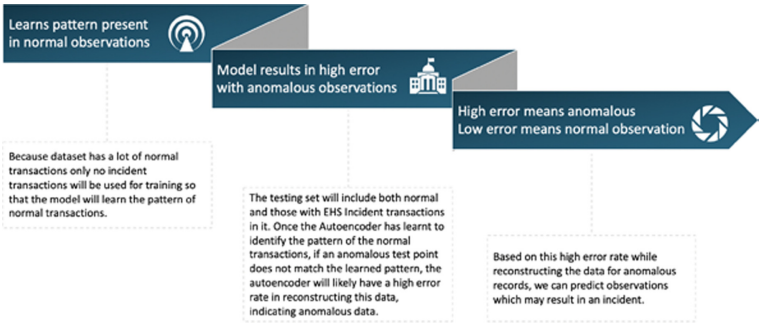

Fig. 3. How Autoencoder detects anomalies

B. **Autoecoder Deep Learning**

An unsupervised neural network uses autoencoders. It is a data compression algorithm that receives the input, processes it through a compressed form, and then produces the output after reconstruction. The object of an autoencoder is to study a representation (encoding) or pattern for a set of data ignoring the noise. The steps involved are shown in Fig. 3.

III. **Automl Classification Using Sampled Data**

H2O Framework Classifiers are used to predict after using its "balanced_classes" function to balance the dataset. An experiment was conducted excluding the "stop-work" event category from the output label and also with the additional "stop-work" event category data added to the output label in the training dataset to uplift the minority "Incident" labeled class further.

Decision tree models and ensemble methods are used to study the input attributes and their relation to the target variable (output label). These methods are built on feature importance, and work by dividing the data into groups that disproportionately represent one class. The tree will keep creating new subsets until it fully comprehends and reflects the link between the variables and the target. Then the dataset is balanced by using the under-sampling technique. The balance classes option within the H2O AutoML platform is used to balance the classes. Further train various classifier models to predict using Auto ML H2O platform and based on the precision/recall matrix [12] the leader model is selected for final predictions (Fig. 4).

IV. **PU Classification Semi-Supervised Learning**

The best model in this case of semi-supervised learning is the one that is able to identify hidden positives within the training and validation data since our goal is to identify if any of the reported incidents by the employees (concern-reports) may lead to a future injury or illness (Near-Misses, Stop-Work or I&I) event.

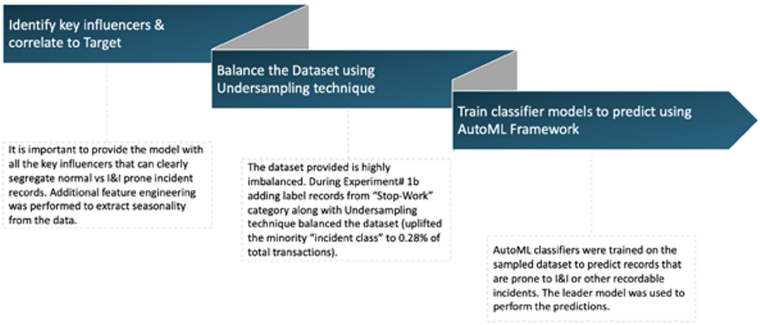

Fig. 4. Under-sampling technique with AutoML Classifier to predict I&I incidents

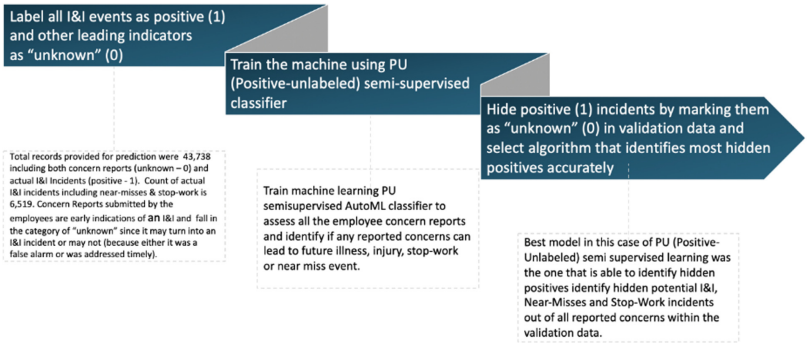

Fig. 5. PU Semi-supervised Classifier to identify any concern reports that may result in future I&I incidents

Positive-unlabeled learning is an important sub-paradigm of semi-supervised learning, where the only available labeled data points are positive. Training dataset used as a combination of 2017–19 I&I (Near-Misses, Stop-Work or I&I) or positive labeled data and 2017–19 Reportable Events (Concern Reports) or unlabeled data. How this algorithm is used on the dataset is shown in Fig. 5.

4 Experiment Setup, Goals, and Results

A list of actual incidents (Illness, Injuries, stop-work, near-misses) and concerns reported are considered for the experimental purpose for the period August 2017 – May 2019 contains around 15000 records of positive incidents. Another list of all timecards submitted by all employees during the same period is also used for the experiment which contains around 9 million records. A single dataset is generated by combining both lists in such a way that around 13000 samples out of 9 million timecards would be labeled as positive (i.e. had an incident) and the rest of the records were unlabeled. September 2018 and February 2019 data are used to validate the performance and accuracy of various models.

The goal of the study is to find:

- Can Machine Learning algorithms correctly predict which of the timecards for Sept 18 and Feb 19 had an I&I incident associated with them?
- How many of the incidents the model can correctly predict?
- How many benign incidents did the model incorrectly label as having an incident

Experiments Performed and Results

As a first step, decision tree models and ensemble methods are used to study the input attributes and their relation to the target variable (output label). Results show that incident records have seasonality. There are specific months and weekdays when the rate is higher. The specific nature of the task and associated customer plays a significant role in the overall distribution of I&I cases.

Experiment A.

The objective of the experiment is to check the presence of an I&I using an amalgamation of the employee's information (hourly employee time card), the employer's information (training hours provided, etc.), the customer's information, and the assignment/project information (location, work conditions, etc.), in which the events labeled as an "incident" are marked "1" otherwise "0". It is observed that the frequency of an I&I event is extremely low (15000 records out of 8.93 million records), which made the importance of these events overlookable to the machine, thereby decreasing the accuracy of the model.

For validation, data from Sept 18 and Feb 19 is used with the below statistics:

- Total records from Sep 18 and Feb 19 to predict I&I cases = 761,641
- Actual Incidents happened from Sep 18 and Feb 19 (excluding "STOP-WORK" events) = 65
- Percentage of positive cases = 0.0085%

The comparative result of all modes is shown in Table 3.

High recall indicates that an algorithm returned the majority of the relevant results, but high precision indicates that an algorithm returned significantly more relevant results than irrelevant ones (Fig. 6). The recall is a measure of comprehensiveness or quantity, whereas precision can be considered as a measure of exactness or quality (Fig. 7). However, the percentage of positive cases in the collection determines the precise relationship between sensitivity and specificity to accuracy which in our case is extremely low (0.0062% in the training dataset and 0.0085% in prediction dataset) (Fig. 8).

Experiment B.

The objective of the experiment is to increase the validity of the model. The inclusion of workplace stoppages in which work on the premises is temporarily halted is added to the dataset as an additional label improving the minority "incidents" class. Using a classification model to discern unusual occurrences along with balancing the data with sampling methods allows the machine to identify the labeled "incidents" as significant, giving these "incidents" more weight in the final accuracy of the model.

For validation, data from Sept 18 and Feb 19 is used with the below statistics:

Table 3. Objective: Given a list of all active projects (including assignment, employee, customer, and nature of work details), train a Machine Learning model that can detect projects that are prone to Illness and Injuries

Model	Total High/Medium Predictions	Actual Incidents from Sep 18 and Feb 19	Correctly Predicted Incidents	Accuracy %	False Alarms	False Alarm Rate
1 – Isolation Forest Untuned	17902	65	14	22%	16602	93%
02 - Isolation Forest w Hyper Param Tuning	6494	65	9	14%	5336	82%
03 - Isolation Forest w Feature Engineering	7819	65	22	34%	6938	89%

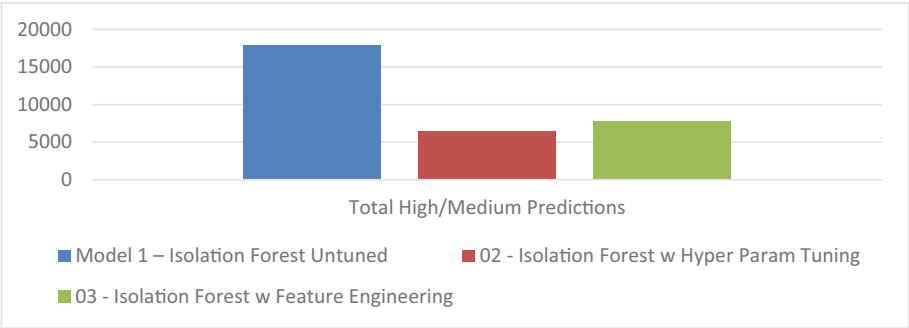

Fig. 6. Prediction of All 3 models

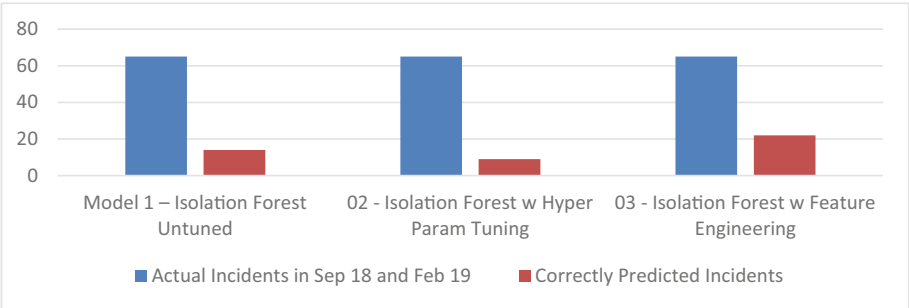

Fig. 7. Actual and predicted incidents of all 3 model

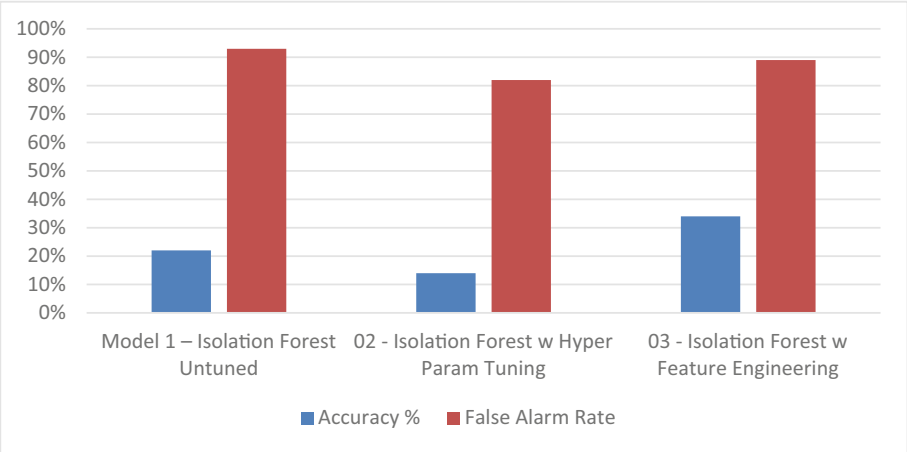

Fig. 8. Accuracy and False Alarm in the percentage of All 3 models

- Total records from Sep 18 and Feb 19 to predict I&I cases = 761,641
- Actual Incidents happened from Sep 18 and Feb 19 (including "STOP-WORK" events) = 2,126
- Percentage of positive cases = 0.28%

The comparative result of all modes is shown in Table 4.

Experiment C.
The primary goal is to take into account the employee's concerns within their "Concern reports". The concern that arises from the employee can result directly in an I&I or any other workplace disruption. These reportable events can be used to predict the possible occurrence of I&Is, which is the goal of experiment C. The hypothesis was that characteristics represented within historic recordable and first aid (I&I) can be utilized to correlate and predict if any of the reportable events (e.g. Near Misses) can turn into future illness or injury. A different dataset with a total of 43,738 records that contains 6,519 labeled near-misses, stop-work, and I&I cases is provided for this experiment C.

- Total records provided for prediction = 43,738
- Actual I&I incidents including near-misses & stop-work = 6,519
- Percentage of actual incidents = 14.9%

The results obtained by using machine learning to assess employee concern reports and identify if any can lead to future illness, injury, stop-work, or near-miss events are shown in Table 5.

Concern Reports submitted by the employees are early indications of an I&I but they do fall in the category of "unknown" where it either turns into an I&I incident or may not (because either it was a false alarm or was addressed timely). The best model in this case of PU (Positive-Unlabeled) semi-supervised learning is the one that is able to identify

Table 4. Objective: Given a list of all active projects (including assignment, employee, customer, and nature of work details), train a Machine Learning model that can detect projects that are prone to Illness and Injuries.

Model	Total High/Medium Predictions	Actual Incidents from Sep 18 and Feb 19	Correctly Predicted Incidents	Accuracy %	False Alarms	False Alarm Rate
Model 1 - Isolation Forest w Feature Engineering	4550	2126	327	15%	2690	59%
Model 2 - Autoencoders Neural Networks	6520	2126	189	8.8%	4430	67%
Model 3 - Classification using Sampled Data	8062	2126	572	34%	535	7%

Table 5. Result of Experiment C

Model	Total High/Medium Predictions	Actual Incidents in the validation set	Correctly Predicted Incidents	Accuracy %	False Alarms	False Alarm Rate
H2O AutoML GBM Classifier (PU Classification Technique)	6290	6519	5215	80%	537	8.5%

hidden positives within the training and validation data. The best model demonstrated the ability to identify hidden potential I&I, Near-Misses, and Stop-Work incidents out of all reported concerns with 80% accuracy (on the validation dataset).

5 Conclusion and Future Work

In the cessation of these findings, the predictive model in experiment B showed promising results, effectively reducing extremely rare occupational Illness and Injury events by 27–34% (recall) in field operations. Using the model built with sampled dataset along with

classification techniques, the predictive model is an early warning system with the ability to predict I&I incidents well in advance with 93% precision.

The Predictive model from experiment C involving employee concern reports, which are otherwise unmanageable due to their volume and unknown outcome, can now be timely assessed and acted upon. The model identifies if any concerns can lead to a future I&I, near-miss, or stop-work incident.

In conclusion, the unpredictability and variance in the category of employee health among field-related operations are effectively reduced and prevented with the usage of these highly precise models, ensuring an increase in field operation safety and security.

Future usage of these models is applicable where physically demanding work is required and can extend to the majority of operations to decrease I&I or work stoppage events.

References

1. Ibrahim, Ima Ilyani, Sarina Muhamad Noor, Noraini Nasirun, and Zulaiha Ahmad. "Favorable working environment in promoting safety at workplace." Journal of ASIAN Behavioural Studies 3, no. 8 (2018): 71–78. https://doi.org/10.21834/jabs.v3i8.279
2. Atan Hood Bin. " Estimation of Occupational Accident and Accident Prevention Cost in Wood Based Industries". PhD diss., University Teknologi Malaysia, 2014.
3. Krishnamurthy, Manikandan, Paramesh Ramalingam, Kumaravel Perumal, Latha Perumal Kamalakannan, Jeremiah Chinnadurai, Rekha Shanmugam, Krishnan Srinivasan, and Vidhya Venugopal. "Occupational heat stress impacts on health and productivity in a steel industry in southern India." Safety and health at work 8, no. 1 (2017): 99–104. https://doi.org/10.1016/j.shaw.2016.08.005
4. Purarjomandlangrudi, Afrooz, Amir Hossein Ghapanchi, and Mohammad Esmalifalak. "A data mining approach for fault diagnosis: An application of anomaly detection algorithm." Measurement 55 (2014): 343–352. https://doi.org/10.1016/j.measurement.2014.05.029
5. Kotsiantis, Sotiris B., I. Zaharakis, and P. Pintelas. "Supervised machine learning: A review of classification techniques." Emerging artificial intelligence applications in computer engineering 160 (2007): 3–24.
6. Fan, Shuoshuo, Guohua Liu, and Zhao Chen. "Anomaly detection methods for bankruptcy prediction." In 2017 4th international conference on systems and informatics (ICSAI), pp. 1456–1460. IEEE, 2017.
7. Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel et al. "Scikit-learn: Machine learning in Python." Journal of machine learning research 12, no. Oct (2011): 2825–2830.
8. https://towardsdatascience.com/credit-card-fraud-detection-using-autoencoders-in-h2o-399cbb7ae4f1
9. Using an AutoML H2O Model to Predict Attrition and LIME to Explain the Predicted Class : by Francis C. Fernandez-Reyes · Jun. 14, 18 · AI Zone · Tutorial : https://dzone.com/articles/using-an-automl-h2o-model-to-predict-attrition-and
10. Suleiman, Dima, and Ghazi Al-Naymat. "SMS spam detection using H2O framework." Procedia computer science 113 (2017): 154-161.
11. H2O.ai:http://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/algo-params/balance_classes.html? highlight=classifier

12. How to Calculate Precision, Recall, and F-Measure for Imbalanced Classification : https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/ . by Jason Brownlee on January 3, 2020 in Imbalanced Classification
13. Bastion Safety Solutions, "Top 10 Reasons — Why workplace safety is Important? " , Jan 22, 2018 https://medium.com/@BastionSafe/top-10-reasons-why-workplace-safety-is-important-8797c978e1f9
14. Borooah, Vani K., John Mangan, and John Hodges. "Determinants of workplace injuries: An econometric analysis based on injuries compensation data for Queensland." Economic Analysis and Policy 28, no. 2 (1998): 149-168.
15. Health and Safety Authority, (November 2012) " Study on the Costs Incurred by Small Businesses as a Result of Workplace Injuries" Indecon International Economic Consultants www.indecon.ie
16. Lu, C.-S., & Yang, C.-S. (2011). Safety climate and safety behavior in the passenger ferry context. Accident Analysis and Prevention, 43(1), 329–341. https://doi.org/10.1016/j.aap.2010.09.001
17. Ismail, Zubaidah, Samad Doostdar, and Zakaria Harun. "Factors influencing the implementation of a safety management system for construction sites." Safety science 50, no. 3 (2012): 418-423. https://doi.org/10.1016/j.ssci.2011.10.001