Deep Learning Models for Health and Safety Risk Prediction in Power Infrastructure Projects

Anuoluwapo Ajayi[1], Lukumon Oyedele[1*] Hakeem Owolabi[1], Olugbenga Akinade[1], Muhammad Bilai[1], Juan Manuel Davila Delgado[1] and Lukman Akanbi[1]

1 Big Data, Enterprise and Artificial Intelligence Laboratory, University of the West of England, Frenchay Campus, Coldhabour Lane, Bristol, United Kingdom.

* Address correspondence to Lukumon Oyedele, University of the West of England, Frenchay Campus, Coldhabour Lane, Bristol BS16 1QY, UK, tel: +44 (0) 117 32 83443 L.Oyedele@uwe.ac.uk

Inappropriate management of Health and safety (H&S) risk in power infrastructure projects can result in occupational accidents and equipment damage. Accidents at work have detrimental effects on workers, company, and the general public. Despite the availability of H&S incident data, utilising them to mitigate accident occurrence effectively is challenging due to inherent limitations of existing data logging methods. In this study, we used a text mining approach for retrieving meaningful terms from data and develop six deep learning (DL) models for H&S risks management in power infrastructure. The DL models include $DNN_{classify}$ (risk or no risk), DNN_{reg1} (loss time), DNN_{reg2} (body injury), DNN_{reg3} (plant & fleet), DNN_{reg4} (equipment), and DNN_{reg5} (environment). An H&S risk database obtained from a leading UK power infrastructure construction company was used in developing the models using the H2O framework of the R language. Performances of DL models were assessed and benchmarked with existing models using test data and appropriate performance metrics. The overall accuracy of the classification model was 0.93. The average R-squared value for the five regression models was 0.92, with Mean Absolute Error (MAE) between 0.91 and 0.94. The presented results, in addition to the developed user-interface module, will help practitioners obtain a better understanding of H&S challenges, minimise project costs (such as third-party insurance and equipment repairs), and offer effective strategies to mitigate H&S risk.

KEYWORDS: Artificial intelligence; deep learning; health and safety risk

1. INTRODUCTION

Health and Safety (H&S) risks are of particular concern in the power infrastructure sector (Aksorn & Hadikusumo, 2008). This is because the power infrastructure sector has high rates of fatality and nonfatal occupational injury, as well as damage to plants and equipment, which require significant repair costs (McDermott & Hayes, 2016). The UK Health and Safety Executive estimated the cost of nonfatal workplace injuries in the UK at £4.8 billion (HSE, 2016). There is also a high risk of illnesses due to exposure to radiation, dust, chemicals and extreme temperatures. These risks are specific to power infrastructure projects because they involve tasks that require working in excavations, trenches, cell tower base stations, scaffolds and operating heavy machinery.

Several Machine Learning (ML) techniques have been adopted to predict H&S risks. Some examples are multiple regression (Tsoukalas & Fragiadakis, 2016), fuzzy methods (Rastogi & Gabbar, 2013), decision trees (Cheng, Leu, Cheng, Wu, & Lin, 2012) and generalised linear models (Esmaeili, Hallowell, & Rajagopalan, 2015), along with fuzzy neural networks (Debnath, Biswas, Sivan, Sen, & Sahu, 2016), fuzzy Bayesian networks (Zhang, Wu, Qin, Skibniewski, & Liu, 2016), bow-tie representation (Jacinto & Silva, 2010), and Bayesian networks (Englehardt, An, Fleming, & Bean, 2003; Papazoglou, Aneziris, Bellamy, Ale, & Oh, 2015). A key challenge associated with most conventional ML and statistical techniques is the considerable effort needed to manually extract attributes from a large pool in the database for them to achieve high prediction accuracy (Esmaeili et al., 2015). As such, an emerging trend within ML has led to the development of Deep Learning (DL) approaches to address the problem of manual extraction of relevant attributes from raw data compiled and merged from various sources. Also, there exist non-linear relationships with high-order interactions among independent and dependent variables making up the raw data.

DL or Deep Neural Network (DNN) is an ML technique that uses multiple processing layers and abstraction levels to learn a data representation. DL with its high learning capacity is efficient at discovering complex structures and can dynamically construct new task-specific attributes from high-dimensional data (LeCun, Bengio, & Hinton, 2015). According to LeCun et al. (2015), these characteristics have enabled it to surpass the performance of existing ML approaches. As such, DL has successfully been used in fields such as vision and image processing (Gharbi, Chen, Barron, Hasinoff, & Durand, 2017), speech recognition (Hinton et al., 2012), and medicine (Al-Rahhal et al., 2016). Other application areas are drug molecule analyses (Ma, Sheridan, Liaw, Dahl, & Svetnik, 2015), building cooling load prediction (Fan, Xiao, & Zhao, 2017), and traffic control (Zhao, Chen, Wu, Chen, & Liu, 2017).

H&S events data in the power infrastructure domain are typically large, heterogeneous and dynamic (Fenrick, Getachew, Fenrick, & Getachew, 2012). Evidence suggests that imbalanced and appreciable missing values are prevalent (Bohle, Quinlan, McNamara, Pitts, & Willaby, 2015). Despite the availability of the H&S events data, utilising them to mitigate accident occurrence effectively is challenging because of the data logging method employed which makes analysis and identification of relevant patterns difficult for humans. Therefore, a text mining approach with the good generalisation ability of DNN will provide a robust mechanism for handling sparse, noisy and nonlinear high-dimensional data (Mamoshina, Vieira, Putin, & Zhavoronkov, 2016). This study, therefore, implements a robust text mining approach and deep neural network method for the PT&D domain and prompts readers to explore its usefulness, especially in predicting H&S risks in power infrastructure construction sites for better decision making. This study, in contrast to other studies (Cheng et al., 2012; Debnath et al., 2016; Esmaeili et al., 2015; Mistikoglu et al., 2015; Tsoukalas & Fragiadakis, 2016), presents a generalised approach to predicting and managing H&S risks to humans, equipment and the environment. The methodology of risk assessment adopted in this study, though using the PT&D infrastructure as a case study, is based on a generic modelling framework, which can easily be customised and extended to other domains when relevant domain knowledge is incorporated. The objectives of the study are to implement a text mining technique and develop robust DNN models (a classification model - $DNN_{classify}$, and five regression models-DNN_{reg1}, DNN_{reg2}, DNN_{reg3}, DNN_{reg4}, DNN_{reg5}) to manage H&S risks and evaluate their accuracies using appropriate metrics.

This work is structured as follows. In Section 2, the materials and methods employed, such as H&S events data, data pre-processing, deep learning and neural networks are discussed. We discuss DNN models' development, benchmarking and results in Section 3, and describe an interface for user interaction with deep learning modules in Section 4. We give concluding remarks and discuss areas of future research in Section 5.

2. RESEARCH OVERVIEW AND DATA DESCRIPTION

In the current study, we use DL as the key technology to predict H&S risks specific to infrastructure projects, since project-related attributes that influence accident occurrence were utilized in modelling the models. We describe DL as a tool, as well as the data pre-processing approach, in subsequent subsections. In Fig. 1, we present the general research outline for the study. Data pre-processing and selection of variables are first carried out to clean H&S event data and identify relevant models' inputs. Subsequently, we construct six DNN models (one classification and five regression models) to set relationships between dependent and independent variables to manage H&S risks. Relevant optimisation techniques are then used to maximise each model's prediction accuracy. Lastly, we evaluate the prediction accuracies of the models on test data using acceptable metrics such as Area Under Curve (AUC), Mean Absolute Error (MAE), Accuracy, Kappa statistic, Sensitivity and R-square (R^2).

2.1. H&S Events Data and Pre-processing

H&S events data covering 17972 reported cases were obtained from a leading UK Power Infrastructure and Distribution company. In Fig. 2, we present percentages of H&S events (incidents and accidents) and damages to plant and fleet for infrastructure projects executed during the period of study. The proportion of fatal accidents, which are usually rarer are not shown in the figure, as only three projects recorded fatal accidents in three years (i.e. 2008, 2009, and 2010).

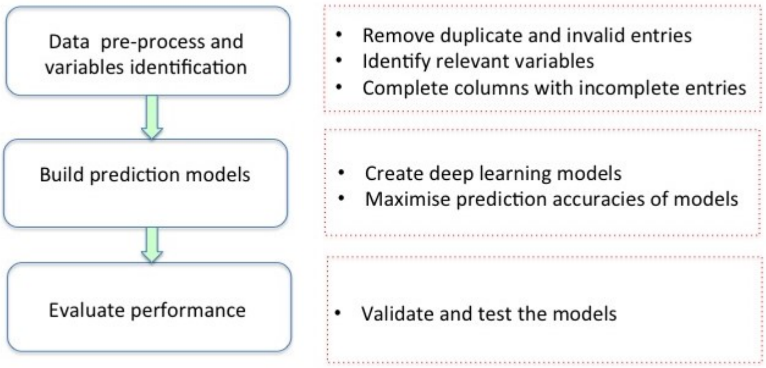


Fig. 1. Research outline

Fig. 2a shows non-fatal accidents (first aid, medical attention beyond first aid, injuries requiring a few days away from work, etc.) and incidents (no reported injury). Accident types "first aid" and "medical beyond first aid" dominate the accident category. The class distribution is, however, not uniform among classes as there are more incidents (majority) compared to non-fatal accidents (minority) as shown in Fig. 2a. Fig. 2b depicts damages to plants and fleets, with damage to "bonnet/bumper" dominating this event. Although, proportions of non-fatal accidents and equipment damage are dropping, determining the primary causes of these incidents is necessary to ensure both human and property safety.

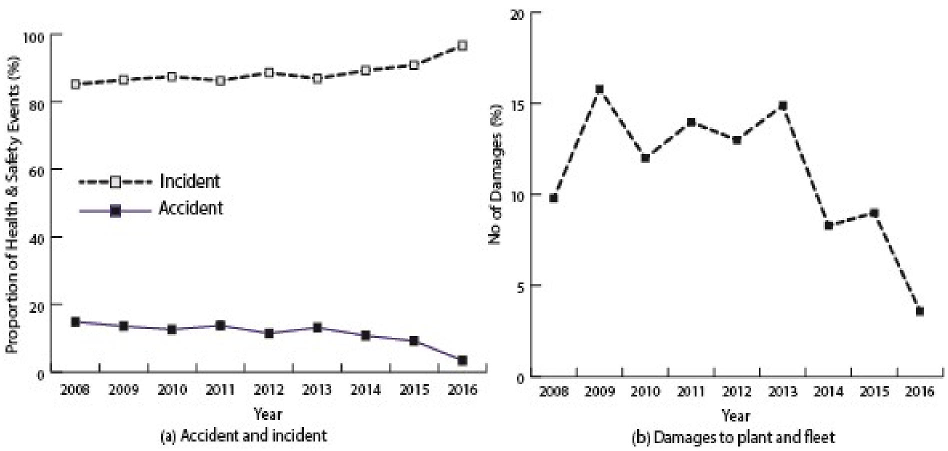


Fig. 2. H&S events proportion

Data cleaning and quality control procedures are implemented to remove invalid and duplicate entries, the remaining data after data cleaning consisted of 16900 events that occurred between January 2008 and June 2016 (90 months). We apply a text-mining approach (Weiss, 2010) to retrieve useful underlying concepts from textual columns (i.e. "*incident_description"*, "*address"*) in the database to create new columns (i.e. "*Shift patterns*", "*climatic condition*", etc.), and used extracted information to complete certain columns with missing entries. To retrieve vital information from text fields, we compute the frequency of each useful word after performing term creation and filtering operations. In the term creation procedure, we tokenized all terms in the string, while in the filtering operation, we carried out the following pre-processing tasks: punctuation and number removal, lower-case conversion, stemming, and stop-word and common-word removal to create the document term matrix (DTM). The DTM allows us to identify meaningful terms along with their corresponding frequencies for further analysis and investigate events using word combinations. For example, filtering co-occurring expressions such as 'slip' and 'ice' or 'icy' from the "*event_description*" column can help infer automatically the event "Slipping on an icy surface."

The complete variables (independent and dependent) used after the text-mining approach are described in Table I. The important variables are selected using a default deep learning model (trained without tuning its parameters), and setting the "*variable_importances*" feature to true, which enables the viewing of the absolute and relative predictive strength of each variable in the prediction task. Dependent variables in Table 1 are italicised. The χ^2coefficient of contingency was calculated to clarify the correlations between the categorical variables. A large degree of freedom was found between variables, and χ^2 was normally distributed. Tests of independence also revealed that $\chi^2 > \chi^2_{(r-1)(c-1),\alpha}$, indicating variables are from the same population and are correlated with each other.

Table I. Predictive and dependent variables for DNN models

Model	Variable name	Explanation
$DNN_{classify}$	Project type	Categorical (overhead line, cabling, or substation)
	Project complexity	Categorical (new build, maintenance, or refurbishment.
	Region	UK regions where projects are constructed i.e. North, Midlands, South East, Scotland, etc.
	Location	Categorical (rural or urban)
	Client	This includes energy company, communications, digital, power supplier contractors, etc.
	Duration	Project duration. Categorical (short, medium, or long)
	Season	Categorical (Autumn, Spring, Summer, or Winter)
	Cost	Categorical (Small: <£250K, Medium: £250K-£1m, Large: £1m-1£0m, Extra large: >£10m)
	Contract status	Categorical: main contractor, or subcontracted, or third party.
	Contract type	Categorical (Target cost, Lump sum, Bill of quantity, Cost reimbursable)

	Accident occurrence	Categorical (Yes or No) (Dependent variable)
DNN_{reg1}	Activity	Categorical up to 25 levels (i.e. Pulling, driving, stringing, loading, foundation, excavating, tower erection, working at height, etc.)
	Equipment	Category up to 20 levels. The tool used for a task (plant equipment, hand tool, etc.).
	Equipment state	Equipment state (Good, moderately good, bad).
	PPE kit type	Condition of the PPE kit. Categorical (Good condition, fairly good, not in good condition)
	Working surface layout	Good condition, moderately in good condition, not in good condition
	Climatic condition	Categorical: Sunny, raining, snowy, windy, clear
	Shift pattern	Categorical: Day, night or weekend
	Distance to site	How long it takes an employee to travel to the site? (Short, average, long)
	Experience	Length of time on the job (<1year, 1-3years, >3 years)
	Day of the week	Day name i.e. Monday, Tuesday, Wednesday, etc.
	Employment contract	Categorical (Temporary or permanent).
	Time	When a task is performed (6am-12pm) or (12pm-19pm).
	Lost time	Number of days (x) an employee is absent from work due to injury, i.e. $0 \leq x \leq 15$. (Dependent variable)
DNN_{reg2}	Activity/Task	Categorical up to 25 levels (i.e. Pulling, driving, stringing, foundation, excavating, tower erection, etc.)
	Equipment	Categorical up to 20 levels. The tool used for a task (plant equipment, hand tool, etc.).
	Equipment state	Equipment state (Good, moderately good, bad).
	PPE kit type	Categorical (Good condition, fairly good, not in good condition)
	Working surface layout	Working surface layout in terms of being spacious and conducive. Categorical (Good, Moderately good, or not in good condition).
	Weather condition	Categorical: Sunny, raining, snowy, windy, hot, cloudy, foggy, etc.,
	Shift patterns	Categorical: Day, night or weekend
	Distance travel to site	Categorical (Short, average, long)
	Experience	Length of time on the job (<1year, 1-3years, >3 years)
	Day of the week	Day name i.e. Monday, Tuesday, Wednesday, etc.
	Employment contract	Categorical (Temporary or permanent).
	Time	When a task is performed (6am-12pm) or (12pm-19pm).
	Body part	A numerical value between 1 and 16 representing body parts i.e. head, back, leg, ankle, face, chest, etc. Sixteen represents no body parts injury. (Dependent variable)
DNN_{reg3}	Location	Categorical: Rural, city, vandals dominated areas.
	Shift patterns	Categorical: Day, night or weekend
	Operator experience	Categorical: Expert, moderately experienced, inexperienced
	Service requirement	Categorical: Regularly (once in 6 months), infrequent (once in two year
	Activity/Task	Categorical: Transporting, uploading, towing, excavating, pulling, etc.
	Plant contract	Categorical: Owned or hired plant.
	Mobilisation technique	How the plant is transported to site i.e. by helicopter, by truck, or driven?
	Fleet/part damaged	Categorical 20 levels (Bumper, Window, Alloy/Wheel, Bucket, etc.) (Dependent variable)
DNN_{reg5}	Time	Categorical. When a task is performed (early in the day) or (later in the day).
	Employment contract	Categorical (Temporary or permanent).
	Working surface layout	Good condition, moderately in good condition, not in good condition.
	Activity/Task	Driving, excavating, Pulling, tower erection, etc.
	Equipment	The tool used for a task (plant equipment, hand tool, etc.).
	Location	Rural or Urban
	Information available on utility location	Categorical (Available or unavailable)

Environment/Utility damaged	These include Farm/Field, Fence, or service tracks damaged, animals killed, etc. (Dependent variable)

Furthermore, to avoid differences in model performance due to imbalanced and unscaled training data, we carry out the following steps. We solve the imbalance problem by re-weighting input data and adding different penalty coefficients to the training errors corresponding to the input. Specifically, by assigning larger penalty coefficient values to examples from minority classes and smaller penalty coefficient values to points from majority classes, to reduce overfitting. In performing scaling, we divide mean-centred variables' values by their standard deviations. Then, the remaining missing data are substituted using the mean imputation method.

2.2. Deep Learning and Neural Networks

DNNs allow computational models, which are comprised of several processing layers, to learn data representations with multiple levels of abstraction (LeCun et al., 2015). DNNs use multiple hidden layers between input and output layers to model complex nonlinear relationships. A DNN is formally defined as follows. Denote the output of a neuron at layer ℓ by h^ℓ, and its input vector from the previous layer by $h^{\ell-1}$. Then, the activation of neurons is defined as $h^\ell = \sigma(b^\ell + W^\ell h^{\ell-1})$, where b^ℓ is a vector of biases, W^ℓ is the matrix of weights, and σ(·) is an element-wise activation function. Activation functions include *Tanh*, *Rectifier*, *Sigmoidal* and *RectifierWithDropout*. The input vector $x=h^o$ is the raw data to be analysed by the network, and the output vector h^ℓ (in the output layer) is used to make predictions. For a multiclass classification task, the model output is determined using Equation (1), where W^ℓ is the matrix of weights, W_i^ℓ is the i[th] row of W^ℓ, $h_i^\ell > 0$, and $\sum_i h_i^\ell = 1$.

$$h_i^\ell = \frac{exp\left(b_i^\ell + W_i^\ell h^{\ell-1}\right)}{\sum_j exp\left(b_j^\ell + W_j^\ell h^{\ell-1}\right)} \qquad (1)$$

For a regression task, the output is determined using Equation (2), where α_{0k} represents the bias applied to the output layer and α_k is the set of weights between the previous layers and the last layer.

$$h_i^\ell = \alpha_{ok} + \alpha_k \sigma\left(b_i^\ell + W_i^\ell h^{\ell-1}\right) \qquad (2)$$

The outputs and the target function y are used together in a cost function $\mathcal{E}(h^\ell, \mathcal{Y})$, which is convex in $b^\ell + W^\ell h^{\ell-1}$. The cost functions for classification and regression tasks are defined using Equation (3), where h_y^ℓ is the network output and y is the desired response.

$$\begin{aligned} \mathcal{E}(h^\ell, y) &= -logh_y^\ell, \quad classification \\ \mathcal{E}(h^\ell, y) &= \left\|y - h_y^\ell\right\|^2, \quad regression \end{aligned} \qquad (3)$$

Techniques that have significantly boosted the success rate of DNN are briefly described as follows:

- Rectified Linear Unit (ReLU) - The most successful techniques in DNNs (LeCun et al., 2015), and a special case of Maxout, which enforces sparse representations and prevents vanishing gradients (Hochreiter, Bengio, Frasconi, & Schmidhuber, 2001). RELU has helped to obtain best results on several benchmark problems across multiple domains.(Dahl, Sainath, & Hinton, 2013)
- Dropout technique – A DNN regularisation scheme for preventing overfitting (Hinton et al., 2012; Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). This technique avoids co-adaptation of hidden units by randomly dropping units during training.
- Cross-entropy objective with Softmax activation - This technique involves applying error functions such as cross-entropy and logistic loss as objectives to be minimised (Mayr, Günter, Unterthiner, & Sepp, 2016). The error functions are often associated with the Softmax activation in the output neurons.

In contrast to Artificial Neural Networks (ANNs), DNNs have complex layers with more training schemes. Table II presents a comparison of ANN and DNN.

Table II. Comparison between DNN and ANN

Feature	DNN	ANN
Architecture	It is complex and deep, with greater than five layers.	Simple and shallow (usually three layers).
Feature extraction	It learns features automatically.	It learns from handcrafted features.
Non-linear function	It uses powerful and optimised functions such as Rectified Linear Unit (ReLU) and Maxout to ensure robustness.	It uses functions such as Tanh and Sigmoidal, amongst others.
Distributed representation	Has exponential advantage due to its distributed representation	Not applicable.
Learning algorithm	Stochastic gradient descent (SGD) for parallel processing.	Mostly uses standard SGD.
Neurons	Has thousands of neurons in each layer to capture all possible facets of input.	Employs a small number of neurons to learn the main pieces of information.

Several DNN architectures abound, with each recording successes in different domains. Examples are feedforward neural networks, recurrent neural networks and the convolutional neural networks. Feedforward neural networks allow data to flow from input to output without looping back. Recurrent neural networks (which are used for language modelling) allow data to flow in any direction. Convolutional DNNs are standard for computer vision applications. We employ the feedforward architecture because of its simplicity and popularity (LeCun et al., 2015).

2.3. Deep Learning for H&S Risk Prediction

A DL network consists of several layers with numerous neurons. It constructs features in neurons and input data are represented as features. Higher layers are used to represent numerous abstract concepts compactly (LeCun et al., 2015). Analysis and modelling of incident databases using practical ML techniques can help predict H&S risks to assist stakeholders in developing programmes and safety practices for H&S risk prevention (Farid, Al-Mamun, Manderick, & Nowe, 2016). Different analytics techniques that are suitable for DL implementation are shown in Table III. Since the domain of this study is predictive analytics, we use two predictive techniques (classification and regression) for H&S risk prediction.

Table III. Machine Learning Techniques

Technique	Application	Algorithm
Regression	Predicts continuous numerical outcomes such as the number of instances of vehicle damage within a timeframe, the number of back injuries suffered by workers and the number of slip/trip incidents in winter.	Linear regression, Naïve Bayes, decision trees, neural networks, support vector machine (SVM), DNN, amongst others
Classification	For delineating classes of output (usually categorical) based on some set of input features. The basic form is a binary classifier with a single output with two labels (Yes and No).	Bayesian probability, ANN, SVM, random forest, DNN, gradient boosted machines (GBM)
Clustering	Explores data to find natural groupings. An example is finding related events that result in a given outcome; for instance, walking on wet ground may cause a trip/slip incident.	K-means, clustering, SVM, Expectation-maximisation, self-organising maps, autoencoders, DNN
Attribute importance	It ranks attributes according to the strengths of their relationships to the target attribute, for instance, by finding the factors that are most associated with worker injury while working on the site.	Minimum description length, decision trees, random forest, DNN
Anomaly detection	Identify unusual or suspicious cases based on deviation from the norm; for example, identifying possible fall accidents based on workers' motion data (e.g., velocity and orientation)	Expectation-maximisation, SVM, DNN, self-organising maps, fuzzy-C means
Association	Finds rules associated with frequently co-occurring items (root-cause analysis), i.e., lower-back injuries among construction workers as a function of lifting heavy objects.	*A priori*, GBM, particle swarm optimisation, DNN, ANN
Feature selection & extraction	Generates new attributes as a linear combination of existing attributes. It is suitable for latent semantic analysis, data compression and pattern recognition.	Principal component analysis, genetic algorithm, Singular Vector Decomposition (SVD), DNN

3. DNN MODELS DEVELOPMENT AND RESULTS

A two two-stage predictive architecture which includes classification and regression is proposed in this study. The architecture first predicts if there is a risk or not; then if there is a risk, regression is performed to quantify the risk, and appropriate actions are then recommended for its mitigation. Thus, the two-stage predictive architecture depicted in Fig. 3, consists of a classification model, which determines the likelihood of an accident and five regression models that predict relevant information such as possible injured body parts, damage to equipment, and environment (trees, water pipes, gas pipes, etc.). The Interpretable Machine Learning (IML) is an R package (R Development Core Team & R Core Team, 2016) that offers a general toolbox for making ML models interpretable. Some examples of interpretability methods can be produced using IML include partial dependence plots, individual conditional expectation (Goldstein, Kapelner, Bleich, & Pitkin, 2015), feature importance, global surrogate tree (Molnar, Casalicchio, & Bischl, 2018), etc. As shown in Fig. 3, feature selection as a specific approach for model interpretation is indicated with the IML selecting a subset of features $x_1, x_2, \ldots x_p, where\ p < n$, that are useful in building a good predictor for each response variable.

3.1. Description and Building of Models

After data pre-processing, we split the dataset (16900) into 60% (training set), 20% (validation set) and 20% (test set) to implement the models. Specifically, 10140 observations to build the models, and the remainder (6760) are employed for validation (3380) and testing (3380). We built six DNN architectures comprising a classification model ($DNN_{classify}$) and five regression models (DNN_{reg1}, DNN_{reg2}, DNN_{reg3}, DNN_{reg4}, and DNN_{reg5}) using the training data in H2O framework using different control parameters. H2O is an open source CRAN package, high-speed, and Java machine learning library software, designed with distributed algorithms scale to big data (Kochura, Stirenko, Alienin, Novotarskiy, & Gordienko, 2017). H2O has an interface to Python, Scala, R, Spark, and Hadoop.

According to Al-Rahhal et al. (2016), finding an optimal structure of deep neural networks is a means of examining its sensitivity. In order to obtain the best validation results for the DNN models, there is need to find an optimal structure of the neural network (number of hidden layers, number of activation units in each layer, and activation functions) and control hyper-parameters, we used a random search. A random search approach is many times more efficient than the grid search method (Tixier, Hallowell, Rajagopalan, & Bowman, 2016). Also, we objectively evaluate all DNN models with respect to architecture since the network architecture plays a major role in improving the classification or prediction accuracy (Lee, Grosse, Ranganath, & Ng, 2009). Control parameters (i.e. Table IV for DNN_{reg1}) were tuned to maximise each model's prediction accuracy on unseen data. For instance, for each DNN model, we apply different hyper-parameter combinations using the random search, and optimal control settings are determined by a 5-fold cross-validation with 10% holdout. The validation step helps to prevent overfitting by comparing the performances of the prediction algorithms that were created based on the training set and selecting the algorithm with the best performance metric. In this case, an algorithm with the least Mean Absolute Error (MAE) is chosen as suggested (Tixier et al., 2016). The accuracies of different DNN_{req1} network structures for instance, with respect to MAE, is depicted in Fig. 4. We settled for the optimal structure with two layers (180 neurons in each layer), Rectifier activation function, $\ell 1 = 1e-4$, $\ell 2 = 1e-6$, and epoch =30. This structure has the least MAE (0.7964) value. Besides, we found out that for this specific regression problem the higher number of neurons were not making a significant difference in the accuracy of the model and therefore we chose fewer neurons to reduce network's complexity. We annotate this optimal topology by appending '2' to the appropriate plot positions. The architectures of other DNN models are determined accordingly, and appropriate optimal hyper-parameter values were obtained.

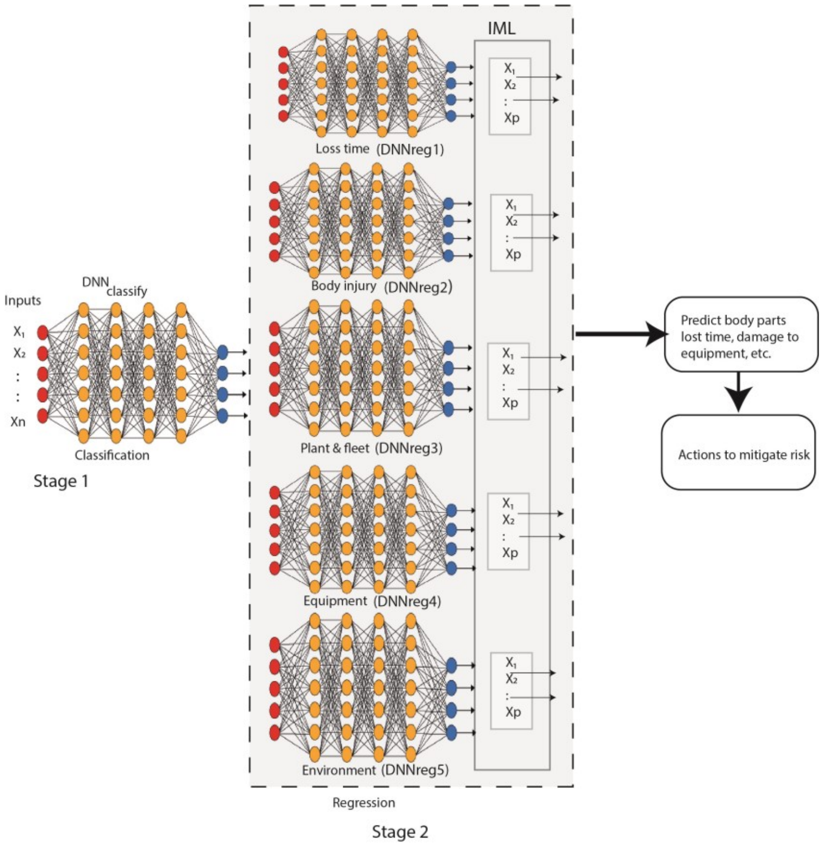


Fig. 3. Conceptual framework of DNN H&S risk prediction

Table IV. Hyper-parameter Combinations

Parameter	List
Activation function	Rectifier, Maxout, RectifierWithDropout, Tanh, TanhWithDropout, etc.
layers	1, 2, 3, 4
neurons	40, 100, 180, 270, 500
rho	0.9, 0.999
epoch	10, 30, 50, 100
epsilon	1e-10, 1e-4
$\ell 1$ regularisation	0, 1e-4, 1e-8, 1e-7
$\ell 2$ regularisation	0,1e-4, 1e-6, 1e-7
input_dropout_ratio	0, 0.05

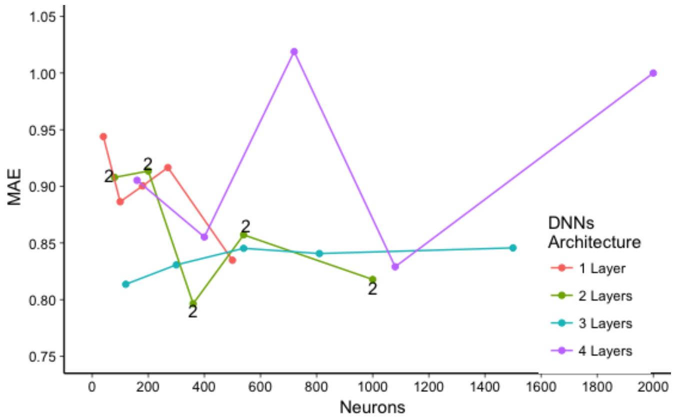


Fig. 4. Accuracy of DNN_{reg1} configurations with MAE as the metric

Although, several optimisation algorithms such as Least-squares methods (Gauss-Newton, Levenberg–Marquardt), quasi-Newton methods (i.e. Broyden–Fletcher–Goldfarb–Shanno (BFGS)) amongst others abound. These methods are computationally too expensive for large NNs (Schmidhuber, 2015). Conjugate gradient (CG), Limited-memory-BFGS (L-BFGS) and other methods are fast alternatives to these algorithms. However, the CG algorithm, in general, requires more cycles to reach the minimum while L-BFGS can overfit on a small training set (Bengio, 2012).

SGD algorithm is a fast training procedure for reducing the cost or loss function (computing a gradient over all training samples), compared with other optimisation techniques (Bottou, 2010). The algorithm computes outputs, errors and the average gradient of observations, and adjusts the weights where necessary. Using a parallelisation technique with a suitable DNN architecture, advantages of the SGD algorithm over L-BFGS increase as the training set size increases (Bengio, 2012). We applied the parallelised SGD (Recht, Re, Wright, & Niu, 2011), which models a lock-free shared-memory system where each processor independently performs stochastic gradient updates. The lock-free stochastic gradient keeps a global result vector and allows each processor to update the vector without considering other processors. Under certain conditions, this asynchronous procedure preserves the convergence of stochastic gradient methods and results in ample speed-ups for many available cores. All models tuning, training and prediction were performed using the H2O framework in R. We adopted H2O because it is a fast and scalable open-source framework for machine learning applications.

3.2. Effect on Small and Bias Data

Naturally, DNN is synonymous with big-data applications (LeCun et al., 2015; Mayr et al., 2016) and it results in overfitting when used on a small dataset. We incorporated a dropout technique, combined with SGD in the training procedure, to reduce overfitting. Data veracity refers to the presence of biases, noise and abnormalities in data (Daniel, 2017). Data imbalance is a form of bias in ML, where the class distribution is not uniform among various classes. For instance, missing data also represent an aberration in data, which could significantly affect prediction. The DNN structures used in this study (insensitive to imbalanced data) are similar to the classic DNN (Larochelle, Bengio, Louradour, & Lamblin, 2009). The H2O framework automatically performs mean imputation for missing values during training.

3.3. Performance Metrics for Model Verification

We measure the predictive accuracies of all models on test cases to assess their generalisation abilities as suggested by Lecun et al. (2015). We use the following performance metrics: accuracy, sensitivity, specificity, AUC, Kappa coefficient, and R-squared. Sensitivity, specificity, and AUC are commonly used to compare a model's predictions against the ground truth.(Gibson & Patterson, 2017; Vallmuur, 2015) There are standard ways of interpreting each metric. For instance, the Kappa coefficient has values less than or equal to one. We categorised these values as defined in (Landis & Koch, 1977) as follows: a value that is less than 0 indicates no agreement, a value between 0 and 0.21 indicates slight agreement, between 0.21 and 0.41 is fair, from 0.41 to 0.61 is moderate, from 0.61 to 0.80 is substantial, and a value between 0.81 and 1 represents perfect agreement. Similarly, for the regression models, we objectively evaluated each model's performance with respect to the different architectures using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). MAE and RMSE are scale-dependent metrics that provide reliable ways to quantify prediction error (Fan et al., 2017). The target is to minimise these metrics to obtain the highest prediction accuracy for the model. These metrics are defined as Equations (4)-(5), where t_i denotes target i, y_i denotes prediction i and N is the number of testing observations.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|t_i - y_i| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(t_i - y_i)^2} \quad (5)$$

We determine the models' complexities by computing the time spent training each model. The computational tool on which the simulation was carried out was a MacBook Pro (Intel Core i7 processor of 2.5 GHz and random access memory of 16 GB) and it was modelled using the R language.

3.4. Model Benchmarking

To proof the worth of deep neural networks in terms of accuracy, we compared the performance of the deep learning classification model ($DNN_{classify}$) and one of the developed DNN regression models (DNN_{reg2}), being the least in respect to prediction accuracy with a null model (neural network with one hidden layer) and a gradient boosted model (GBM). GBM is a generalisation of tree boosting that attempts to produce an accurate and effective procedure for data mining. GBM is a powerful and popular technique with proven ability for problems relating to predictions in various fields (Liu, Zhang, & Shen, 2012; Liu, Li, Tan, Zhu, & Wang, 2014). It is also worth mentioning that all the three models were trained on the same dataset. We omitted part of the dataset from the training set and used it for testing the prediction performance of the null model, DNN_{reg2} and GBM models. In the first part, a regression problem where body part injured is tackled, and in the second part a classification problem to predict the possibility of a risk occurring. For the first problem, the tuning parameters for the optimal GBM model, in this case, are the number of trees (200), interaction depth (8), and learning rate (0.1). We illustrate a regression problem by predicting injury to body parts to compare the performance of the models. Table V summarises the Mean Square Error (MSE), Mean Absolute Errors (MAE), Root Mean Square Error (RMSE), and R squared for the null model, DNN_{reg2}, and GBM models. The prediction errors are computed by comparing values predicted by the model with the response variable (body part injured) in the test set. Fig. 5 shows error frequency distribution for the three models (Null, DNN, and GBM) respectively. It can be seen that the error distribution for the DNN and GBM models are highly concentrated near zero, while that of the Null model is relatively dispersed. However, the DNN's prediction accuracy for predicting hazards at construction sites is superior to GBM and null models.

Table V. Prediction Errors

Model	MSE	MAE	RMSE	R^2
Null	5.547	1.773	2.355	0.785
DNN	1.487	0.744	1.219	0.928
GBM	2.021	0.932	1.422	0.913

Table VI summarises the three ML techniques classification performance using confusion matrix. The null model shows that 1026 false values (i.e., no risk) correctly classified as false, while 77 are wrongly categorised as true. Similarly, 874 true values are categorised as true, while 123 are wrongly misclassified. The $DNN_{classify}$ correctly classified 1107 false values (i.e., no risk), and 55 wrongly labelled as true. It also correctly labelled 896 true values but wrongly misclassified 42 as false (no risk). Likewise, GBM classified correctly 1099 false values, 59 false values as true, correctly labelled 892 true values, and wrongly misclassified 40 true values as false.

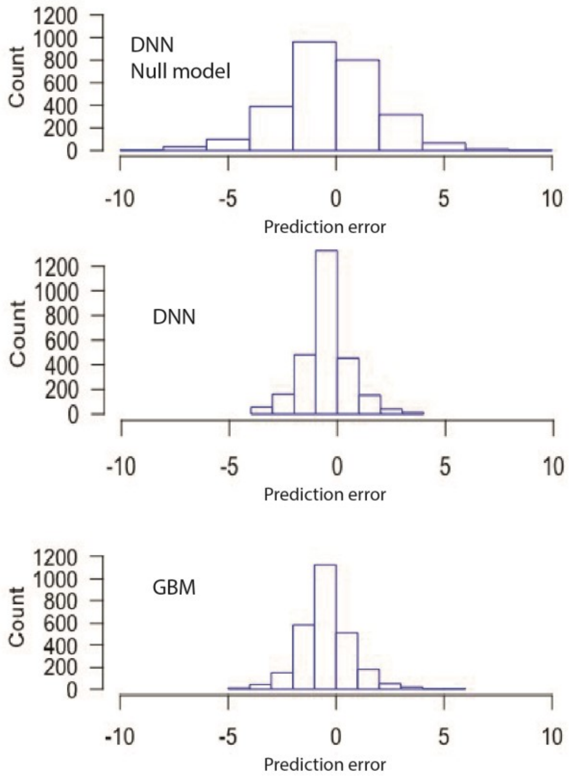


Fig. 5. Prediction error histograms (Null, DNN, GBM)

Table VI. Classification Confusion Matrix

Model	Actual/Predicted	No risk	Risk	Error
Null	False	1026	77	0.075
	True	123	874	0.141
	Totals	1149	951	0.095
DNN	False	1107	55	0.049
	True	42	896	0.047
	Totals	1149	951	0.046
GBM	False	1099	59	0.054
	True	50	892	0.056
	Totals	1149	951	0.052

3.5. Models Interpretability

Interpretability is an essential condition for machine learning models applied in areas such as medicine, financial markets, and law (Chen, Song, Wainwright, & Jordan, 2018). The IML implements many model-agnostic methods (i.e. surrogate tree, Shapley value, Feature extraction, partial dependence, etc.) to interpret a DNN model. However, due to space constraints interpretation of models with few model-agnostic methods are discussed in the study. Two perspectives are used to interpret the developed DNN models. The first perspective is global, which includes feature importance and measuring interactions between predictors. The second perspective is local, which tries to explain why an individual prediction was made for a given observation.

The feature importance measure was computed by calculating the increase in the model's prediction error after permuting a feature. A feature is "important" if the model error increases by permuting its values, because the model relied on the feature for its prediction, while a feature is "unimportant" if permuting its values keeps the model error unchanged. The outputs of some of these are shown in Fig. 6. For example, the ranking of predictors for $DNN_{classify}$, DNN_{reg2} and DNN_{reg3} in order of their importance are depicted in Figs 6a, 6b, and 6c. Feature "experience" is the most important variable for DNN_{reg2} with an error increase of 5.82 after permutation. Similarly, predictor "activity" is the most influential for DNN_{reg3} with an error increase of 4.87 after permutation. The feature selection as determined by IML for all the models has the following as the most significant predictors for PT&D project risks management are project type, complexity, equipment, activity, duration, climatic condition, location and working surface. Many of these predictors have been used in previous studies to predict health and safety risks. For instance, project type, location, activity (Törner & Pousette, 2009) were used for prediction. The work activity, complexity, location, equipment, climatic condition have also been used to estimate the spatial distribution of work accident risk (Bailey, Cordeiro, & Lourenço, 2007; Raviv, Shapira, & Fishbain, 2017).

Measuring interactions determine how strongly features interact with each other. To measure interactions, the H-statistic proposed by(Friedman & Popescu, 2008) is used. This statistic measures how much of the variation of the predicted outcome depends on the interaction of predictors. The interaction strength is zero when there is no interaction at all and one if all variations of the predicted outcome depend on a given interaction. Interactions between predictors of DNN_{reg1} and DNN_{reg4} are depicted in Fig. 6(d) and 6(e). The predictor "*employment_contract*" exhibits the strongest interaction signal for DNN_{reg1} and the remaining other relevant features contributing to lost days (*loss time*) are climatic conditions, equipment, lineman's experience, distance to sites amongst others. For DNN_{reg4}, the predictor "project_type" has the strongest interaction signal while other important features contributing to equipment damage are location, project complexity, equipment state, amongst others. The partial dependence plot (PDP) and individual conditional expectation (ICE) curves can be used to show effects of a relevant predictor (i.e. experience) on the response (body part injured) for its entire range of values (i.e. experience levels: 1-fresh, 2-mid-level experts, or 3-experts) of a lineman. Fig. 6(f) depicts PDP in thick yellow (averages of all observations for each response class) and ICE curves to compare the marginal impact of the feature "experience "on the occurrence of an injury to body parts (body injury) while a lineman is carrying out a construction task.

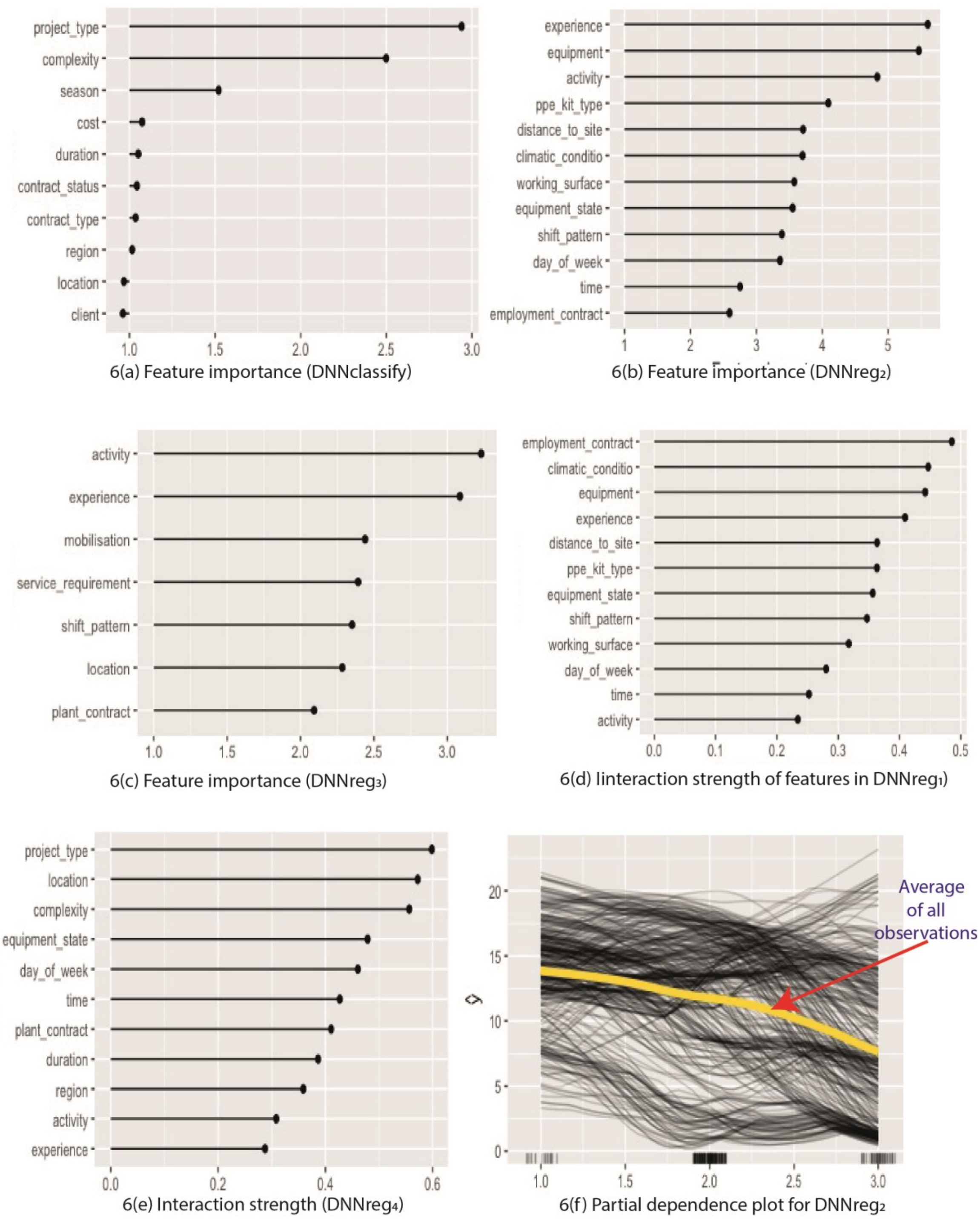


Fig. 6. DNN models interpretation (Feature importance, Interaction strength and Partial dependence)

The DNN model (DNNreg$_2$) captures a non-linear and a monotonic decrease in the occurrence of an injury to body parts. Experience of a lineman on the task reduces the probability of injury to body parts. In addition, a local interpretable model-agnostic explanations (LIME) (Ribeiro, Singh, & Guestrin, 2016) was implemented (Figs 7(a) and 7b) to provide local explanations for models. For instance, Fig. 7(a) fits a local model to the DNN$_{reg2}$ model for the observation (i.e. a lineman) with the lowest probability for ankle/feet injury by looking at ten features in the DNN_{reg2} model that are most influential. As shown in Fig. 7a, features activity (i.e. wiring), equipment state (i.e. good state), time (i.e. morning), and equipment (i.e. pliers) have little influence on this lineman having an ankle/feet injury while working on a given workplace terrain.

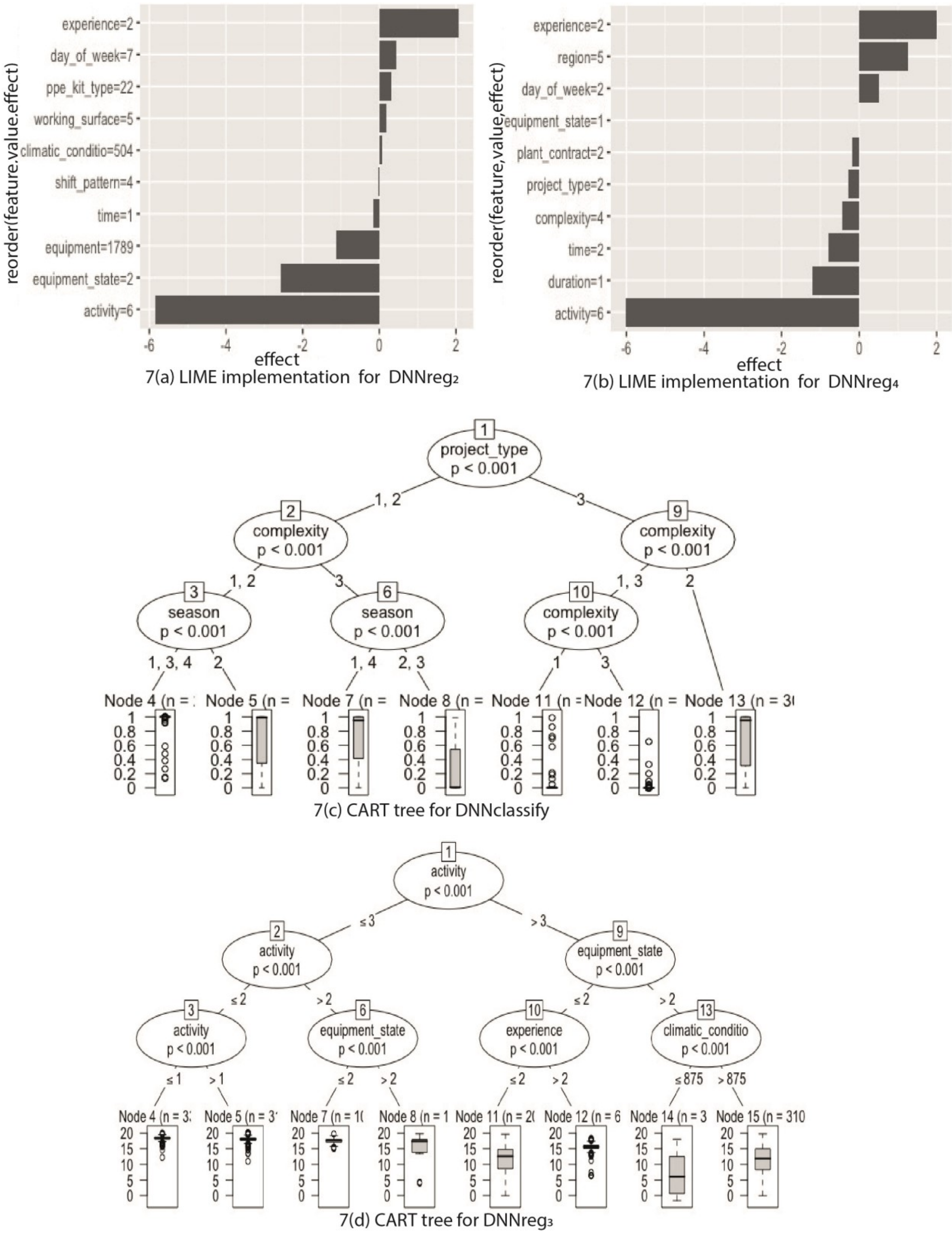


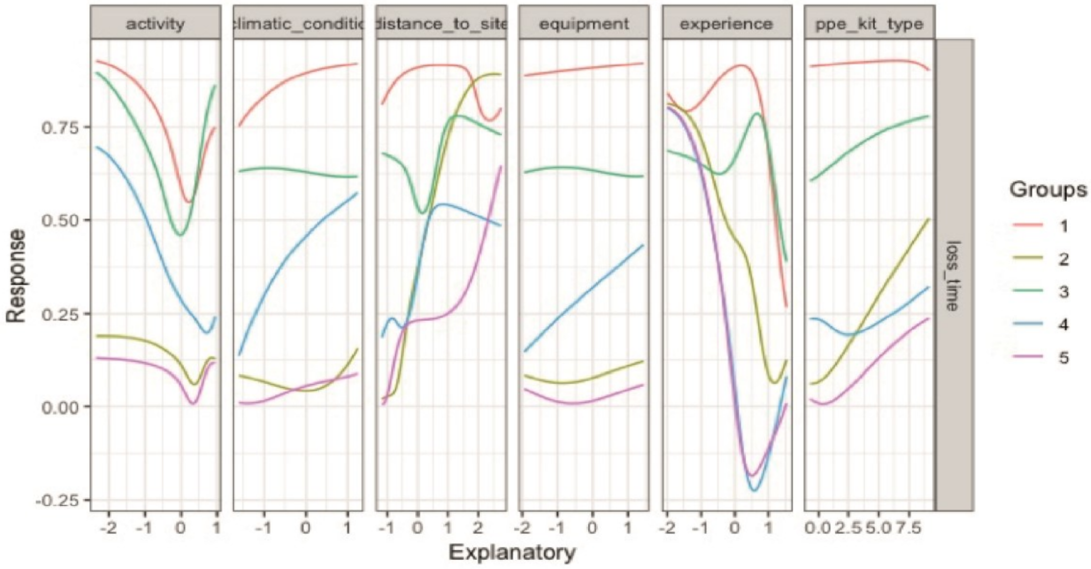
Fig. 7. Interpretation of DNN models (LIME and Surrogate models)

However, features such as "*experience*", "*day*", "*working surface*", and "*PPE kit type*" have a sizable influence on the occurrence of an ankle/feet injury for this observation. Similarly, a surrogate model was used to interpret the deep learning models (Figs 7c and 7d). Fig. 7(c) for example, shows a decision tree to mimic the DNN_classify model in respects to accident (risk) probability, with “project_type”, “complexity”, “cost”, and “season” being the most discriminative features for risk prediction.

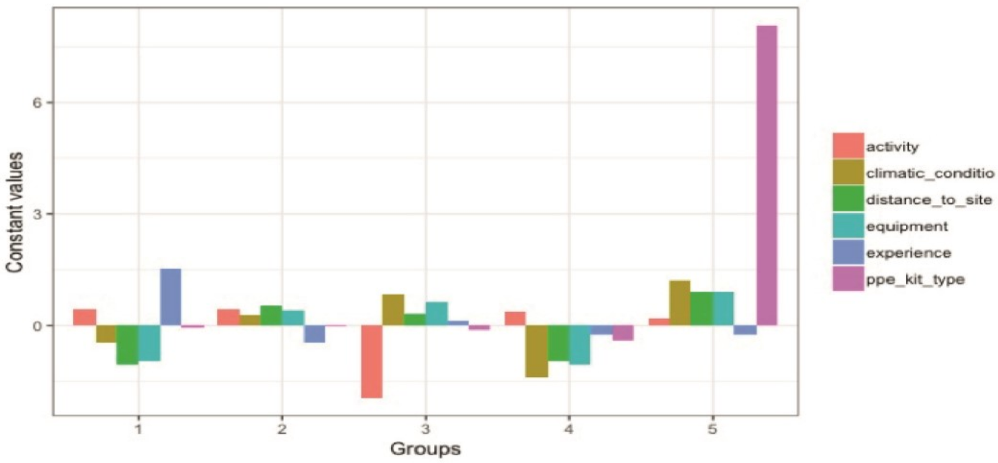
As depicted in Fig. 7(c), an overhead or underground PT&D cabling project ($type \in \{1,2\}$) either a new build or refurbished i.e. $complexity \in \{1,2\}$, has probability of risk around 0.97, around 0.75 (if less costly i.e. cost=1) or around 0.98 (if costlier i.e. $cost \in \{2,3,4\}$). For a new build substation project requiring repairs i.e. type=3 and $complexity \in \{1,3\}$ and $season \in \{1,2,3\}$, the probability of a risk is around 0.00, and around 0.73 (if season =4). The probability of a risk is around 0.98 for project type= 3 (substations) and complexity =2 (refurbished). These decision rules can help safety managers to better identify risks accurately and provide mitigation plans based on project characteristics. For instance, if an accident is expected to occur on site with a given probability (say 73%) safety managers having this information can adjust project schedules to alter this estimation or eliminate the risk. Fig. 7d is interpreted in a similar manner.

A sensitivity analysis to identify predictors that potentially influence response variables was done using the Lek profile method (Gevrey, Dimopoulos, & Lek, 2003). The *lekprofile* function (Beck, 2018) was used to evaluate the effects of predictors by returning a plot of model predictions across the range of values for each predictor with the remaining explanatory variables constantly held while evaluating the effects of each predictor. A typical illustration is made to show the influence of predictors on the response variable lost days (loss time) at a typical PT&D construction site. Key factors potentially relating to "loss time" are extracted by the deep learning model (in this case, $DNNr_{eg1}$) from the dataset. They include lineman's experience, activity, distance to the site, climatic condition, equipment, and the type of the PPE kit. We scaled and centred the predictors and scaled the response variable to 0–1. The result from the *lekprofile* function (Fig. 8) shows non-linear responses that vary by different groupings of the data. Values for each variable in the different unevaluated groups (based on clustering) show that there were no apparent patterns between groups, with the exception being group five that had higher values for the PPE type. As the interpretable aspect of models developed in the study has revealed, the key features H&S managers should consider when implementing a robust safety strategy for the power infrastructure project. The features are “activity”, “equipment”, “experience”, “project_type”, “project_complexity”, location, etc. Though, these attributes have been identified in previous studies (Cheng et al., 2012; Cheng, Lin, & Leu, 2010;

Chi & Han, 2013) as sources of accidents, this study has also identified additional key attributes namely "distance_to_site"," shift pattern", and "PPE_kit_type".



8a Grouping of unevaluated explanatory variables by cluster means (five groups)



8b Values at which explanatory variables are held constant for each group

Fig. 8. Sensitivity analysis of DNN_{reg1} using the Lek profile method

Long hours of commuting to sites, long shift sequences, and long hours are associated with increased fatigue, decreased alertness and concentration, increased errors, heart attack, musculoskeletal disorders, and resentment at work (Hoehner, Barlow, Allen, & Schootman, 2012; Kivimäki, Nyberg, Batty, Madsen, & Tabák, 2018). A human error which depends on sleep-related factors is an essential factor in work accidents. Contractors may supply sub-standard or inappropriate kits that may not last or put the wearer in risk. In tackling these issues, additional staff training increased inspections, and flexible working practices should be done to prevent accidents. Also, PT&D organisations should endeavour to make available gyms to stimulates and refreshes the brain.

Besides, state-of-the-art technological innovations should be embraced. For example, remote-controlled equipment and intelligent machines can significantly alleviate operators' exposure to hazardous sources.

3.6. Prediction Ability of Deep Learning for H&S Risks

To verify all models ($DNN_{classify}$, DNN_{reg1}, DNN_{reg2}, DNN_{reg3}, DNN_{reg4}, DNN_{reg5}) applicability, they were tested on the test data (20% of the H&S event data) that were not replicates or used in the training process to determine their generalisation and prediction abilities. The prediction results of all the DNN models are noted and recorded. The confusion matrix and other metrics depicted in Table VII illustrate the performance of $DNN_{classify}$ to predict whether or not accidents will occur given any PT&D project parameters. The model achieved an accuracy of 0.9325 (~ 93%), and an estimated AUC value of 0.9331 for the proportion of correctly ranked "positive"-"negative" pairs. Figs. 9 and 10 depict both DNN_{reg1} and DNN_{reg2} prediction accuracies tested on randomly sampled 16 PT&D infrastructure projects executed between January 2015 to December 2016. Fig. 9 depicts the number of days ("Loss time") that an employee is absent from work due to injury, as predicted by DNN_{reg1} (R^2=0.946). In Fig. 9, we observe that the majority of values of the dependent variable "Lost time" (shown between the two horizontal dashed lines) in the test data are between 1 and 4 days, with the following frequencies: 22% (1 day), 17% (2 days), 11% (3 days), and 17% (4 days), giving a total of 67% of lost days between 1 and 4.

There is, therefore, the need to reduce these lost times (days) during the project execution. Thus, the proposed H&S risk management tool provides a platform for stakeholders to investigate both current and future PT&D projects concerning probable risks and implement appropriate avoidance strategies before accidents occur.

Table VII. Confusion Matrix and other Performance Metrics

		Actual	
Predicted		No risk	Yes risk
	No risk	2507	59
	Yes risk	183	837

Performance metrics	
Accuracy	0.9325
Kappa	0.8279
Sensitivity	0.9320
Specificity	0.9342
AUC	0.9331

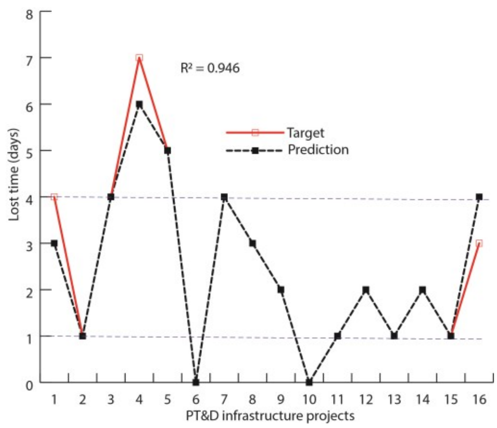


Fig. 9. Predicted vs actual number of lost days (DNN_{reg1})

Similarly, in Fig. 10, we depict the ability of DNN_{reg2} ($R^2 = 0.928$) to predict the injured body part of linemen from 16 H&S event test data consisting of overhead lines projects where a specific activity "tower erection" was carried out. Correctly predicted body parts from the test data are shown in green, while incorrect predictions are indicated in red. DNN_{reg2} predicted correctly injured body parts for approximately 70% of the projects except for those labelled 2, 7, 8, 11, and 15, but despite this, the predicted body parts are in proximity to targets with respect to their locations in the body. For instance, for project 7, 'neck' as target but 'shoulder' was predicted. Similarly, for project 8, 'knee' is the target but 'foot' predicted. In summary DNN_{reg2} (Body injury) predicted the body parts Fingers, Ankle, Back/Buttocks, and Knee, which are frequently prone to injuries. This result is in agreement with various studies (Aasa, Barnekow-Bergvist, Angquist, & Brulin, 2005; Fan et al., 2014; Sanchez et al., 2015). These results established the good generalisation ability of the model: for a given input, the model could reasonably predict the dependent variables that defined incidents in power infrastructure projects. In terms of the training time, the complexities of all the deep learning architectures for the six models are less complicated in terms of the reduced training times used to build them. We attribute this to the parallelised SGD used as the training algorithm, in addition to today's Graphics Processing Unit (GPU)-based computers, which have a million times the computational power of a desktop (Schmidhuber, 2015). Thus, the developed DNN models will execute conveniently on most machines. Prediction accuracies of all DNN models on the test data are depicted in Table VIII. From the results of predictive modelling, we discover that the key sources of incidents in power infrastructure projects are "caught in/between", "struck by/between", "cutting", "driving on uneven ground" and "falling objects". These sources cause injuries to body parts such as fingers/thumbs, back/buttocks, hands and ankles, which are essential for carrying out plant- and equipment-related operations such as lifting, loading, cutting, and pulling (Chi & Han, 2013).

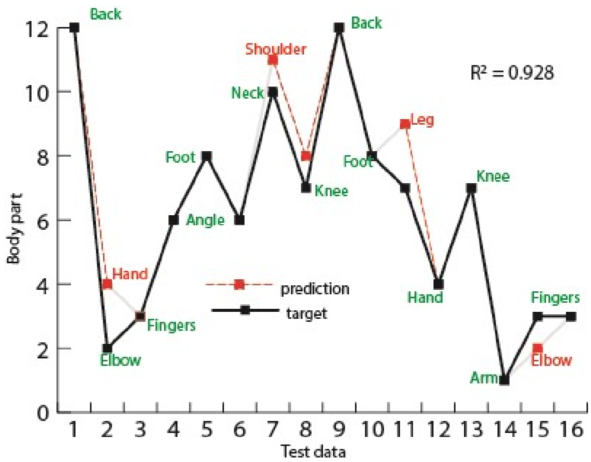


Fig. 10. Predicted vs. actual injured body parts (DNN_{reg2})

Table VIII. Prediction Accuracy of Models

Name	Four key predictors	Type	Performance metrics	Values
$DNN_{classify}$	Project type, complexity, season, and duration	Classification	Accuracy	0.9325
			AUC	0.9331
DNN_{reg1}	Experience, activity, climatic conditions, working surface layout	Regression	MAE	0.7105
			R^2	0.9463
DNN_{reg2}	Experience, activity, distance to site, working surface layout	Regression	MAE	0.7440
			R^2	0.9282
DNN_{reg3}	Location, activity, operator experience, PPE kit type	Regression	MAE	0.6075
			R^2	0.9356
DNN_{reg4}	Activity, operator's experience, equipment's age, equipment state	Regression	MAE	0.6121
			R^2	0.9382
DNN_{reg5}	Working surface layout, location, activity, equipment	Regression	MAE	0.6302
			R^2	0.9472

In Table IX, we show the proportions of injuries to the five most frequently injured body parts. Accordingly, using appropriate protective equipment such as gloves, helmets, and boots should be encouraged to reduce the number of incidents. Similarly, there is a need for mechanisation or automated lifting and loading equipment that requires minimal human interaction. Excavation operations are the source of damage to utility services in 45% of incidents. This is attributed to lack of data on positions of buried utilities, ground terrains, and misjudgements by machine operators. Relevant data on soil conditions should be explored before excavating. Additionally, using automated detection tools to locate utility service positions should be encouraged. Damages to vehicles (windows, windscreens, bumpers, bonnets, front and rear lights, scratches, dents and punctured tires) represent the highest proportion of damage incidents to plant and fleet (30%). Sources (ranked by prevalence) are "vandalism by a third party", "weather", "uneven terrain", "animals", "road traffic accident", and "drivers' errors". Safety knowledge management and training should be intensified to ensure that workers can identify safety signs and hazards. The ability to perceive hazards is strongly related to safety performance (Zhou, Goh, & Li, 2015).

Table IX. Proportions of Injuries to Body Parts

Body part	Proportion
Fingers/Thumbs	21%
Back/Buttocks	13%
Hand	10%
Ankle	09%
Knee	08%

3.7. Implication for Practice

This study offers some implication for practice in terms of the following:

1. Efficient H&S risk assessment- The most common problem of risk management practices in the pre-project stage is ineffective risk identification practices. The tool developed in this study provides an effective way to forecast and visualise risk-related events that may emerge during the PT&D project execution. The tool adopts a robust text mining approach that is used to process and extract meaningful patterns from the dataset. Currently, risk assessment practices involve a high level of subjectivity due to over-reliance on intuition, judgment or the individual experience of decision-makers when predicting H&S risks and their impacts on project goals. The tool, therefore, offers H&S managers precise methodology data to assess the risk level of any PT&D project. The assessment will reduce inconsistencies and vagueness in risk ratings. Besides, the tool provides information about the status of all projects in respect to probable H&S risks and appropriate mitigation strategies. Thus, the tool will allow H&S managers to oversee several PT&D projects at the same time efficiently. In addition, the models developed have also revealed the key features that H&S managers should consider when implementing a robust safety strategy for a safe environment. In addition to "activity", "equipment", "experience", "project_type", "project_complexity", location, etc., this study has also identified additional key attributes namely "distance_to_site"," shift pattern", and "PPE_kit_type."
2. A user-interface for H&S risk analysis and new data for future study- The developed deep learning models are incorporated into a software module with a rich user output, which enables H&S managers to make practical, informed decisions in the field. Similarly, the tool allows for effective methods of collection and storage of new data to address the data limitation problem caused by inefficient data logging methods currently employed. The new data could be used to support continuous deep

reinforcement learning to improve the H&S risk prediction and exhaustive sensitivity analysis for performance verification. Predictions and follow-ups by H&S managers will improve risk management.

3. Future tool development for H&S risk analysis- This study could also influence the H&S risk management practices in the power infrastructure environment. Although several studies suggest that deep learning and Big Data technologies capability are critical for efficient risk management, deep learning techniques for H&S risk management are often ignored in the construction safety research community. The increasing success of deep learning techniques in other fields and the governments' commitment to H&S reforms have compelled more industry practitioners to integrate robust risks management practices into the H&S software. This study, therefore, provides a clear direction on how to achieve this by combining deep neural networks and big data technology for H&S risks management in a PT&D environment. This study also has vast implications for H&S software developers. The recent advances in machine learning techniques and Big data technologies show that innovation within the HSE industrial practices requires H&S compliance. Besides, complex, and repetitive power infrastructure tasks need to be automated to achieve the expected reliability and efficiency. As such, the framework employed in this study and the H&S risk tool development process serves as a blueprint for developing H&S risk-enabled software for PT&D H&S risk management and related tasks.

4. MODEL USER INTERFACE

We develop a front-end system called Incident Reporting for Power Infrastructure Projects (IRePIP) using the Python language, which will enable users to make informed decisions regarding H&S events. This system allows input parameters to be specified by users. The interface triggers DNN models to predict the occurrence likelihoods of H&S events. The system notifies stakeholders of probable risks to humans, equipment and the environment associated with certain operations (excavating, wire pulling, loading), along with appropriate recommendations to mitigate such risks. For instance, as shown in Fig. 11, the body parts that are most likely to be injured are graphically indicated to allow safety managers to prioritise H&S risk factors according to their likelihoods of occurrence. Thus, adequate attention is paid to these risk factors when controlling incidents to achieve a safer environment.

The system communicates probable consequences and recommends appropriate plans (improved equipment and maintenance, improved training, work procedures, etc.) to mitigate identified incidents, even before mobilising staff to the sites.

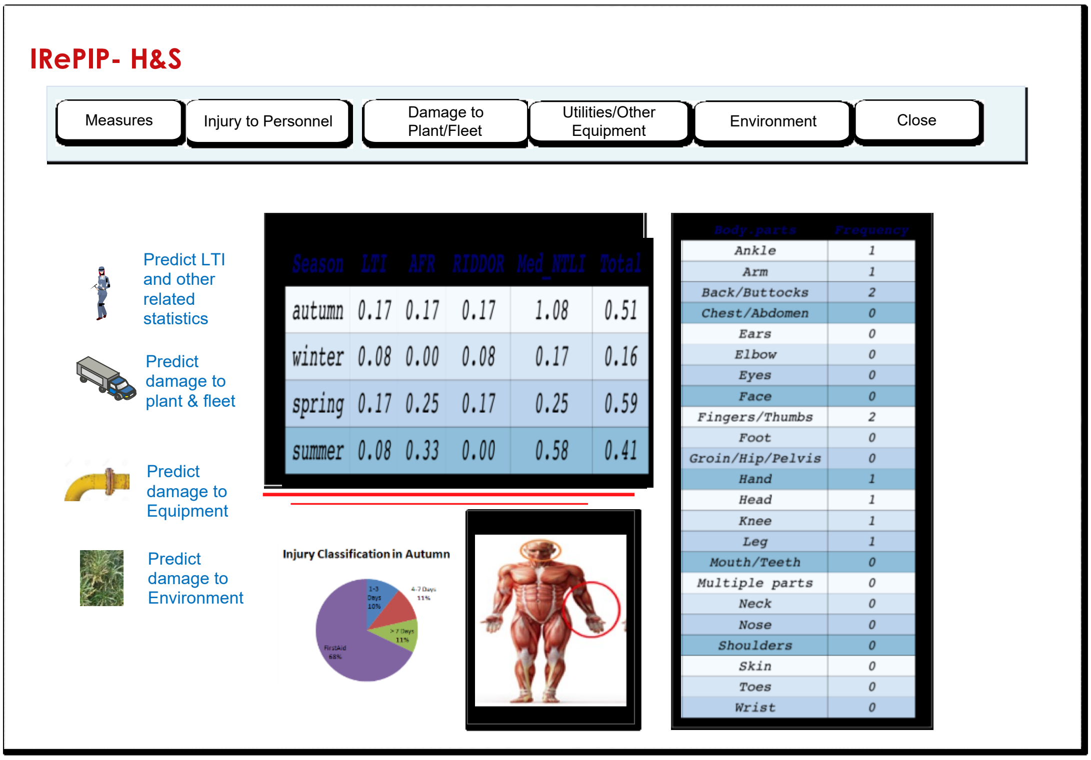


Fig. 11. User interface

5. CONCLUSIONS

Despite the availability of H&S incident data, utilising them to effectively mitigate accident occurrence is challenging because of the data logging method that is employed for recording incidents. Free-text format is often used to describe incidents in databases. For a database that has grown to a considerable size, analysing and identifying relevant information may be difficult for humans. Similarly, in the H&S research community, new AI techniques (deep neural networks, convolution neural networks, recurrent neural network, etc.) have often been ignored despite their considerable attention in other research fields (LeCun et al., 2015). This study explored the deep neural network method in power infrastructure projects and prompted the readers to explore its usefulness especially in predicting H&S risks at power infrastructure construction sites. In contrast to traditional machine learning and artificial intelligence approaches, the deep learning has demonstrated due to its success that predictive accuracy can be enhanced, and expert feature engineering dispensed with, by fitting highly flexible models that are capable of learning novel representations. We implemented DNN models and a user interface to predict H&S risks in power infrastructure projects to minimise costs (third-party insurance, repairs to utilities and equipment). We utilised a text mining approach for the following: to process free-text columns, investigate events using word combinations, and complete columns with missing entries. We fill in the remaining missing entries (less than 10%) in the database using the mean imputation method. In the context of this study, we perform risk assessment by giving higher priority to accidents that result in nonfatal injury or damage to the plant, fleet, underground utilities or other facilities. The DNN models are developed coherently using the H2O framework in R software. We evaluated the DL models on unseen data and obtained high prediction accuracies. The classification model achieved an AUC of 0.93, while the regression models had mean absolute errors in the range of 0.91 to 0.94. Regarding the computational aspect, the training times required by the models are negligible, which is evidence of the method's computational efficiency. This research is valuable for stakeholders who need information regarding probable H&S risks to design strategies and safety measures to mitigate such risks.

The adopted approach proved that employing DL in engineering and related tasks is timely. This is because the industry is becoming conscious of the need to collect massive amounts of unstructured data and to elicit meaningful values for decision-making. As such, using DL will offer an edge in extracting useful insights from huge data sets. In future work, we hope to develop a mobile version and interface the server with real-time data sources (such as British Geological Survey, Google terrain and Metrological office). The biggest issue with the deep learning is its black-box problem (how outputs are arrived at), and theories to interpret its results are still unavailable. Despite this challenge, most machine learning researchers still consider deep learning as the most effective and supervised machine learning approach due to its high predictive accuracy (Bengio, 2012; LeCun et al., 2015). In the future, we will explore the use of visualisation techniques to interpret deep learning outputs. This study implies that the deep learning models developed have revealed key predictors that H&S managers in power infrastructure should consider when implementing a safety strategy to help reduce accidents; the integration of DNN with IML will also assist managers and future scientists in deep learning models' interpretation. Also, the rich user-interface developed will provide an opportunity of not only predicting injuries to personnel but providing a holistic platform for predicting damages to equipment, environment, and plant and fleet, which will help to manage H&S in an effective and efficient manner.

A major limitation of this study is that we used a single point of data collection. We hope to obtain more data from several heterogeneous sources and carry out in-depth analyses on other types of DL architectures. Furthermore, techniques such as genetic algorithms and fuzzy logic will be used in conjunction with DL to enhance insights into data.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to Innovate UK and EPSRC (Grant Ref EP/S031480/1) for providing the financial support for the research.

REFERENCES

Aasa, U., Barnekow-Bergvist, M., Angquist, K., & Brulin, C. (2005). Relationship between work-related factors and disorders in the neck-shoulder and low-back region among female and male ambulance personnel. *Journal of Occupational Health*, *47*, 481–489.

Aksorn, T., & Hadikusumo, B. (2008). Critical success factors influencing safety program performance in Thai construction projects. *Safety Science*, *46*(4), 709–727.

Al-Rahhal, M. ., Bazi, Y., Alhichri, H., Alajlan, N., Melgani, F., & Yager, R. R. (2016). Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, *345*, 340–354.

Bailey, T. C., Cordeiro, R., & Lourenço, R. . (2007). Semiparametric modeling of the spatial distribution of occupational accident risk in the casual labor market, Piracicaba, Southeast Brazil. *Risk Analysis*, *27*(2), 421–431.

Beck, M. . (2018). NeuralNetTools: Visualization and Analysis Tools for Neural Networks. *Journal of Statistical Software*, *85*(11), 1–20.

Bengio, Y. (2012). Practical Recommendations for Gradient-Based Training of Deep Architectures. In G. Montavon, G. B. Orr, & K. Müller (Eds.), *Lecture Notes in Computer Science* (pp. 437–478). Springer Berlin Heidelberg.

Bohle, P., Quinlan, M., McNamara, M., Pitts, C., & Willaby, H. (2015). Health and well-being of older workers: Comparing their associations with effort–reward imbalance and pressure, disorganisation and regulatory failure. *Work & Stress*, *29*(2), 114–127.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Y. Lechevallier & G. Saporta (Eds.), *Computational Statistics proceedings of the nineteenth International Conference* (pp. 177–187). Paris, France.

Chen, J., Song, L., Wainwright, M. J., & Jordan, M. I. (2018). *Learning to Explain: An Information-Theoretic Perspective on Model Interpretation*. Retrieved from http://arxiv.org/abs/1802.07814

Cheng, C., Leu, S., Cheng, Y., Wu, T., & Lin, C. (2012). Applying data mining techniques to explore factors contributing to occupational injuries in Taiwan's construction industry. *Accident Analysis and PreventionPrevention*, *48*, 214–222.

Cheng, C., Lin, C., & Leu, S. (2010). Use of association rules to explore cause-effect relationships in occupational accidents in the Taiwan construction industry. *Safety Science*, *48*(4), 436–444.

Chi, S., & Han, S. (2013). Analyses of systems theory for construction accident prevention with specific reference to OSHA accident reports. *International Journal of Project Management*, *31*(7), 1027–1041.

Dahl, G. E., Sainath, T. N., & Hinton, G. E. (2013). Improving deep neural networks for LVCSR using rectified linear units and dropout. *IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 8609–8613). Vancouver, BC, Canada.

Daniel, B. K. (2017). *Big data and learning analytics in higher education: Current theory and practice* (1st ed.). Switzerland: Springer.

Debnath, J., Biswas, A., Sivan, P., Sen, K., & Sahu, S. (2016). Fuzzy inference model for assessing occupational risks in construction sites. *International Journal of Industrial Ergonomics*, *55*, 114–128.

Englehardt, J. D., An, H., Fleming, L. E., & Bean, J. A. (2003). Analytical predictive bayesian assessment of occupational injury risk: Municipal solid waste collectors. *Risk Analysis*, *23*(5), 917–927.

Esmaeili, B., Hallowell, M. R., & Rajagopalan, B. (2015). Attribute-based safety risk assessment. II: Predicting safety outcomes using generalized linear models. *Journal of Construction Engineering and Management*, *141*(8), 1–11.

Fan, C., Xiao, F., & Zhao, Y. (2017). A short-term building cooling load prediction method using deep learning algorithms. *Applied Energy*, *195*, 222–233.

Fan, Z., Silverstein, B., Bao, S., Bonauto, D., Howard, N., & Smith, C. (2014). The association between combination of hand force and forearm posture and incidence of lateral epicondylitis in a working population. *Human Factors*, *56*(1), 151–165.

Farid, D., Al-Mamun, M., Manderick, B., & Nowe, A. (2016). An adaptive rule-based classifier for mining big biological data. *Expert Systems with Applications*, *64*, 305–316.

Fenrick, S. A., Getachew, L., Fenrick, L., & Getachew, S. (2012). Cost and reliability comparisons of underground and overhead power lines. *Utilities Policy*, *20*(1), 31–37.

Friedman, J. H., & Popescu, B. E. (2008). Predictive learning via rule ensembles. *Annals of Applied Statistics*, *2*(3), 916–954.

Gevrey, M., Dimopoulos, I., & Lek, S. (2003). Review and Comparison of Methods to Study the Contribution of Variables in Artificial Neural Network Models. *Ecological Modelling*, *160*(3), 249–264.

Gharbi, M., Chen, J., Barron, J. T., Hasinoff, S. W., & Durand, F. (2017). Deep bilateral learning for real-time image enhancement. *ACM Transactions on Graphics*, *36*(4), 1–12.

Gibson, A., & Patterson, J. (2017). *Deep Learning* (1st ed.). Sebastopol, CA: O'Reilly Media, Inc.

Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E. (2015). Peeking In- side the Black Box: Visualizing Statistical Learning with Plots of Individual Condi- tional Expectation. *Journal of Computational and Graphical Statistics*, *24*(1), 44–65.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., … Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In J. Kolen & S. Kremer (Eds.), *A Field Guide to Dynamical Recurrent Networks* (pp. 237–244). New York: IEEE.

Hoehner, C. M., Barlow, C. E., Allen, P., & Schootman, M. (2012). Commuting distance, cardiorespiratory fitness, and metabolic risk. *American Journal of Preventive Medicine*, *42*(6), 571–578.

HSE. (2016). Health and safety at work Summary statistics for Great Britain. Retrieved April 14, 2017, from http://www.hse.gov.uk/statistics/overall/hssh1516.pdf?pdf=hssh1516

Jacinto, C., & Silva, C. (2010). A semi-quantitative assessment of occupational risks using bow-tie representation. *Safety Science*, *48*, 973–979.

Kivimäki, M., Nyberg, S. T., Batty, G. D., Madsen, I. E. H., & Tabák, A. G. (2018). Long Working Hours and Risk of Venous Thromboembolism. *Epidemiology*, *29*(5), e42–e44.

Kochura, Y., Stirenko, S., Alienin, O., Novotarskiy, M., & Gordienko, Y. (2017). *Performance Analysis of Open Source Machine Learning Frameworks for Various Parameters in Single-Threaded and Multi-Threaded Modes*. (Ml), 1–15. Retrieved from http://arxiv.org/abs/1708.08670

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Larochelle, H., Bengio, Y., Louradour, J., & Lamblin, P. (2009). Exploring strategies for training deep neural networks. *Journal of Machine Learning Research*, *10*, 1–40.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. Y. (2009). Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, *54*(10), 95–103.

Liu, M., Zhang, D., & Shen, D. (2012). Ensemble sparse classification of Alzheimer’s disease. *NeuroImage*, *60*(2), 1106–1116.

Liu, Y., Li, B., Tan, R., Zhu, X., & Wang, Y. (2014). A gradient-boosting approach for filtering de novo mutations in parent-offspring trios. *Bioinformatics*, *30*(13), 1830–1836.

Ma, J., Sheridan, R. P., Liaw, A., Dahl, G. E., & Svetnik, V. (2015). Deep neural nets as a method for quantitative structure-activity relationships. *Journal of Chemical Information and Modeling*, *55*, 263–274.

Mamoshina, P., Vieira, A., Putin, E., & Zhavoronkov, A. (2016). Applications of deep learning in biomedicine. *Molecular Pharmaceutics*, *13*(5), 1445–1454.

Mayr, A., Günter, K., Unterthiner, T., & Sepp, H. (2016). DeepTox:Toxicity prediction using deep learning. *Frontiers in Environmental Science*, *3*, 1–15.

McDermott, V., & Hayes, J. (2016). “We’re still hitting things”: The effectiveness of third party processes for pipeline strike prevention. *Proceedings of the Eleventh International Pipeline Conference (pp*. 1–10). Calgary, Alberta, Canada.

Mistikoglu, G., Gerek, I. H., Erdis, E., Usmen, P. E. M., Cakan, H., & Kazan, E. E. (2015). Decision tree analysis of construction fall accidents involving roofers. *Expert Systems with Applications*, *42*(4), 2256–2263.

Molnar, C., Casalicchio, G., & Bischl, and B. (2018). iml: An R package for Interpretable Machine Learning. *The Journal of Open Source Software*, *3*(27), 1–2.

Papazoglou, I. I. A. I., Aneziris, O., Bellamy, L., Ale, B. J. M. B., & Oh, J. J. I. H. J. (2015). Uncertainty Assessment in the Quantification of Risk Rates of Occupational Accidents. *Risk Analysis*, *35*(8), 1536–1561.

R Development Core Team, R., & R Core Team, R. (2016). *R: A Language and Environment for Statistical Computing*. Retrieved from https://www.r-project.org/

Rastogi, A., & Gabbar, H. A. (2013). Fuzzy-logic-based safety verification framework for nuclear power plants. *Risk Analysis*, *33*(6), 1128–1145.

Raviv, G., Shapira, A., & Fishbain, B. (2017). AHP-based analysis of the risk potential of safety incidents: Case study of cranes in the construction industry. *Safety Science*, *91*, 298–309.

Recht, B., Re, C., Wright, S., & Niu, F. (2011). Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In J. Shawe-Taylor, R. S. Zemel, P. Bartlett, F. C. Pereira, & K. Weinberger (Eds.), *Advances in Neural Information Processing Systems* (pp. 693–701). Granada, Spain.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?” Explaining the Predictions of Any Classifier". *20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, (pp. 1135–1144). San Francisco, CA, USA.

Sanchez, A., Iglesias-Rodriguez, F., Fernandez, R., de Cos Juez, F., Fernandez, P., de Cos Juez, F., … Su, A.

(2015). Applying the k-nearest neighbor technique to the classification of workers according to their risk of suffering musculoskeletal disorders. *International Journal of Industrial Ergonomics*, *52*, 92–99.
Schmidhuber, J. (2015). Deep learning in neural networks : An overview. *Neural Networks*, *61*, 85–117.
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.
Tixier, A. J. P., Hallowell, M. R., Rajagopalan, B., & Bowman, D. (2016). Application of machine learning to construction injury prediction. *Automation in Construction*, *69*, 102–114.
Törner, M., & Pousette, A. (2009). Safety in construction - a comprehensive description of the characteristics of high safety standards in construction work, from the combined perspective of supervisors and experienced workers. *Journal of Safety Research*, *40*(6), 399–409.
Tsoukalas, V. D., & Fragiadakis, N. G. (2016). Prediction of occupational risk in the shipbuilding industry using multivariable linear regression and genetic algorithm analysis. *Safety Science*, *83*, 12–22.
Vallmuur, K. (2015). Machine learning approaches to analysing textual injury surveillance data: A systematic review. *Accident Analysis and Prevention*, *79*, 41–49.
Weiss, S. M. (2010). *Text mining: Predictive methods for analyzing unstructured information*. New York: Springer.
Zhang, L., Wu, X., Qin, Y., Skibniewski, M. J., & Liu, W. (2016). Towards a Fuzzy Bayesian Network Based Approach for Safety Risk Analysis of Tunnel-Induced Pipeline Damage. *Risk Analysis*, *36*(2), 278–301.
Zhao, Z., Chen, W., Wu, X., Chen, P. C. Y., & Liu, J. (2017). LSTM network: A deep learning approach for short-term traffic forecast. *IET Intelligent Transport Systems*, *11*(2), 68–75.
Zhou, Z., Goh, Y., & Li, Q. (2015). Overview and analysis of safety management studies in the construction industry. *Safety Science*, *72*, 337–350.