TESTING FOR KNOWLEDGE: MAXIMISING INFORMATION OBTAINED FROM FIRE TESTS BY USING MACHINE LEARNING TECHNIQUES

Arjan Dexters, School of Engineering, University of Edinburgh, United Kingdom

Rolff Ripke Leisted, Department of Civil Engineering, Technical University of Denmark

Ruben Van Coile, Department of Structural Engineering, Ghent University, Belgium

Stephen Welch, School of Engineering, University of Edinburgh, United Kingdom

Grunde Jomaas, School of Engineering, University of Edinburgh, United Kingdom

ABSTRACT

A machine learning (ML) algorithm was applied to predict the onset of flashover in 1:5 scale Room Corner Test experiments with sandwich panels. Towards this end, a penalized logistic regression model was chosen to detect the relevant variables and consequently provided a tool that can be used to make predictions of unseen samples. The method indicates that a deeper understanding of the contributing factors leading to flashover can be achieved. Furthermore, it allows a more nuanced ranking than currently offered by the commonly used classification methods for reaction to fire tests. The proposed methodology shows a substantial value in terms of guidance for future large and intermediate scale testing. In particular, it is foreseen that the method will be extremely useful for assessing and understanding the behaviour of innovative materials and design solutions.

INTRODUCTION

Fire-classification of materials is used in conjunction with relevant legislation with the intent to be part of a design that can assure adequate fire safety in buildings. The classification of a product should in principle be deduced from its behaviour to fire in a scenario that represent the end-use situation, i.e., a full-scale test. However, as such tests are often costly and labour intensive, a tendency exists to try to predict full-scale fire behaviour based on small-scale testing. In order to justify such a scaling methodology a thorough understanding of the fire behaviour is necessary [1]. While this is currently the case for many single burning items, a knowledge gap persists on the interaction of fire [illegible]ded to accurately classify such materials.

The full-scale Room Corner Test (RCT) [2] used to be the standard for classification of linings in a variety of countries. However, because it requires rather large samples, it was neither considered to be cost- nor time efficient and it was therefore replaced with the new European intermediate-scale Single Burning Item (SBI) test. The concept of scaling-down seems justified, as in 87 per cent of the cases the full-scale fire growth behaviour is captured adequately by the intermediate-scale [3]. Nevertheless, for materials such as sandwich panels, linear systems and polycarbonate panels, the correlation proved to be less accurate [3]. New tests were developed for some specimens, e.g., linear systems, while for others the dependency for classification remained with the SBI test. As such, it is foreseen that dangerous situations could arise due to the possible misclassification of these materials. The latter is further magnified as the demand for innovative materials, e.g., sandwich panels, rises due to the surge in the passive housing market. For these reasons, the relevance of the RCT remains, and the need for an accurate screening method based upon bench-scale test results, or an improvement of the intermediate-scale test, must be researched to provide the industry with an accurate, easy and cost-effective method for quality control and product development.

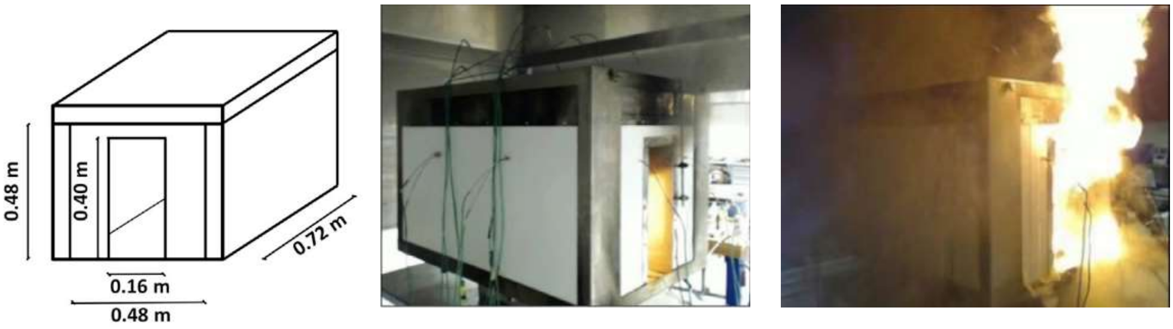
Various screening models have been developed to predict the occurrence of flashover in the RCT based on bench-scale test results [4–6]. Although these models showed promising results, especially after the adjustments by Hansen et al. [5], they still failed to instil enough confidence to reduce the dependence on large-scale and intermediate-scale testing. The reasons for this could be the dependency on rigid and empirical derived equations, which are only valid for a certain application range. These non-learning algorithms might have been viable a decade or more ago, but in an industry that is challenged daily with unprecedented scenarios they are losing relevance at a staggering pace. Looking at fire safety engineering (FSE) from a holistic point of view, the goal is to solve a problem with many context-dependent variables [1] by extracting knowledge from a multitude of fields [7] such as, material science, fluid dynamics and civil engineering. Thus, getting closer to a solution inherently means that researchers have to become more and more specialised, up to the point where it is difficult, if not impossible, to put all the pieces together [8]. Finally, recent efforts made to derive a new intermediate-scale test [9] highlighted that a tool which can identify experimental configurations for their knowledge benefit would significantly reduce research time and cost, and also augment the possibility of success.

For the aforementioned reasons, the need emerges for an approach that can adapt at a pace equivalent to the speed the industry changes with, and, at the same time, is able to find underlying patterns in vast, multi-parameter data sets. One possibility encompasses the application of machine learning (ML), which has already proven its merit in many fields. A foremost advantage of a ML algorithm is that it possesses the capability to learn by way of observation and experience, rather than by using rigid prescribed equations. In summary, a ML algorithm will write its own appropriate model based upon the data it is presented with. To elucidate the latter, Pedro Domingos [8] made a striking low-tech analogy between ML and farming: "Learning algorithms are the seeds, the data is the soil, and the learned programs are the grown plants. The machine learning expert is like a farmer, sowing the seeds, irrigating and fertilizing the soil, and keeping an eye on the health of the crop but otherwise staying out of the way." As such, the algorithm itself can be surprisingly simple but can easily prompt different programs that are magnitudes more complex for varying inputs and without the interference of the ML expert. The simplicity inherently means that ML algorithms can analyse complex problems and vast amounts of variables within seconds, whereas conventional techniques quickly get overwhelmed, either by the limitations of computational-time or computational-space or just because they are too complex to be understood by humans.

The currently proposed model uses ML techniques to predict flashover or no-flashover for a material exposed to three different burning intensities within the physical confines of a 1:5 scale model of the RCT [2]. Many different ML algorithms can be applied to predict a binary output, of which most are referred to as a black-box, meaning that nobody really knows what happens in the in-between state of giving the algorithm the data and the algorithm coming up with new correlations. As such, it is not surprising that a defiance exists in a field where mathematical proof usually is demanded in order to trust computer simulations. For this reason, a logistic regression model was chosen, which is by far the most transparent ML environment. As such, the focus of this study is to provide an insight in basic ML concepts while at the same time providing proof-of-concept that ML can be an easy-to-use tool in fire safety engineering (FSE) to identify relevant parameters, provide guidance for ongoing experimental research and provide a more detailed classification method than currently available.

THE EXPERIMENTAL SETUP AND DATA

Figure 1a depicts the dimensions of the 1:5 scale enclosure. The latter has a width, length and height of 0.48 m, 0.72 m and 0.48 m, respectively, and an opening that was 0.40 m (height) by 0.16 m (width). The enclosure boundary was made from a commercially available sandwich-panel with polyisocyanurate (PIR) foam core with a thickness of either 6 or 10 mm. The 14 experimental observations ($m = 14$) that were used as input for the ML algorithm are summarized in Table 1, which in the remainder of the manuscript is referred to as the historical data set. For more detailed information about the experiments, see Leisted et al. [9].



a) b) c)

Figure 1: a) Dimensions of the 1:5 scale enclosure. b) Experiment resulting in no-flashover. c) Experiment resulting in flashover. Taken from Leisted et al. [9].

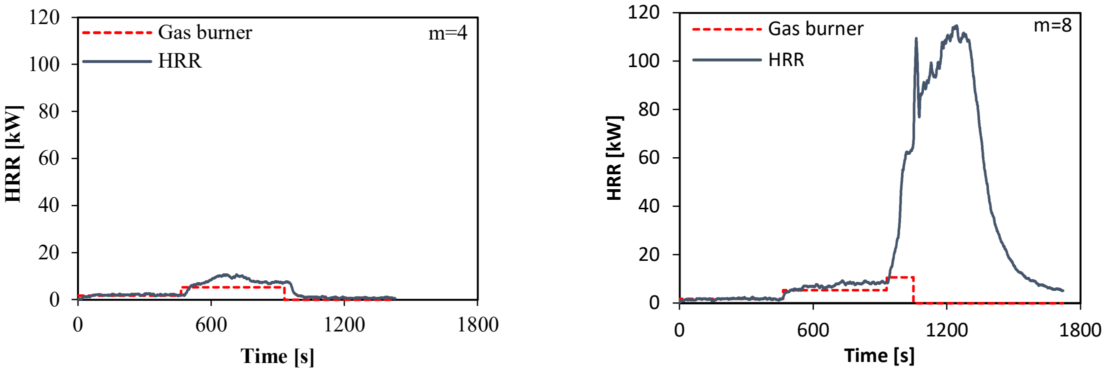
The output variables flashover, $y = 1$, or no-flashover, $y = 0$, were recorded for one type of material (PIR). The following five ($n = 5$) input features x_j $(1 \leq j \leq n)$, i.e., random variables, were varied: The three burning intensities x_{1-3}, the thickness of the material x_4 and the planned burning time of each burning intensity x_5. The input and output can take particular values x_j^i and y^i for every ith observation $(1 \leq i \leq m)$. The time to flashover t_{fo} is listed as an informative feature for the reader but will not be used in the ML model. Figure 1b and 1c show photographs from an experiment with no flashover and with flashover, respectively.

Table 1: The experimental data which will be used for the ML algorithm, i.e., the historical data set. Reproduced from Leisted et al. [10].

m	x_1 [kW]	x_2 [kW]	x_3 [kW]	x_4 [mm]	x_5 [s]	y [-]	t_{fo} [s]
1	179	537	0	10	600	0	-
2	179	537	0	6	465	0	-
3	179	537	0	10	600	0	-
4	179	537	0	6	465	0	-
5	179	537	1074	10	600	1	1270
6	179	537	1074	6	465	1	1009
7	179	537	1074	10	600	1	1249
8	179	537	1074	6	465	1	963
9	179	179	179	10	600	0	-
10	179	537	537	10	600	1	1322
11	537	537	0	10	600	1	847
12	537	537	0	6	465	1	775
13	179	537	1074	6	465	1	1010
14	179	537	1074	6	465	1	1066

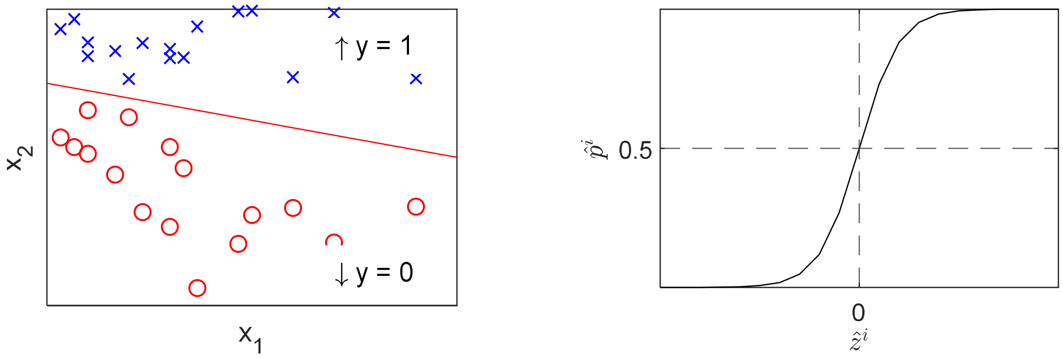
The data is an updated version of the complete data set which was performed by Leisted et al. [9] to derive a new intermediate-scale test which was hypothesized to provide a better correlation for the RCT than the SBI test, with respect to sandwich panels. The particular values of Table 1 are a direct result of keeping a constant Froude number, and as such provide a correlation between intermediate-scale and large-scale. Originally, also the presence of a joint in the specimen build up and the burner location in the enclosure were varied. These aspects were not considered when defining features for the ML analysis as only a few data points were available. Furthermore, one training example was omitted due to poor burner mounting, causing it to have a slight outwards angle. This resulted in a divergence of the output variable, when compared to its two equivalents. As the burner angle is considered vertical in the experimental setup, this observation was omitted.

Figure 2 shows the output of the burner and the corresponding HRR, which was measured using oxygen consumption calorimetry, for test numbers four and eight. Although the onset of flashover was visually observed during the experiments, i.e, flames propagating outside the boundaries of the enclosure, it can also be deduced from the graph as a sudden spike in the HRR. Figure 2b shows that the third burner intensity was turned off at the moment of flashover, i.e., the burning time was smaller than the planned duration x_5. Nevertheless, as the time to flashover t_{fo} is not an a priori known variable, the ML input x_5^8 remains unchanged, as can be seen from the historical data set.



a) *b)*

Figure 2: The burner output and the corresponding HRR profile for test number a) four and b) eight.



a) *b)*

Figure 3: Graphical interpretation of a) the decision boundary $\hat{f}$, represented by the red line, which separates y=1 from y=0 and b) the sigmoid function.

THE MACHINE LEARNING ENVIRONMENT

As a convention, the output variable of interest, flashover, is denoted as *y=1* and no-flashover as *y=0*. The model divides the six-dimensional space, defined by the input features, into a flashover and no-flashover zone with a so-called decision boundary $\hat{f}$. By doing this, it is implicitly assumed that a true division f exists, which corresponds with test setups for which x_{1-5}^{i} results in a 50 per cent flashover chance. As f is almost always unknown, the goal of the machine learning algorithm is to make an approximation $\hat{f}$ of f. In two dimensions, i.e., for two input features, and considering a linear decision boundary, $\hat{f}$ represents a line $\theta_1 x_1 = \theta_2 x_2$, see Figure 3a. The regression coefficients θ_j determine the general position and direction of $\hat{f}$ in space. Whereas, the ideal regression coefficients $\theta_{j,ideal}$ represent the best location and direction for $\hat{f}$ to separate the two classes.

The Decision Boundary

Equation (1) shows the general form of $\hat{f}$ for a first order linear model. Note that an "extra feature", i.e., the bias unit $x_0^i = 1$, is added to the feature set to accompany the intercept regression coefficient θ_0. Which brings the total number of regression coefficients and features to $n + 1$.

$$\hat{f} = 0 \iff \hat{\theta}_0 x_0 + \hat{\theta}_1 x_1 + \cdots + \hat{\theta}_n x_n = 0 \tag{1}$$

Once the regression coefficients θ_j are known, Equation (2) can be used to predict the location $\hat{z}^i$ of the ith training example relative to$\hat{f}$. Meaning that if $\hat{z}^i = 0$ the observation is situated on $\hat{f}$. Otherwise, the observation is a certain distance removed from $\hat{f}$.

$$\hat{z}^i \overset{?}{=} 0 \Leftrightarrow \hat{\theta}_0 x_0^i + \hat{\theta}_1 x_1^i + \cdots + \hat{\theta}_n x_n^i \overset{?}{=} 0 \tag{2}$$

Substituting the feature values of an observation from Table 1 in Equation (2) will result in an output value $\hat{z}^i \in \mathbb{R}$. As the output of interest is binary, i.e., flashover or no-flashover, the sigmoid function, Equation (3), is used to scale $\hat{z}^i$ to a value $0 < \hat{z}^i < 1$, as shown in Figure 3b. It should be noted that other functions exist which have the same effect, but the logistic function is preferred due to its traceability, interpretability and smoothness. The obtained value can be interpreted as a measure of confidence in the prediction, i.e., the probability $\hat{p}^i$ that for the ith training example the combination of x_j^i results in flashover $(y = 1)$. Samples situated on $\hat{f}$ return a value of $\hat{p}^i = 0.5$ and samples far removed from $\hat{f}$ return a value close to zero or close to one (signifying high confidence in the prediction).

$$\hat{p}^i = \frac{1}{1 + e^{-\hat{z}^i}} \tag{3}$$

In practice, the following boundary conditions are made in conjunction with Equation (3) to come to the actual predicted output for the ith training example $\hat{y}^i$, i.e., flashover or no-flashover.

$$\begin{cases} \hat{z}^i \geq 0 \text{ then } 0.5 \leq \hat{p}^i < 1 \text{ and } \hat{y}^i = 1 \\ \hat{z}^i < 0 \text{ then } 0 < \hat{p}^i < 0.5 \text{ and } \hat{y}^i = 0 \end{cases} \tag{4}$$

Cost Function for Unregularized Logistic Regression

The ideal values for the regression coefficients $\hat{\theta}_{j,ideal}$ are those which minimize the difference between f and $\hat{f}$, i.e., minimize the cost function. The cost function, used for this paper, is represented by Equation (5).

$$J(\hat{\theta}_j) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^i \log(\hat{p}^i) + (1 - y^i)\log(1 - \hat{p}^i)\right] \tag{5}$$

The right-hand part of the equation is usually referred to as the log likelihood (log lik) function, as shown in Equation (6).

$$J(\hat{\theta}_j) = -\frac{1}{m}\left[\log \text{lik}\left(\hat{\theta}_j\right)\right] \tag{6}$$

Equation (6) is usually minimized with mathematical and statistical programs, e.g., Matlab, Python, R, etc. that have a build-in optimization algorithm. The process of optimizing the cost function is commonly referred to as fitting the model.

Model Performance and the Deviance and R^2 Metric

In order to determine $\theta_{j,ideal}$ while still being able to report on the model performance, the historical data set is split into two parts: The training set and the test set. As such, the cost on the training set J_{train} can be calculated by replacing m in Equation (5) with the amount of observations allocated to the training set m_{train}. The training set is then used to fit the model and the test set to report on the ability of the fitted model to accurately predict flashover or no-flashover on unseen samples, i.e., the generalization error. The reason for using the test set is that, the data examples used to calculate J_{train} do not classify as unseen anymore and would thus give an optimistic approximation of the generalization error. The deviance on the test set, see Equation (7), is a metric which is commonly used to quantify the generalization error for logistic regression [11]. It denotes the difference between the fitted model and the ideal model, i.e., the saturated model. As such, the higher the deviance, the worse the performance of the model. It should be noted that, when D_{test} is evaluated over multiple lists a conservative approach is usually taken and the minimum D_{test} plus one standard error, $\min. D_{test} +$ 1SE, is considered to be the most parsimonious model [12], which in the remainder of the manuscript is referred to as the ideal scenario.

$$\begin{aligned} D_{test} &= -2\sum_{i=1}^{m_{test}}\left[y_{test}^{i}\log\left(\hat{p}_{test}^{i}\right) + \left(1-y_{test}^{i}\right)\log\left(1-\hat{p}_{test}^{i}\right)\right] \\ &\qquad + 2\sum_{i=1}^{m_{test}}\left[y_{test}^{i}\log\left(y_{test}^{i}\right) + \left(1-y_{test}^{i}\right)\log\left(1-y_{test}^{i}\right)\right] \\ &= -2\log \text{lik}\left(\hat{\theta}_j\right) \end{aligned} \tag{7}$$

The model without any features is referred to as the null-model, i.e., the worst model, and makes predictions solely with the intercept regression coefficient $\hat{\theta}_0$. The deviance of the null-model D_0 is calculated with Equation (9) and can be used as a benchmark for D_{test}. As D_0 and D_{test} might be difficult to interpret, especially due to the dependency on the amount of observations, they can be used to derive the R^2 value ($0 \leq R^2 \leq 1$). A value of unity ($R^2 = 1$) represents a perfect fit, while $R^2 = 0$ signifies a scenario where the features do not add anything to the regression.

$$R^2 = 1 - \frac{D_{test}}{D_0} \tag{8}$$

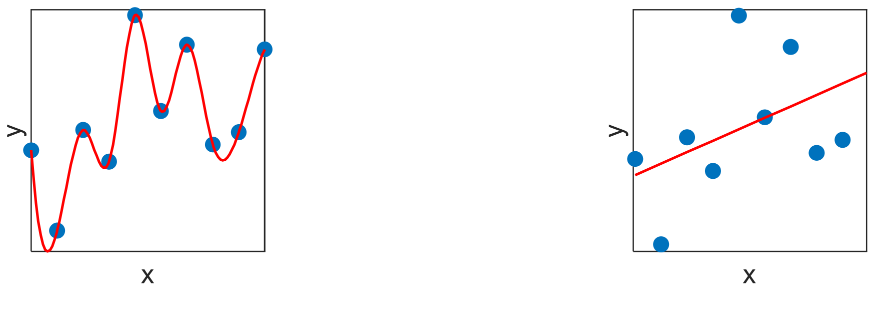
$$\text{with } D_0 = -2\log \text{lik}\left(\hat{\theta}_0\right) \tag{9}$$

In order to avoid an exceptionally good (or bad) allocation of observations the procedure is randomized. As such, The model was fitted and D_{test} was calculated as the average over a 1000 randomly generated training lists $m_{train} = 8$ and test lists $m_{test} = 6$. The values found for D_{test} and D_0 were respectively ≈ 13 and ≈ 10. It can thus be concluded that the fitted model performed worse than the null-model. The causes and possible solutions for this phenomenon are further elaborated in the next section.

Bias and Variance

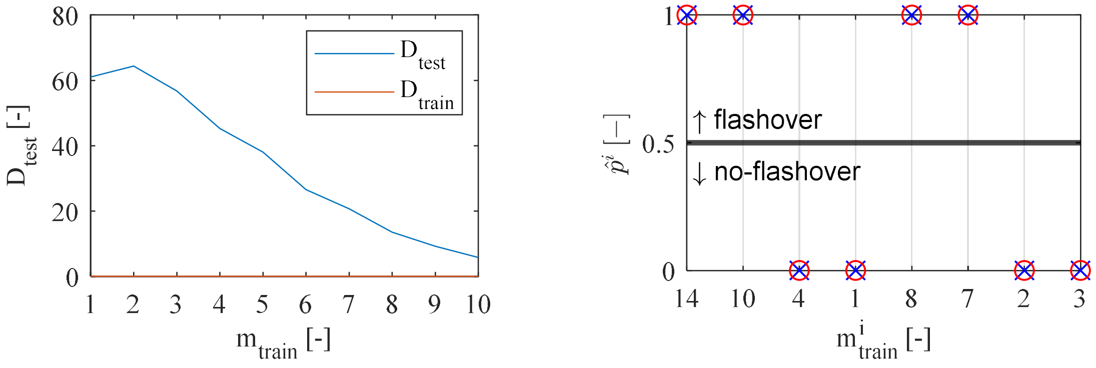
To improve the performance of the model the type of error is determined first. This is particularly important as the type will dictate the possible solutions. A high variance error signifies a $\hat{f}$ which is too flexible. As such, the model will find a pattern that is not actually true in the real world [8], see Figure 4a.

On the other hand, a model suffering from high bias will not be flexible enough to capture the intricacies of a training set, see Figure 4b.



a) *b)*

Figure 4: A schematic illustration of a decision boundary $\hat{f}$, red line, which represents a a) high variance/low bias and b) high bias/low variance scenario.



a) *b)*

Figure 5: a) The learning curves of the unpenalized model indicate a high variance/low bias error. b) The probability of flashover on the training set predicted with the unpenalized model coincides with the experimentally observed values and implies that the model suffers from high-dimensionality.

At this point it should be clear that a decrease in bias will inevitably mean an increase in variance and vice versa. As such, the ideal situation is a trade-off between the two types of error. A learning curve allows to assess whether a model suffers from high bias or high variance. In addition, learning curves are also a tool to determine the effect of an increasing training set size on the model performance. To construct the learning curve the amount of m_{train} was varied from one to ten while the amount of m_{test} was kept constant at four. For each new training list size, the model was fitted and D_{train} and D_{test} were calculated as the average over a 1000 randomly selected lists.

From Figure 5a it can be seen that D_{train} is approximately zero for every training set size. Whereas, the high value for D_{test} implies that the model fails to generalize to unseen samples. As such, there remains a large gap between D_{test} and D_{train}, which is typical for a high variance/low bias case. In addition, Figure 5b shows that the predicted probability of flashover $\hat{p}^i$, indicated by the red circles, perfectly matches the experimentally observed output y^i, indicated by the blue crosses, for every training examples m^i_{train} of one random training list. This is an indication that the model suffers from high-dimensionality, which in turn would explain the high variance/low bias error. As such, the next section will further elaborate the concept of high-dimensionality.

Considerations in High-Dimensionality

A high-dimensionality problem refers to the case where the amount of observations m is smaller, equal or only slightly higher, than the number of features n [12]. Table 1 shows that there are no more than five observations in the least prevalent class, i.e., no-flashover. Whereas some rules of thumb advise a minimum of 10-20 observations of the least prevalent class per feature considered [13]. According to this rule of thumb, to evaluate all five features, approximately 50-100 no-flashover observations would be needed. This suggests that the model is too complex for the recorded number of observations. The reason for the earlier mentioned high variance/low bias error can thus be attributed to the lack of observations relative to the number of features. As high dimensionality problems are becoming more and more frequent, mainly due to the vast feature collection possibilities of the internet [8], numerous solutions were developed, of which one is explored in the next section.

Cost Function for Regularized Logistic Regression

In order to solve the high-dimensionality problem the choice was made to apply subset selection, i.e., evaluating the effect of deleting certain features. For the model at hand approximately 32 (2^n) different subsets exist. As such, a shrinkage method was applied to avoid having to identify every possible subset and consequently run the model ≈ 32 times. Shrinkage effectively introduces a shrinkage penalty $P_\alpha(\hat{\theta})$ to the cost function applied to the training set, see Equation (10) [14]. For $\alpha = 0$, the regression coefficients of non-predictive features are reduced towards zero, i.e., ridge-regression. For $\alpha = 1$ the regression coefficients of non-predictive features are reduced to exactly zero, i.e., lasso regression. A value of $0 < \alpha < 1$ represents an elastic-net regression, which can be seen as a trade-off between ridge-regression and lasso-regression. The reason for evaluating α is that, it is difficult to know a priori which regression method will perform best. Lasso-regression will outperform ridge-regression when only a few features are related to the response and vice versa. The tuning parameter λ controls the trade-off between the log-likelihood function and the shrinkage penalty. For $\lambda \rightarrow \infty$, all coefficients will be near or exactly zero, which defines the null-model. For $\lambda \rightarrow 0$ the effect of the shrinkage penalty becomes negligible and the cost function is again approximated by Equation (5).

$$J(\hat{\theta}_j) = -\frac{1}{m}\sum_{i=1}^{m}\left[\log\text{lik}\,(\hat{\theta}_j)\right] + \lambda P_\alpha(\hat{\theta}_j)$$
$$\text{where } P_\alpha(\hat{\theta}_j) = \sum_{j=1}^{n}\left[\frac{1}{2}(1-\alpha)\hat{\theta}_j^2 + \alpha\left|\hat{\theta}_j\right|\right] \qquad (10)$$

It is advised to use the standardise features $\tilde{x}_j^i = x_j^i / \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(x_j^i - \bar{x}_j^i\right)^2}$ in combination with shrinkage to make sure all inputs have a standard deviation of one [15]. As such, the magnitude of the regression coefficients will only be affected by the size of λ and not by the scaling differences between the features.

It should be noted that, applying subset selection on a limited historical dataset could result in the deletion of information that might be relevant. Therefore, a preference exists to increase the number of observations in order to resolve high dimensionality. Although this is not allows possible, the fire safety community should strive towards an easily accessible databases in which experimental results are compiled. Recent steps towards this goal were undertaken by Naser [7], who compiled a library of 12,000 data-points for fire-tested timber members.

Cross-Validation

The introduction of the hyperparameters α and λ gives rise to another problem. Namely that, for every possible combination of α and λ the model must first be fitted on the training set. After which, the best model can be chosen as the one that minimizes D_{test}. As such, the test set cannot be used anymore to calculate the generalization error because the test observations do not classify as truly unseen anymore, i.e., they were used to establish the ideal hyperparameter combination. In order to fit the model, determine the ideal hyperparameters and be able to calculate the generalization error, the historical data set must be split into three parts. The training set to fit the model, the cross-validation (CV) set to determine the ideal hyperparameters, and the test set to calculate the generalization error. Unfortunately, the available data set is not large enough to be split in three ways while still allowing enough data for training and cross-validation. For this reason, it was decided to only split the data into a training and CV set and use D_{cv} as an approximation of D_{test}. In contrast to the training set, the CV set was only used to establish the hyperparameters. As such the model never truly 'learns' from them and thus D_{cv} will be a better approximation of D_{test} than D_{train}.

In addition, Leave-One-Out Cross-Validation (LOOCV) was applied [15], rather than randomly assigning observations to different lists. For LOOCV, the model is fitted on $m-1$ data examples and the remaining i^{th} data example is used to calculate the cross-validation deviance $D_{cv,i}$. The process is then repeated m times, with for every run a different data example to be used as CV. The total D_{cv} is calculated as the average over m observations [15]. The advantage of LOOCV is the absence of randomness in allocating observations to sub sets and the possibility to fit the model on almost the complete data set. With the LOOCV method the null-model deviance was found to be ≈ 21, which will be used as a benchmark in the following section.

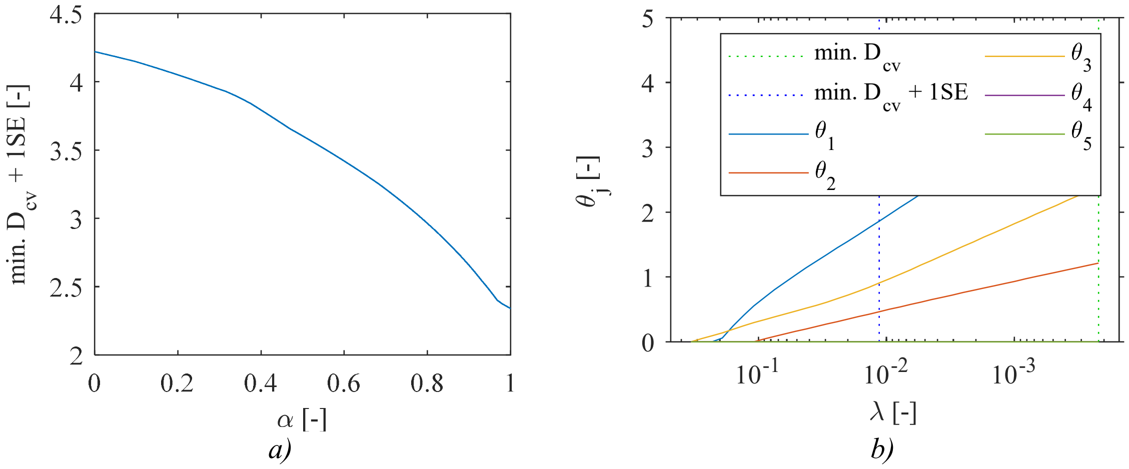


Figure 6: The model was evaluated for different values of α and λ and it was found that a) lasso-regression ($\alpha = 1$) in combination with b) $\lambda \approx 0.01$ resulted in the best performance. For which b) lasso-regression effectively reduces the regression coefficients θ_{4-5} to zero.

RESULTS AND DISCUSSION

The model was fitted with the LOOCV method by minimizing Equation (10) for different combinations of α and λ. The ideal scenario for every α is depicted by Figure 6a, from which it can be seen that a ridge-regression model ($\alpha = 1$) gives the best performance. Figure 6b shows that for $\alpha = 1$ the most parsimonious model is obtained with $\lambda \approx 0.01$, which effectively reduces the regression coefficients θ_{4-5} to zero, i.e., lasso-regression considers the features x_{4-5} not relative for the prediction of flashover or no-flashover. As such the model is reduced from five degrees of freedom (df) to three. With the lasso-regression model it was possible to obtain a model performance of min. $D_{cv} + 1\,SE \approx 2$, or a $R^2 \approx 0.91$. Which is a considerable improvement compared to the null-model with $D_0 \approx 21$.

A careful interpretation of the subset selection is necessary as the machine learning algorithm does not know the principles of fire safety engineering, and thus solely makes conclusions based on the data it is presented with. In particular, by implementing the shrinkage parameter it is implicitly assumed that the emphasis of the model is directed towards making predictions on unseen samples based upon the current historical data set, and not so much on explaining the underlying correlations between the variables of the historical dataset itself. The difference can be found in the fact that for explanatory modelling the goal is to reduce the bias as much as possible, i.e., make accurate predictions on the training set. Whereas for predictive modelling the objective is to reduce both bias and variance, for which it might be necessary to sacrifice some theoretical accuracy. The latter was accurately described by Shmueli as: "To explain or to predict" [16].

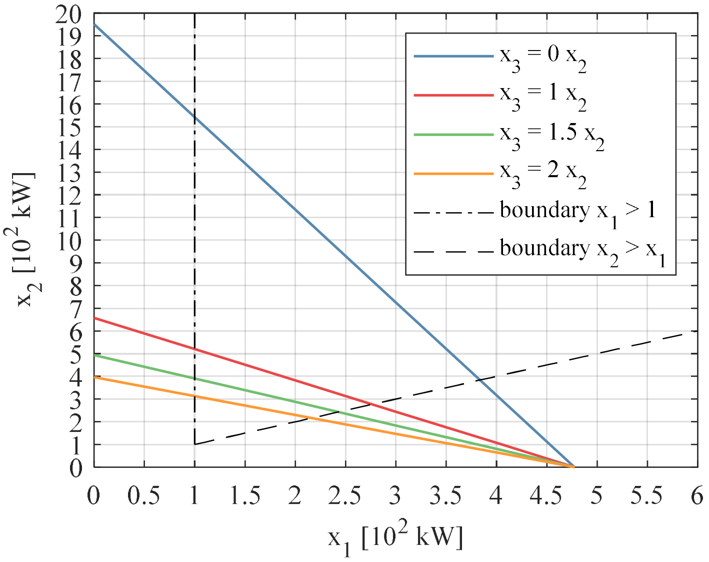


Figure 7: The model was fit on the complete historical dataset to determine the final definition of the regression coefficients. Different assumptions were made for burning intensity three x_3. The resulting curves divide the space into a flashover, above the coloured curve, and a no-flashover zone, below the coloured curve.

To obtain the final definition of the regression coefficients the complete historical dataset was used for training. The 3D decision boundary flashover/no-flashover corresponding with the final regression coefficients is visualized in the 2D plot of Figure 7 for various magnitudes of the third burning intensity x_3. For these definitions of x_3, the values for x_1 and x_2 resulting in a probability of flashover of 0.5 were iteratively calculated with Equation (3). As such, each coloured line of Figure 7 divides the space into a flashover zone, situated above the line and denoting a $\hat{p}^i > 0.5$, and a no-flashover zone, situated below the line and denoting a $\hat{p}^i < 0.5$. Due to the limited amount of observations any extrapolations which does not comply with the following conditions should be treated with caution: $x_1^i > 100$ kW, $x_2^i > x_1^i$, $6 < x_4^i < 10$ and $x_5^i > 465$. Figure 7 can be used to define future experiments which would result in the greatest knowledge benefit for the ongoing research and for the model, i.e., updating the model with testing conditions close to the decision boundaries will result in an increase of confidence in a region where the model currently has a low measure of confidence.

The explanatory capacities of the model were also researched by evaluating the effect of an increasing/decreasing thickness and burning time on the probability of flashover for an unpenalized model, i.e., a model that perfectly predicts the training set. It was found that more stringent conditions were predicted to avoid flashover for an increasing x_{4-5}. The latter seems to be in correspondence with the literature on the subject: Equation (11) represents the general heat conduction equation when heat losses through convection and radiation are ignored [17].

$$\dot{q}'' = h(T_g - T_a) \tag{11}$$

Where T_a is the ambient temperature, T_g is the gas temperature and the heat transfer coefficient h is defined as $\sqrt{k\rho c/\pi t}$ for transient conditions and as k/δ for steady-state conditions. As such, an increase of t or δ will result in a decrease of the conduction losses $\dot{q}''$, and thus make flashover more likely.

CONCLUSION

As the SBI test cannot guarantee an accurate classification for sandwich panels the dependency on the full-scale RCT remains. As the latter is nor time nor cost efficient the need arises to derive a specific intermediate-scale test. Recent work by Leisted et al. [10] showed promising results based upon a constant Froude number and a 1:5 scale model of the RCT. The methodology derived in this study demonstrates how ML can assists such research by providing an easy-to-use tool that can determine the relevant parameters and derive experimental tests which maximize the knowledge benefit.

The model showed a considerable improvement compared to the null-model. This indicates that the model is ready to be applied on unseen samples as it was able to make accurate predictions for the cross-validation set. As such, the model can be used to predict the occurrence of flashover with high confidence. On the other hand, the model can be used to identify critical combinations of burner intensities, i.e., the decision boundary. Nevertheless, the limited data set inevitably means that the algorithm cannot capture all the physics, as it can only learn from the data it is presented with. As such, future predictions should be used in combination with engineering judgement and within the boundaries prescribed by the historical data set.

With the LOOCV and lasso-regression method the ideal hyperparameters were determined. As a result, two of the five features were found to be non-predictive, with respect to the given historical data set. By letting the algorithm determine which features are relevant for the response it was shown how machine learning can be used to achieve a deeper understanding of the contributing factors that induce flashover. It can be argued that for the small data set used here the same can be achieved by simple reasoning. Nevertheless, identical principles, and for that matter the same algorithm, can be used for a database that is several magnitudes larger.

It was demonstrates how ML can be used in ongoing experimental research to define experiments which result in the greatest knowledge benefit. Furthermore, the presented model tests an exposure beyond what is prescribed in the ISO standard for the RCT. As such weaknesses of specimens, which might be overlooked in the original configuration, can be determined and used to provide guidance for future large-scale and intermediate-scale testing.

The vast amount of data needed to learn an algorithm is a nuisance to overcome. In particular for FSE, where the destructive nature of experiments is inherently associated with considerable costs. To make the application of ML viable, the currently available test results need to be compiled in databases and new fire tests must be defined for their knowledge benefit. A common database will not only allow learners to learn but will also provide the possibility to accurately compare the performance of different algorithms.

Contemporary classification methods, e.g., Euroclasses, subdivide the space in a discrete number of intervals which makes it impossible to determine how good, or how bad, a material effectively is within its own class. The predicted probability of flashover $\hat{p}^i$ for every CV observation could be used to establish a more nuanced ranking of various materials and configurations in the form of a continuous number. It should be noted though that, for it to be possible to build a ranking system based upon $\hat{p}^i$ the confidence in the model must be large, i.e., $\hat{p}^i$ values around 0.5 are due to the specimen or configuration and not due to a badly fitted model. The latter could not be demonstrated for the derived ML algorithm, as the lack of data made it impossible to calculate D_{test}.

The flexibility of ML algorithms is unmatched in current models. As such, it might proof to be the solution for an ever-changing application of innovative materials and design solutions. Furthermore, it is foreseen that when enough data becomes available the ML algorithm will be able to produce more accurate results than contemporary models.

In a future phase it is envisioned to incorporate data from small scale testing (e.g., cone calorimeter) to make predictions for intermediate-scale and large-scale tests (the RCT and SBI test). In addition, it is foreseen that the integration of the time dependent small-scale HRR curves will make it possible to accurately predict the large scale HRR for a variety of materials, designs and scenarios.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of the BRE Trust for this project.

REFERENCES

[1] J. L. Torero, Scaling-Up fire, *Proc. Combust. Inst.*, 2013, **34**, 99–124.

[2] British Standard Institution, *Draft BS ISO 9705-1:2016--Reaction to fire tests -- Room corner test for wall and ceiling lining products -- Part 1: Test method for a small room configuration*, 2016.

[3] B. Messerschmidt, The Capabilites and Limitations of the Single Burning Item (SBI) test, *FireSeat*, University of Edinburgh, United Kingdom, 2008, 70–81.

[4] U. Wickstrom and U. Goransson, Full-scale/Bench-Scale correlations of wall and ceiling linings, *Fire Mater.*, 1992, **16**, 15–22.

[5] A. S. Hansen and P. J. Hovde, Prediction of time to flashover in the ISO 9705 room corner test based on cone calorimeter test results, *Fire Mater.*, 2002, **26**, 77–86.

[6] B. Ostman and L. D. Tsantaridis, Correlation between cone calorimeter data and time to flashover in the room fire test, *Fire Mater.*, 1994, **18**, 205–209.

[7] M. Z. Naser, Fire resistance evaluation through artificial intelligence - A case for timber structures, *Fire Saf. J.*, 2019, **105**, 1–18.

[8] P. Domingos, *The master algorithm : how the quest for the ultimate learning machine will remake our world*, Basic Books, 1st edn., 2015.

[9] R. R. Leisted, M. X. Sørensen and G. Jomaas, Experimental study on the influence of different thermal insulation materials on the fire dynamics in a reduced-scale enclosure, *Fire Saf. J.*, 2017, **93**, 114–125.

[10] R. R. Leisted, The Fire Performance of Steel-faced Insulation Panels with Stone Wool or Polymer Cores (Unpublished Doctoral Dissertation), Technical University of Denmark, 2018.

[11] J. A. Dobson, *An Introduction To Generalized Linear Models*, CRC Press Company, 2002.

[12] T. Hastie, R. Tibshirani and J. Friedman, *The Elements of Statistical Learning*, 2nd edn., 2017.

[13] F. E. J. Harrell, *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Springer US, 2nd edn., 2015.

[14] J. Friedman, T. Hastie and R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent, *J. Stat. Softw.*, 2010, **33**, 1–22.

[15] G. James, D. Witten, T. Hastie and R. Tibshirani, *An Introduction to Statistical Learning*, 2017.

[16] G. Shmueli, To Explain or to Predict?, *Stat. Sci.*, 2010, **25**, 289–310.

[17] B. Karlsson and J. Quintiere, *Enclosure Fire Dynamics*, CRC Press, 2000.