Testing for knowledge: Application of machine learning techniques for prediction of flashover in a 1/5 scale ISO 137841 enclosure

Citation for published version:
Dexters, A, Leisted, RR, Van Coile, R, Welch, S & Jomaas, G 2020, 'Testing for knowledge: Application of machine learning techniques for prediction of flashover in a 1/5 scale ISO 137841 enclosure', *Fire and Materials*. https://doi.org/10.1002/fam.2876

Digital Object Identifier (DOI):
10.1002/fam.2876

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Fire and Materials

TESTING FOR KNOWLEDGE: APPLICATION OF MACHINE LEARNING TECHNIQUES FOR PREDICTION OF FLASHOVER IN A 1/5 SCALE ISO 13784-1 ENCLOSURE

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

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

Ruben Van Coile, Department of Structural Engineering and Building Materials, 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 algorithm was applied to predict the onset of flashover in archival experiments in a 1/5 scale ISO 13784-1 enclosure constructed with sandwich panels. The experiments were performed to assess whether a small-scale model could provide a better full-scale correlation than the single burning item test.

To predict the binary output a logistic regression model was chosen as machine learning environment. Because results indicated a high variance/low bias issue regularization was applied. It was found that lasso-regression significantly reduced the amount of variance at a negligible increase in bias.

With the regularized model, it was possible to discern the predictive variables and determine the decision boundary. In addition, a methodology was put forward on how to use the decision boundary to iteratively update the learning algorithm. As a result, it was shown how a learning algorithm can be used to facilitate ongoing experimentation. At first as a crude guideline, and in later stages, as an accurate prediction algorithm.

It is foreseen that, by iteratively updating the algorithm, by compiling existing and new experiments in databases, and by applying fire safety knowledge, the final learned algorithm will be able to make accurate predictions for unseen samples and test conditions.

KEYWORDS

Machine learning, fire tests, sandwich panels, fire classification, flashover

LIST OF SYMBOLS AND ABREVIATIONS

LOWERCASE

f	True decision boundary
$\hat{f}$	Estimated decision boundary

i	Observation number i from the historical dataset
l_m	The model-scale enclosure length [m]
l_f	The full-scale enclosure length [m]
m_{test}^i	Observation i from the test set
m_{train}^i	Observation i from the training set
$\hat{p}^i$	The predicted probability of flashover for observation i
$\hat{p}^{i*}$	The predicted probability of flashover for an experiment i^* which has not been performed yet
t_{fo}	Time to flashover
t_M	The model-scale burner time step [s]
t_F	The full-scale burner time step [s]
x_j^i	Input value for predictor j and for observation i [kW; m; s]
x_j^{i*}	Input value for predictor j and for an experiment i^* which has not been performed yet [kW; m; s]
$\boldsymbol{x}^i$	$[(N+1) \times 1]$ column vector containing all the values of the features for observation i, including the bias unit [kW; m; s]
$\boldsymbol{x}^{i*}$	$[(N+1) \times 1]$ column vector with feature values for an experiment i^* which has not been performed yet [kW; m; s]
$\boldsymbol{x}_{DB}^{i*}$	$[(N+1) \times 1]$ column vector with feature values that are an element of the decision boundary for an experiment $i*$ which has not been performed yet [kW; m; s]
$\bar{x}_j$	Mean of input variable j [kW; m; s]
y^i	The observed output value for observation i
$\hat{y}^i$	The predicted output value for observation i
$\hat{z}^i$	The location of the i^{th} observation relative to $\hat{f}$

UPPERCASE

D_0	Deviance of the null model
D_{cv}	Deviance on the cross-validation set
D_{test}	Deviance on the test set, i.e. the approximation of the generalization error
$J(\hat{\boldsymbol{\theta}})$	Unregularized cost function
$J(\hat{\boldsymbol{\theta}})_{test}$	Unregularized cost function on test set
$J(\hat{\boldsymbol{\theta}})_{train}$	Unregularized cost function on training set
$J_\lambda(\hat{\boldsymbol{\theta}})$	Regularized cost function
M	Total amount of observations
M_{test}	Total amount of observations allocated to the test set
M_{train}	Total amount of observations allocated to the training set
N	Total amount of input variables
$P_\alpha(\hat{\boldsymbol{\theta}})$	Shrinkage penalty including lasso, ridge and elastic-net regression
$\dot{Q}_M$	The model-scale gas burner heat release rate [kW]
$\dot{Q}_F$	The full-scale gas burner heat release rate [kW]
$\boldsymbol{X}$	$[M \times (N+1)]$ matrix with all observations, i.e. with on every row the transpose of $\boldsymbol{x^i}$ [kW; m; s]
X_0	Bias unit with value $x_0^i = 1$
X_j	Input variable j, also referred to as predictor j or feature j [kW; m; s]
Y	The variable on which the prediction is to be made, i.e. the output variable

GREEK

α	Tuning, or hyper, parameter II
λ	Tuning, or hyper, parameter I
θ_j	True value for the regression coefficient j

$\hat{\boldsymbol{\theta}}$	$[(N+1)\times 1]$ vector containing the values of the regression coefficients, including the intercept $\hat{\theta}_0$
$\hat{\theta}_0$	Estimated intercept regression coefficient
$\hat{\theta}_j$	Estimated value for the regression coefficient j

ABBREVIATIONS

CV	Cross-validation	ML	Machine learning
HRR	Heat release rate	RCT	Room corner test
LogLik	Log likelihood function	SBI	Single burning item test
LOOCV	Leave-one-out cross-validation	SE	Standard error

INTRODUCTION

Fire-classification of materials is a central element for ensuring safe building design. The classification of a product should in principle be arrived at based on its reaction to fire in a test that represents the end-use situation (often full-scale). However, as large-scale tests are often costly and labour intensive, a tendency exists to try to predict full-scale fire behaviour based on small-scale testing[1–6]. In order to justify such a scaling methodology, a thorough understanding of the fire behaviour is necessary[7]. While this is currently the case for many single burning items, a knowledge gap persists for the interaction of a growing fire and combustible linings in an enclosure. Therefore, large-scale testing is still needed to accurately classify such materials.

The full-scale room corner test (RCT)[8] 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)[9] test. The concept of scaling-down seems justified, as in 87 per cent of the cases the full-scale fire growth behaviour could be captured adequately by the intermediate-scale test[10]. Nevertheless, for materials such as sandwich panels, linear systems (e.g. cabling and piping) and polycarbonate panels, the correlation proved to be less accurate[10]. New tests have been developed for some specimens, but the defining test for classification of sandwich panels remained with the SBI test. As such, it is foreseen that unwanted situations could arise due to the possible misclassification of sandwich panels. The risk is further magnified as sandwich panels are frequently used as free-standing or frame-mounted construction elements rather than as linings. Free-standing or frame-mounded sandwich panels should be tested conform the large-scale ISO 13784-1[11] standard to represent the end-use situation. Therefore, the correlation between the SBI test and the ISO 13784-1 can be further questioned[12].

For the aforementioned reasons, the dependency on large-scale testing remains for an accurate classification of sandwich panels[13]. Therefore, a new reduced scale test is needed to provide the industry with an accurate and time- and cost-effective method for quality control and product development. To this end, recent work by Yoshioka et al.[14] and Leisted[15] studied the correlation between various scale-models of the ISO 13784-1 and the full-scale compartment. In particular, Yoshioka et al.[14] researched the predictive capabilities of a 1/3 scale-model and Leisted[15] researched the application of both a 1/2 and 1/5 scale-model. The work by Leisted[15] is especially relevant for this paper as it has highlighted that a tool which can both identify experimental configurations for their knowledge benefit, and discern relevant parameters, would significantly reduce research time and cost, and therefore also augment the possibility of a successful research outcome.

One possibility to develop a tool which can aid ongoing experimentation 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. Meaning that, a relatively simple learning algorithm can prompt different learned algorithms, which can be magnitudes more complex, for varying datasets and without the interference of the ML expert. The simplicity inherently means that the learning algorithm can analyse complex problems and large amounts of variables, whereas conventional techniques quickly get overwhelmed, either by the limitations of computational-time or computational-space or because they are too complex to be understood by humans.

The currently proposed learning algorithm uses ML techniques to aid ongoing experimentation to derive a new intermediate-scale test procedure for sandwich panels with regards to the ISO 13784-1 standard. The regularized logistic regression model predicts flashover or no-flashover for a polyisocyanurate (PIR) sandwich panel exposed to various burning intensities within the physical confines of a 1/5 scale model of the standardised ISO 13784-1 enclosure.

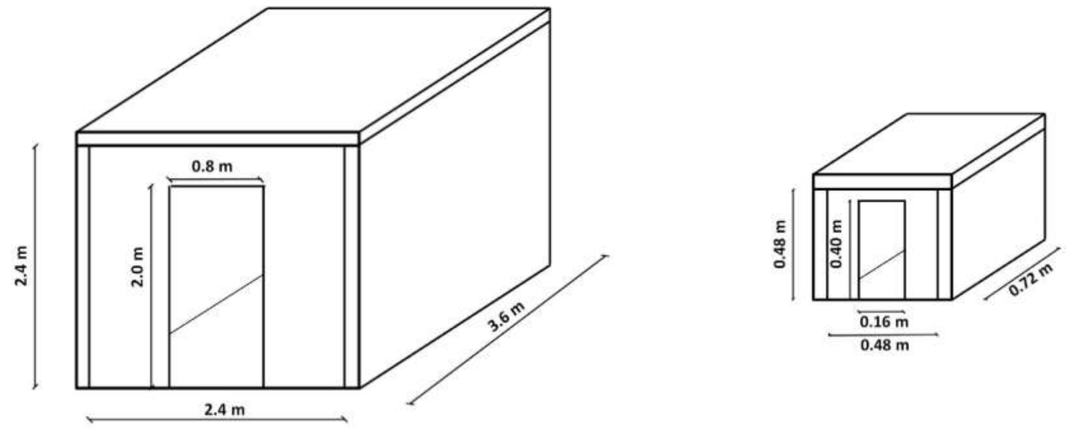
EXPERIMENTAL SETUP

The experimental data used for the ML algorithm is a selected version of the complete dataset that was performed by Leisted[15] in an attempt to develop a screening method for the ISO 13784-1 enclosure. In particular, Leisted[15] researched whether a scale model of the ISO 13484-1 would provide a better correlation than the SBI test, with respect to sandwich panels. Towards this end, different experiments were conducted at 1/2 and 1/5 scale, but because only a few 1/2 scale experiments were available, the remainder of this paper will focus on the 1/5 scale experiments.

Geometrical scaling of the compartment and the gas burner

Figure 1a shows the geometry of the full-scale ISO 13784-1 enclosure and Figure 1b shows the 1/5 scaled enclosure. The full-scale prescribed compartment dimensions were 3.60 m x 2.40 m x 2.40 m (L x W x H) with a door opening of 0.80 m x 2.00 m (W x H), whereas the 1/5 scale enclosure and door dimensions were scaled to respectively 0.72 m x 0.48 m x 0.48 m (L x W x H) and 0.16 m x 0.40 m (W x H). The walls and ceiling of both enclosures were made of a commercially available sandwich-panel with PIR foam core, for which the reference, i.e. full-scale, experiments had a material thickness of 0.10 m and the scaled experiments had a material thickness of either 0.06 or 0.10 m.

The dimensions of the gas burner, and the height of the top of the gas burner to the inert floor, were also scaled with a 1/5 factor. That is, the prescribed ISO 13784-1 burner of 170 mm x 170 mm x 200 m (L x W x H) was scaled to 34 mm x 34 mm x 40 mm (L x W x H). To ensure an even gas flow, the internal compartment height of the burner was increased from 40 mm to 100 mm and the excess burner height was allowed to protrude trough the inert floor.



a) *b)*

Figure 1: a) Internal dimensions of the full-scale enclosure. b) Internal dimensions of the 1/5 scale enclosure. Reused with permission from Leisted[15]*.*

Froude scaling of the gas burner HRR and gas burner duration

To ensure that a correlation would exists across the scales, the Froude scaling technique was used to scale down the size of the fire with respect to the geometric scaling of the enclosure[15]. In particular, the gas burner HRR and the gas burner duration were scaled with respectively Equation (1)[16] and Equation (2)[16].

$$\frac{\dot{Q}_M}{\dot{Q}_F} = \left(\frac{l_M}{l_F}\right)^{5/2} \tag{1}$$

$$\frac{t_M}{t_F} = \left(\frac{l_M}{l_F}\right)^{1/2} \tag{2}$$

Here $\dot{Q}_M$ is the model-scale HRR, $\dot{Q}_F$ is the full-scale burner HRR, t_M is the model-scale burner timestep, t_F is the full-scale burner timestep, l_F is the full-scale length and l_M is the model-scale length. It should be noted that, for Equation (1) l_M and l_F refer to the compartment length, whereas for Equation (2) l_M and l_F refer to the material thickness.

The aforementioned equations were applied to the ISO 13784-1 standard, which specifies a gas burner HRR equal to 100 kW for the first 10 min, 300 kW for the next 10 min, and 0 kW for the last 10 min, i.e. the observation period. The end of the test is either the 30 min mark or the onset of flashover, whichever occurs first. Equation (1) was applied to the prescribed ISO 13784-1 gas burner HRR, which resulted in a scaled gas burner HRR of 1.79 kW during the first time step, and 5.37 kW during the second time step. Furthermore, the duration of each burner intensity was determined with Equation (2) to be either 465 s or 600 s for insulation thicknesses of 0.06 m and 0.10 m, respectively. Lastly, research[17,18] showed that fires in commercial premises were often much higher than the prescribed 300 kW and had a longer duration than 30 min. Therefore, Leisted[15] added two gas burner regimes with a third burner intensity and two regimes with a continuous burner output to allow for more (severe) testing conditions. The burner regimes were denoted with three subsequent numbers, one for every time step, that can take the following values: 0 for 0 kW, 1 for 1.79 kW, 5 for 5.37 kW and 10 for 10.74 kW. The scaled stepwise burner regimes are depicted in Figure 2a and the scaled continuous burner regimes in Figure 2b.

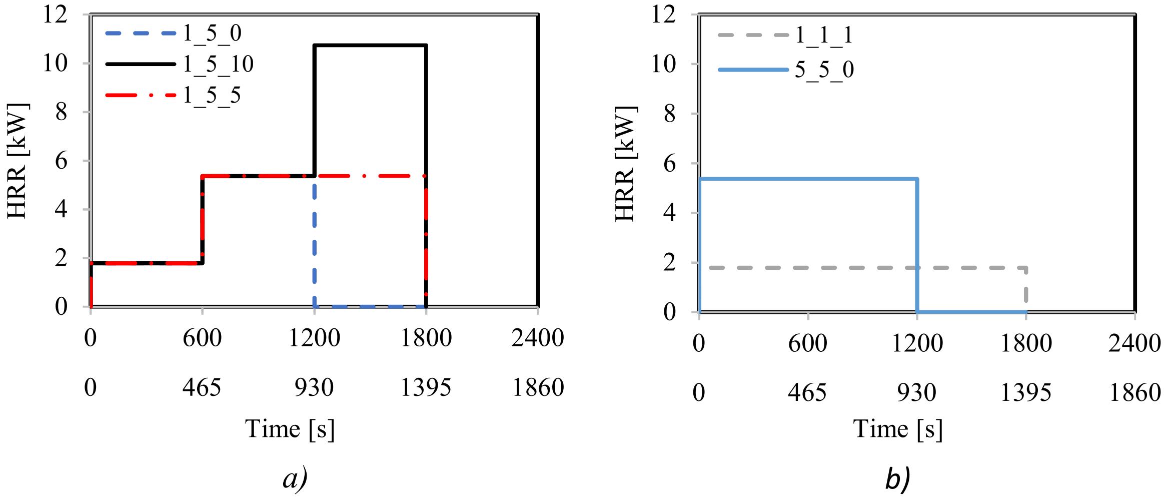


Figure 2: a) The stepwise burner regimes and b) The continuous burning regimes for the 0.10 m material thickness (timestep of 600 s) and the 0.06 m material thickness (timestep of 465 s).

Experimental data from the 1/5 scale experiments

The sandwich panels were exposed in the 1/5 scale model to the aforementioned scaled burner regimes and the specimen HRR was recorded with the oxygen consumption theory. Figure 3a and b depicts the specimen HRR for respectively the 0.10 m and 0.06 m thick material.

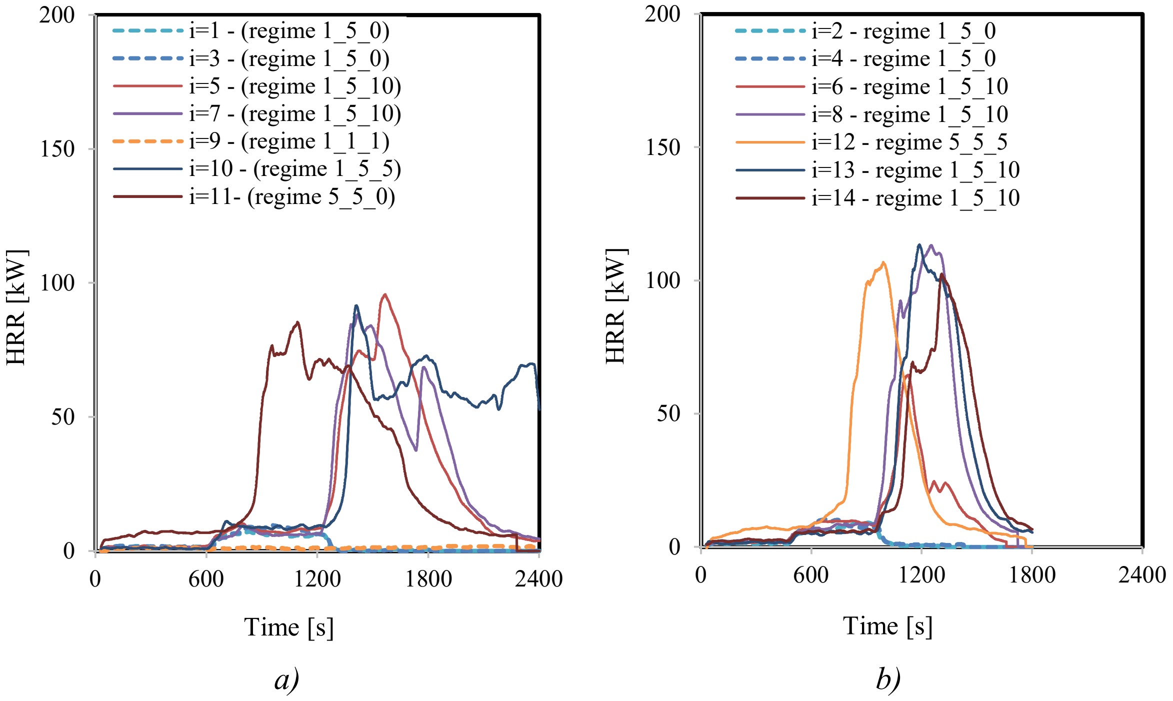
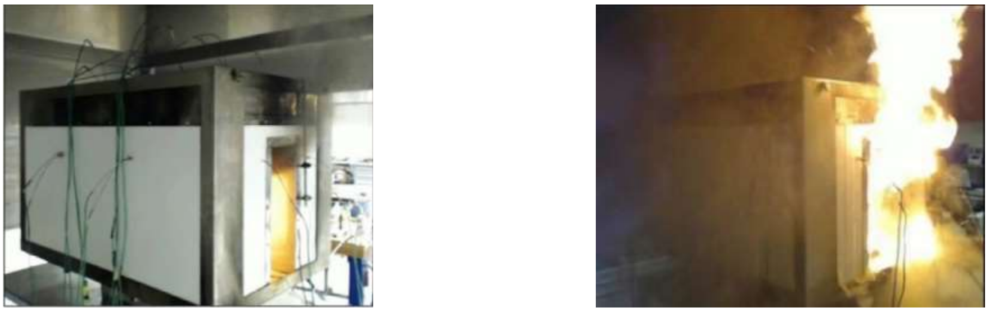


Figure 3: The specimen HRR profile when exposed to different gas burner regimes for a) The 0.10 m thick sandwich panel and b) The 0.06 m thick sandwich panel.

The dashed lines indicate those tests which did not lead to flashover, i.e. test number 1, 3 and 9 for the 0.10 m thick material and test number 2 and 4 for the 0.06 m thick material. 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. It should be noted that, on the onset of flashover the burner was turned off regardless of the predefined burner regime. Figure 4a and b show graphs from a 1/5 scaled experiment with no-flashover and with flashover, respectively.



a) *b)*

Figure 4: a) 1/5 scale experiment resulting in no-flashover. b) 1/5 scale experiment resulting in flashover. Reused with permission from Leisted[15].

THE MACHINE LEARNING ENVIRONMENT

The 14 experimental observations ($M = 14$) depicted in Figure 3 are a selected subset of all the 1/5 scale experiments performed by Leisted[15]. 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 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.

The experimental observations used for the ML algorithm are summarized in Appendix Table 1, which in the remainder of the manuscript is referred to as the historical dataset. The time to flashover t_{fo} is listed as an informative feature for the reader but will not be used in the ML model as it is not an a priori known variable. Capital letters are used to denote the output variable Y and the five $(N = 5)$ input variables X_j $(1 < j < N)$. The values of the variables are denoted with lowercase letters x_j^i and y^i for every i^{th} observation $(1 \leq i \leq M)$ and j^{th} input variable. The following list summarizes the input and output variables together with the boundaries defined by the historical dataset:

- The output variable Y, with $y^i \in \{0,1\}$ for respectively flashover and no-flashover.
- Three burning intensities X_{1-3}, with $x_{1-3}^i \in \{0\text{ kW};\ 1.79\text{ kW}; 5.37\text{ kW}; 10.74\text{ kW}\}$.
- The thickness of the material X_4, with $x_4^i \in \{0.06\text{ m}, 0.10\text{ m}\}$.
- The planned duration of each burning intensity X_5, with $x_5^i \in \{465\text{ s}, 600\text{ s}\}$.

Many different ML algorithms can be applied to predict a binary output, i.e. flashover or no-flashover, of which most are referred to as a black box, meaning that it is practically impossible to fully comprehend all the implications of what happens in the in-between state of giving the algorithm the data and the algorithm coming up with a prediction. Therefore, the choice was made to use a logistic regression model, which is by far the most transparent trained ML algorithm. In other words, after performing the regression all input can be clearly ranked in terms of importance and all input can be measured in terms of effect on the output.

The logistic regression model divides the five-dimensional space, defined by the input features X_j, into a flashover and no-flashover zone with a so-called estimated decision boundary (DB) $\hat{f}$. By doing this, it is implicitly assumed that a true division f exists, which corresponds with test setups i for which input vector $\boldsymbol{x}^i = \{x_1^i; x_2^i; \ldots; x_j^i\}$ results in a 50 per cent flashover chance. As f is almost always unknown, the goal of the ML algorithm is to make an approximation $\hat{f}$ of f. In two dimensions, i.e. for two input features, and considering a linear DB, $\hat{f}$ represents a line $\hat{\theta}_0 + \hat{\theta}_1 X_1 = \hat{\theta}_2 X_2$, see Figure 5. The estimated regression coefficients $\hat{\theta}_j$ and the estimated intercept regression coefficient $\hat{\theta}_0$ determine the position and direction of $\hat{f}$ in space and are an approximation of the ideal regression coefficients θ_j, which in turn represent the location and direction of f in space.

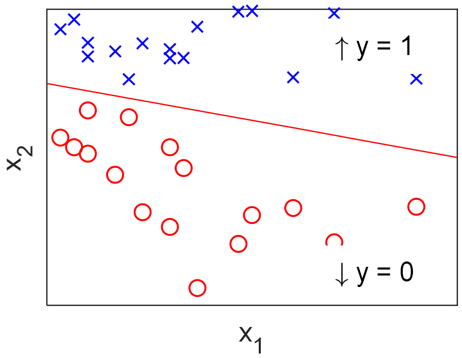


Figure 5: Graphical interpretation of the decision boundary $\hat{f}$, represented by the red line, which separates flashover $(y^i = 1)$ *from no-flashover* $(y^i = 0)$.

The Decision Boundary

Equation (3) 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 with $x_0^i = 1$, is added to the feature set to accompany the estimated intercept regression coefficient $\hat{\theta}_0$, which allows the use of matrix notations.

$$\hat{f} = 0 \Leftrightarrow \hat{\theta}_0 X_0 + \hat{\theta}_1 X_1 + \cdots + \hat{\theta}_N X_N = 0 \tag{3}$$

Once the regression coefficients are estimated, Equation (4) can be used to predict the location $\hat{z}^i$ of the i^{th} training example relative to$\hat{f}$. Meaning that if $\hat{z}^i = 0$ the observation is situated on $\hat{f}$. Otherwise, the observation is situated either above or below $\hat{f}$.

$$\begin{aligned} \hat{z}^i \overset{?}{=} 0 &\Leftrightarrow \hat{\theta}_0 x_0^i + \hat{\theta}_1 x_1^i + \cdots + \hat{\theta}_N x_N^M \overset{?}{=} 0 \\ &\Leftrightarrow \boldsymbol{X}\hat{\boldsymbol{\theta}} \overset{?}{=} 0 \end{aligned} \tag{4}$$

Substituting the feature values of an observation from Appendix Table 1 in Equation (4) 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 (5), is used to scale $\hat{z}^i$ to a value $0 < \hat{z}^i < 1$, as shown in Figure 6.

$$\hat{p}^i = \frac{1}{1 + e^{-\hat{z}^i}} \tag{5}$$

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[19]. The obtained value can be interpreted as a measure of confidence in the prediction, i.e. the probability $\hat{p}^i$ that for the i^{th} training example the input vector $\boldsymbol{x}^i$ results in flashover ($y^i = 1$). Observations situated on $\hat{f}$ return a value of $\hat{p}^i = 0.5$ and observations far removed from $\hat{f}$ return a value close to zero or close to one (signifying high confidence in the prediction).

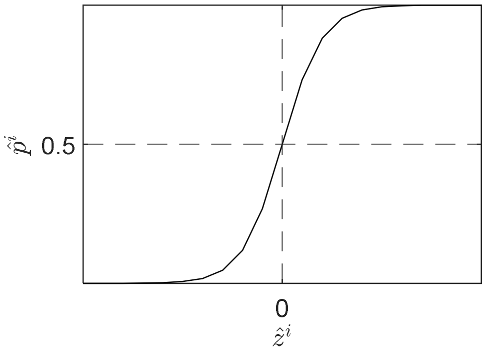


Figure 6: Graphical interpretation of the sigmoid function.

In practice, the following interpretations are made in conjunction with Equation (5) to come to the actual predicted output for the i^{th} training example $\hat{y}^i$, i.e. flashover or no-flashover.

$$\begin{cases} \text{if } 0.5 \leq \hat{p}^i < 1 \text{ then } \hat{y}^i = 1 \\ \text{if } 0 < \hat{p}^i < 0.5 \text{ then } \hat{y}^i = 0 \end{cases} \tag{6}$$

Cost Function for Unregularized Logistic Regression

The estimated values for the regression coefficient matrix $\hat{\boldsymbol{\theta}}$ are those which minimize the difference between f and $\hat{f}$, i.e. minimize the cost function $J(\hat{\boldsymbol{\theta}})$. The cost function, used for the model, is represented by Equation (7).

$$J(\hat{\boldsymbol{\theta}}) = -\frac{1}{M}\sum_{i=1}^{M}\left[y^i \log(\hat{p}^i) + (1 - y^i)\log(1 - \hat{p}^i)\right] \tag{7}$$

The right-hand part of the equation is usually referred to as the log likelihood (log lik) function, as shown in Equation (8).

$$J(\hat{\boldsymbol{\theta}}) = -\frac{1}{M}\left[\text{log lik}\left(\hat{\boldsymbol{\theta}}\right)\right] \tag{8}$$

Equation (8) is usually minimized with mathematical and statistical programs, e.g. Matlab, Python, R, that have built-in optimization algorithms. 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 $\widehat{\boldsymbol{\theta}}$ while still being able to evaluate the model performance, the historical dataset is split into two parts: The training set and the test set. As such, the cost on the training set $J(\widehat{\boldsymbol{\theta}})_{train}$ can be calculated with Equation (7) when only taking into account those observations which are allocated to the training set, and by replacing M with the total amount of training observations M_{train}. The training set is then used to fit the model and the test set is used to report on the anticipated ability of the fitted model to accurately predict flashover or no-flashover on "unseen" observations, i.e. the approximation of the generalization error. The reason for using the test set is that the data examples used to calculate $J(\widehat{\boldsymbol{\theta}})_{train}$ do not classify as unseen anymore, and thus give an optimistic approximation of the generalization error. The deviance on the test set D_{test}, see Equation (9), is a metric which is commonly used to approximate the generalization error for logistic regression[20]. 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 simplest model, defined by the minimum D_{test} plus one standard error (SE), is considered to be the most parsimonious model[21], which in the remainder of the manuscript is referred to as the ideal scenario.

$$\begin{aligned} D_{test} &= -2\sum_{i=1}^{M_{test}}\left[y_{test}^{i}\log\left(\hat{p}_{test}^{i}\right)+\left(1-y_{test}^{i}\right)\log\left(1-\hat{p}_{test}^{i}\right)\right]\dots \\ &\dots+2\sum_{i=1}^{M_{test}}\left[y_{test}^{i}\log\left(y_{test}^{i}\right)+\left(1-y_{test}^{i}\right)\log\left(1-y_{test}^{i}\right)\right] \\ &= -2\log\text{lik}\,(\widehat{\boldsymbol{\theta}}) \end{aligned} \tag{9}$$

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 (10) 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 ($0 \leq R^2 \leq 1$), see Equation (11). 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.

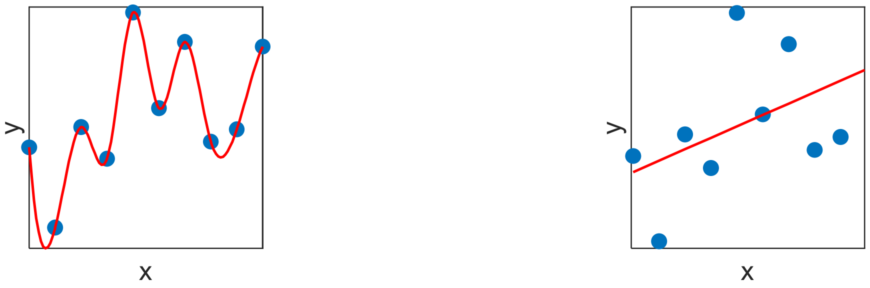
$$D_0 = -2\log\text{lik}\,(\hat{\theta}_0) \tag{10}$$

$$R^2 = 1 - \frac{D_{\text{test}}}{D_0} \tag{11}$$

In order to avoid an exceptionally good (or bad) allocation of observations to the test or training set 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 $\approx$ 13 and $\approx$ 10. It can thus be concluded that the fitted full 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

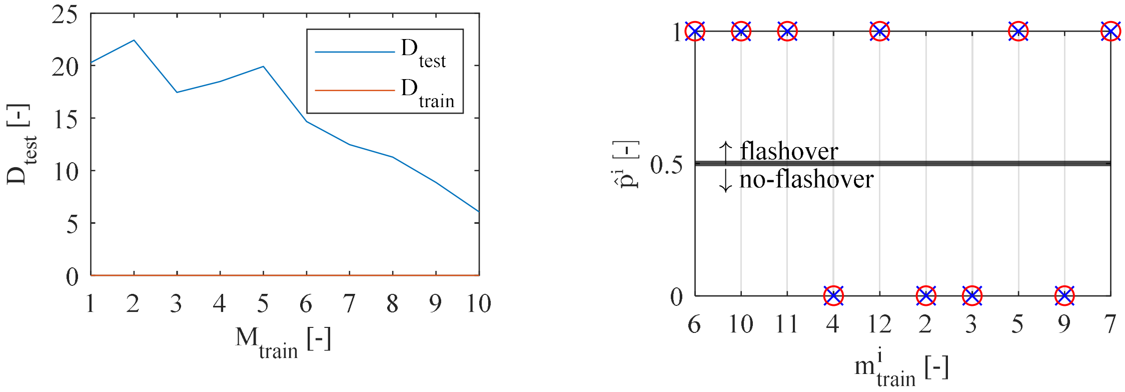
To improve the performance of the model the type of error is determined first. This is particularly important as the type will dictate the possible solutions. A high variance error signifies a $\hat{f}$ which is too flexible. As such, the model will find a pattern that is not actually true in the real world[22], see Figure 7a. 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 7b.



a) *b)*

Figure 7: Illustration of a decision boundary $\hat{f}$, red line, which represents a) A high variance/low bias and b) A high bias/low variance scenario.

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, or a priori knowledge, 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.



a) *b)*

Figure 8: a) The learning curves of the unregularized model indicate a high variance/low bias error. b) The predicted probability of flashover $\hat{p}^i$ (red circles) on the training set with the unregularized model coincides with the experimentally observed values (blue crosses) and implies that the model suffers from high dimensionality.

From Figure 8a 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 observations. 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 8b 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. Strictly speaking, high dimensionality refers to the case where the amount of observations M is smaller than the number of features N[21]. Because, many of the same considerations apply when M is only slightly larger than N, the next section will further elaborate the concept of high dimensionality.

Considerations in High-Dimensionality

Appendix 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[23]. 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 large feature collection possibilities of the internet[22], 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 $\approx$ 32 times. Shrinkage effectively introduces a shrinkage penalty $P_\alpha(\hat{\theta}_j)$ to the cost function applied to the training set, see Equation (12)[24]. For $\alpha = 0$, the estimated regression coefficients of non-predictive features are reduced towards zero, also referred to as ridge-regression. For $\alpha = 1$ the regression coefficients of non-predictive features are reduced to exactly zero, also referred to as 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 $P_\alpha(\hat{\theta}_j)$. 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 represented by Equation (7).

$$\begin{gathered} J_\lambda(\hat{\boldsymbol{\theta}}) = -\frac{1}{M}\sum_{i=1}^{M}\left[\log \text{lik}\,(\hat{\boldsymbol{\theta}})\right] + \lambda P_\alpha(\hat{\theta}_j) \\ \text{where } P_\alpha(\hat{\theta}_j) = \sum_{j=1}^{N}\left[\frac{1}{2}(1-\alpha)\hat{\theta}_j^2 + \alpha\left|\hat{\theta}_j\right|\right] \end{gathered} \tag{12}$$

It is advised to standardise the features with Equation (13) when applying shrinkage to make sure all inputs have a standard deviation equal to one and a mean equal to zero[19]. 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.

$$\tilde{x}_j^i = \frac{x_j^i - \bar{x}_j}{\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(x_j^i - \bar{x}_j\right)^2}}, \qquad for\ j = 1 \ldots N \tag{13}$$

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 always 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,[25] 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 approximate 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 approximate the generalization error, the historical dataset 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 dataset 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 the CV set and thus D_{cv} will be a better approximation of D_{test} than D_{train}. Nevertheless, as there is no subset to calculate the approximation of the generalization error the unbiased performance of the model cannot be reported. It should be noted that, the absence of D_{test} is not a problem for the model at hand, as the goal is to provide a guideline for the user on which experiment to conduct next, rather than providing the industry with a finished prediction tool for flashover or no-flashover.

Leave-One-Out Cross-Validation (LOOCV) was applied[19], 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. The process is then repeated M times, with for every run a different data example to be used as CV, and consequently D_{cv} is calculated as the average over M observations[19]. The advantage of LOOCV is the absence of randomness in allocating observations to subsets and the possibility to fit the model on almost the complete dataset. With the LOOCV method the null model deviance was found to be ≈ 21, which will be used as a benchmark in the following section.

RESULTS AND DISCUSSION

To determine the hyperparameters the model was fitted with the LOOCV method by minimizing Equation (12) for different combinations of α and λ. It was found that a lasso-regression model ($\alpha = 1$) gives the best performance in combination with a tuning parameter λ of approximately 0.01. The lasso-regression model with $\lambda = 0.01$ effectively reduces the regression coefficients $\hat{\theta}_{4-5}$ to zero, i.e. lasso-regression considers the features X_{4-5} not relevant for the prediction of flashover or no-flashover. As such the model is reduced from five to three features. With the lasso-regression model it was possible to obtain a model performance on the CV set of $D_{cv} \approx 2$, 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 ML 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 observations based upon the current historical dataset, 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 focus is to reduce the bias, 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[26] as: “To explain or to predict”.

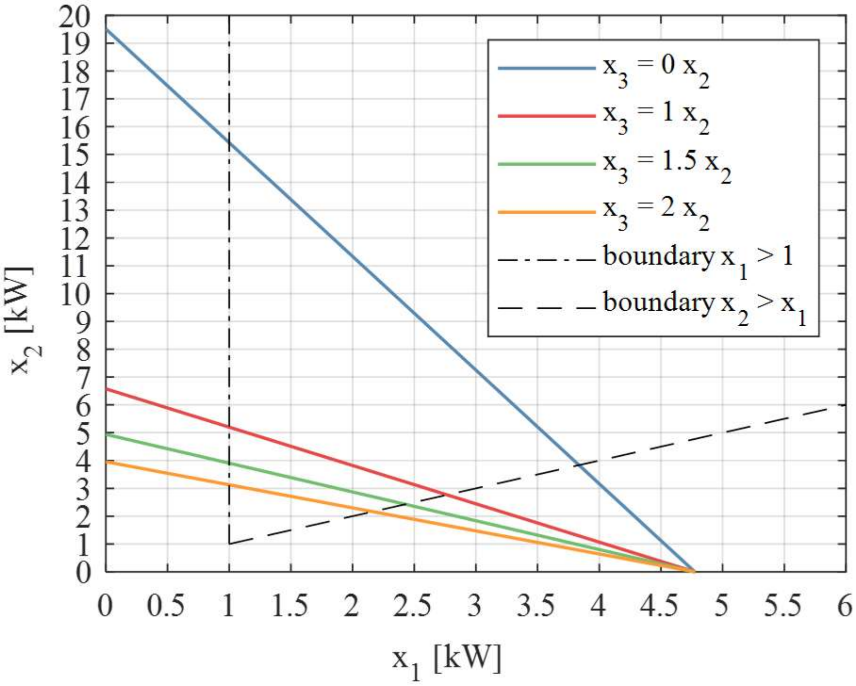
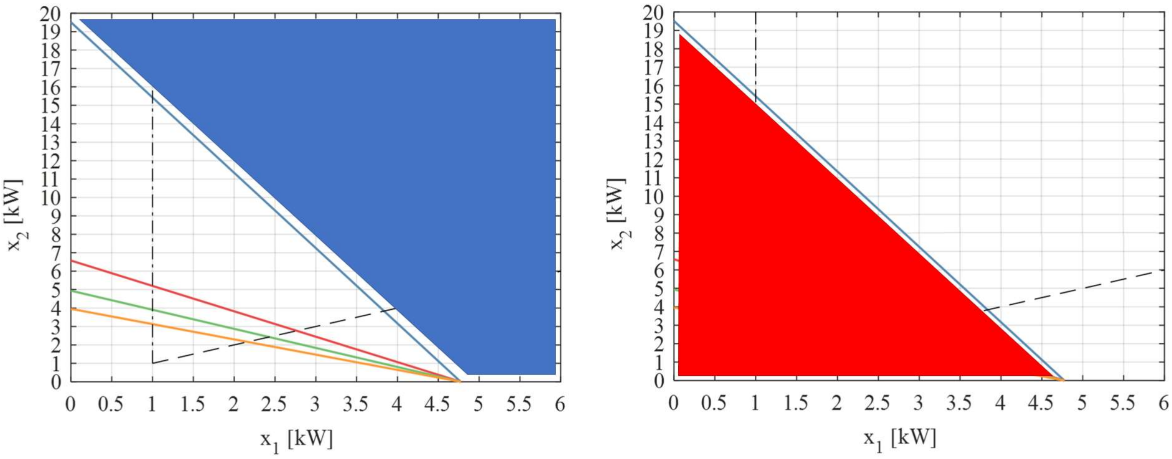


Figure 9: The model was fit on the complete historical dataset to determine the final definition of the regression coefficients. For the trained model, different assumptions were made for burning intensity three x_3. The resulting curves divide the space into a flashover zone, above the colored curve, and a no-flashover zone, below the colored curve.

Lastly, in order to arrive at the learned algorithm, i.e. determine the final regression coefficients, the complete historical dataset was used for training, in conjunction with the earlier defined hyper-parameters. To allow for a two-dimensional plot, the decision boundary was calculated for a set of fixed values for *X₃*. In other words, new feature combinations $\boldsymbol{x}^{i*}$ were determined, with $x_3^{i*} = 0$, $x_3^{i*} = x_2^{i*}$or $x_3^{i*} = 2x_2^{i*}$, such that the learned model assesses the probability of flashover as $\hat{p}^{i*} = 0.5$, see Equation (5). As such, each coloured line of Figure 9 divides the space into a flashover zone, see Figure 10a, situated above the line and denoting a $\hat{p}^{i*} \geq 0.5$, and a no-flashover zone, see Figure 10b, 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: $x_1^i > 1$ kW, $x_2^i > x_1^i$, $6 < x_4^i < 10$ and $x_5^i > 465$.



a)

b)

Figure 10: For a third burning intensity equal to zero, the following areas can be discerned in Figure 9: a) In blue, the input vectors for which the model predicts a higher than 50 percent chance of flashover. b) In red, the input vectors for which the model predicts a lower than 50 percent chance of flashover, i.e. no flashover.

It is important to understand that herein the aforementioned "learned algorithm" relates to the (historical) dataset available at a given point in time. In other words, the final regression coefficients are not truly 'final' from this point onwards, but are rather intended to be updated as more data becomes available. In practice, the algorithm will thus alternate between being learned and learning, as elaborated in the following paragraph.

The now learned model can be used to define future experiments that would result in the greatest knowledge benefit for the user and for the algorithm itself. That is, the decision boundary (DB) identifies the test conditions with input vectors $\boldsymbol{x}_{DB}^{i*} = \{x_0^{i*}; x_1^{i*}; \ldots; x_j^{i*}\}$, as shown in Figure 11, for which new experiments are predicted to have a 50 percent probability of flashover and a 50 percent probability of no-flashover. In addition, the further removed from the DB, the higher is the confidence of the learned model in predicting the outcome of the experiment (flashover above the DB and no-flashover below the DB). Therefore, the user and model will gain (almost) no additional knowledge from experiments with input values far removed from the DB, as the outcome of the experiment will be known' a priori. On the other hand, the experimental outcome of $\boldsymbol{x}_{DB}^{i*}$ may lead to new knowledge in an area previously unexplored. This can be explained by the fact that the probability of flashover and no-flashover being equal to each other, in a zone with limited available test date, is close to zero. Meaning that, the model requires further evaluations in this area to update its prediction accuracy. After every experiment $\boldsymbol{x}_{DB}^{i*}$, the newly obtained knowledge can be used iteratively to update the learned algorithm, and as such the learned algorithm becomes again a learning algorithm. The significant benefit of using machine learning, and thus the model set out herein, is that the model does not need recoding or rewriting to accommodate a changing dataset. That is, for every update the ML algorithm will re-evaluate the values for α and λ and change them accordingly. For example, if at one point more variables that are correlated to the output (rather than not) are analysed, the model will prefer ridge-regression over lasso-regression by changing α to a value closer to zero.

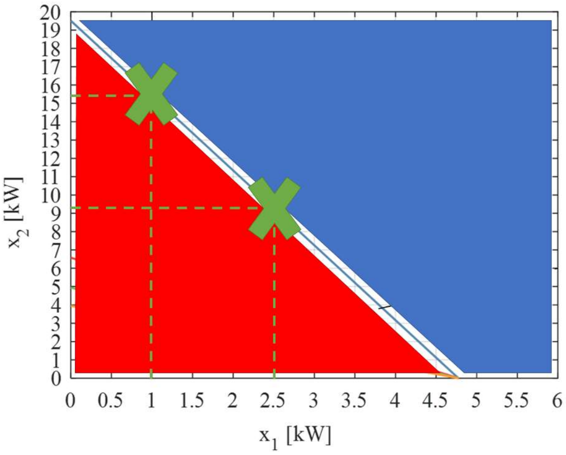


Figure 11: For a third burning intensity equal to zero in Figure 9, the decision boundary is the interface between the blue, flashover, and red, no-flashover, area. The interface denotes the input vectors, green crosses, for which the new test condition $\boldsymbol{x}_{DB}^{i*}$ is predicted to have a 50 percent chance of flashover and a 50 percent chance of no-flashover.

CONCLUSION

As the SBI test cannot guarantee an accurate classification for sandwich panels the dependency on the full-scale ISO 13784-1 test remains. As the latter is time nor cost efficient the need arises to derive a specific intermediate-scale test for screening purposes. Recent work by Leisted[15] showed promising results based upon a constant Froude number and a 1/5 scale model of the ISO 13784-1 test. The tool derived in this study demonstrates how ML can assists such work by providing an easy-to-use tool that can determine the relevant parameters and derive experiments which maximize the knowledge benefit.

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 dataset. By letting the algorithm determine which features are relevant for the response it was shown how machine learning can be used to discern the relevant experimental parameters. As a result, the user can decide either to delete the non-predictive features (and thus (possibly) introduce other experimental parameters) or amend the pre-defined parameter constraints. It can be argued that for the small dataset used herein, the added value over simple reasoning is small. Nevertheless, over time, new collaborations and ideas will inevitably lead to more parameters. Whereas a non-learning algorithm needs to be rewritten for every new input, the learning-algorithm evolves with the database without the interference of the machine-learning expert.

The logistic regression model was used to create the decision boundary (DB) to identify the combinations of input values that are predicted to result in a 50 per cent chance of flashover and a 50 percent chance of no-flashover. By identifying the areas where the model, and likely the user, have low confidence, it was demonstrated how the ML algorithm can help identify new experimental setups for their knowledge benefit. In addition, with every new experiment the learning-algorithm can iteratively update its DB and progress from a guidance tool to a more accurate prediction method. Nevertheless, a limited dataset inevitably means that the algorithm cannot capture all the physics, as it can only learn from the data it is presented with. As such, predictions with a limited database, i.e. at an early stage, should be used in combination with engineering judgement and within the boundaries prescribed by the historical dataset.

The flexibility of machine learning algorithms is unmatched in current models. Thanks to this, it might prove to be part of the solution for an ever-changing application of innovative materials and design solutions. Nonetheless, to arrive at a fully learned model that can be used universally, i.e. an algorithm for which the regression coefficients are permanently fixed, a large amount of data is needed. The tool presented herein partly overcomes the challenges associated with limited data as it is foreseen that the algorithm will develop as the database grows. In other words, the algorithm will make more crude predictions at first, and increasingly more accurate predictions as more data becomes available. In addition to the improved fidelity, the algorithm is expected to become more and more valuable with time due to the increasing complexity and size of the available dataset.

Nevertheless, the end-goal of the research, for which the work presented herein is considered to provide a valuable contribution, is to create a screening tool that can be used by anyone to predict the output of large-scale and intermediate-scale tests. For example, it can be used for. the single burning item test and the room corner test, as well as for various compartment geometries and materials based on parameters obtained from bench scale tests such as the cone calorimeter. Therefore, all the currently available fire test results need to be compiled in a database and new reaction to fire tests must be defined for their knowledge. Still, due to various limitations such as anonymity issues and the destructive nature of reaction to fire tests, this widespread data sharing platform is not deemed viable in the foreseeable future. Therefore, the envisioned algorithm will have to be a symbiosis between fire safety science and machine learning. This symbiosis will allow current science to fill in the knowledge gaps inherent to a limited database, and in turn allow machine learning to complement and enhance the knowledge-base in fire safety engineering.

ACKNOWLEDGMENTS

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

REFERENCES

1. Hansen AS, Hovde PJ. Prediction of time to flashover in the ISO 9705 room corner test based on cone calorimeter test results. *Fire Mater*. 2002;26:77-86. doi:10.1002/fam.788

2. Ostman B, Tsantaridis LD. Correlation between cone calorimeter data and time to flashover in the room fire test. *Fire Mater*. 1994;18:205-209. doi:10.1002/fam.810180403

3. Wickstrom U, Goransson U. Full-scale/Bench-Scale correlations of wall and ceiling linings. *Fire Mater*. 1992;16:15-22. doi:10.1002/fam.810160103

4. Messerschmidt B, Van Hees P, Wickstrom U. Prediction of SBI (single burning item) test results by means of cone calorimeter test results. *Conf Proc Interflam*. 1999:11-21.

5. Van Hees P, Hertzberg T, Hansen AS. *SP Rapport 2002:11 - Development of a Screening Method for the SBI and Room Corner Using the Cone Calorimeter*. Sweden, Boras: Fire Technology; 2002.

6. Hansen AS. No fire without smoke-Prediction models for heat release and smoke production in the SBI test and the Room Corner test based on Cone Calorimeter test results. *Doktoravhandlinger ved NTNU, 1503-8181; 46*. 2002.

7. Torero JL. Scaling-Up fire. *Proc Combust Inst*. 2013;34(1):99-124. doi:10.1016/j.proci.2012.09.007

8. 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.

9. British Standard Institution. *BS EN 13823:2010+A1:2014--Reaction to Fire Tests for Building Products. Building Products Excluding Floorings Exposed to the Thermal Attack by a Single Burning Item, European Committee for Standardisation*.; 2014.

10. Messerschmidt B. The Capabilites and Limitations of the Single Burning Item (SBI) test. *FireSeat*. 2008:70-81.

11. International Standards Organization. *ISO 13784-1:2014--Reaction to Fire Test for Sandwich Panel Building Systems. Small Room Test*. Brussel; 2014.

12. Axelsson J, Van Hees P. New data for sandwich panels on the correlation between the SBI test method and the room corner reference scenario. *Fire Mater*. 2005;29(1):53-59. doi:10.1002/fam.879

13. Johansson P, Van Hees P. *Development of a Test Procedure for Sandwich Panels Using ISO 9705 Philosophy*.; 2000.

14. Yoshioka H, Tanaka Y, Nishio Y, et al. Self-standing Compartment Fire Tests on Sandwich Panels. *Fire Sci Technol*. 2016;35(1):19-38. doi:10.3210/fst.35.19

15. Leisted RR. The Fire Performance of Steel-faced Insulation Panels with Stone Wool or Polymer Cores. [Doctoral Thesis]. Kgs. Lyngby, DK: Department of Civil Engineering, Technical University of Denmark; 2020.

16. Li YZ, Hertzberg T. Scaling of internal wall temperatures in enclosure fires. *Orig Artic J Fire Sci*. 2015;33(2):113-141. doi:10.1177/0734904114563482

17. Crewe RJ, Hidalgo JP, Sørensen MX, et al. Fire Performance of Sandwich Panels in a Modified ISO 13784-1 Small Room Test: The Influence of Increased Fire Load for Different Insulation Materials. *Fire Technol*. 2018;54:819-852. doi:10.1007/s10694-018-0703-5

18. Leisted RR, Sørensen MX, Jomaas G. 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. doi:10.1016/j.firesaf.2017.09.004

19. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning, with Applications in R*.; 2017.

20. Dobson JA. *An Introduction To Generalized Linear Models*. CRC Press Company; 2002.

21. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2nd ed.; 2017.

22. Domingos P. *The Master Algorithm : How the Quest for the Ultimate Learning Machine Will Remake Our World*. 1st ed. Basic Books; 2015.

23. Harrell FEJ. *Regression Modeling Strategies With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. 2nd ed. Springer US; 2015.

24. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1-22.

25. Naser MZ. Fire resistance evaluation through artificial intelligence - A case for timber structures. *Fire Saf J*. 2019;105:1-18. doi:10.1016/j.firesaf.2019.02.002

26. Shmueli G. To Explain or to Predict? *Stat Sci*. 2010;25(3):289-310. doi:10.1214/10-STS330

APPENDIX

Appendix Table 1: The experimental data which will be used for the ML algorithm, i.e. the historical dataset. Reproduced from Leisted[15]

i	X_1 [kW]	X_2 [kW]	X_3 [kW]	X_4 [m]	X_5 [s]	Y [-]	t_{fo} [s]
1	1.79	5.37	0	0.10	600	0	-
2	1.79	5.37	0	0.06	465	0	-
3	1.79	5.37	0	0.10	600	0	-
4	1.79	5.37	0	0.06	465	0	-
5	1.79	5.37	10.74	0.10	600	1	1270
6	1.79	5.37	10.74	0.06	465	1	1009
7	1.79	5.37	10.74	0.10	600	1	1249
8	1.79	5.37	10.74	0.06	465	1	963
9	1.79	1.79	1.79	0.10	600	0	-
10	1.79	5.37	5.37	0.10	600	1	1322
11	5.37	5.37	0	0.10	600	1	847
12	5.37	5.37	0	0.06	465	1	775
13	1.79	5.37	10.74	0.06	465	1	1010
14	1.79	5.37	10.74	0.06	465	1	1066