Predicting Non-Residential Building Fire Risk Using Geospatial Information and Convolutional Neural Networks

Jake Anderson-Bell [a], Calogero Schillaci [b]* , Aldo Lipani [a]

[a] *Department of Civil, Environmental and Geomatic Engineering, Gower Street, London, University College London*

[b] *Department of Agricultural and Environmental Sciences, Milan Via Celoria 2, University of Milan*

*calogero.schillaci@unimi.it

Abstract

Building fire risk prediction is crucial for allocation of building inspection resources and prevention of fire incidents. Existing research of building fire prediction makes use of data relating to local demography, crime, building use and physical building characteristics, yet few studies have analysed the relative importance of predictive features. Furthermore, image features relating to buildings, such as aerial imagery and digital surface models (DSM), have not been explored. This research presents a multi-modal hybrid neural network for the prediction of fire risk at the building level using the London Fire Brigade dataset. The inclusion of traditional and novel image features is assessed using Shapley values and an ablation study. The ablation study found that while building use is the most effective contributor of classification performance, demographic features, apart from social class, are detrimental. Moreover, while the DSM did not lead to any significant improvement in classification performance, the inclusion of the aerial imagery feature lead to a 4% increase in median validation ROC AUC. The final model presented achieved an ROC AUC of 0.8195 on the test set.

1. Introduction

Fire related incidents impact human communities across the globe, posing a threat to human life, damaging property, and hindering productivity in their wake. Recent advances in fire safety have aided a decrease in the total number of fires in the United

Kingdom from 473,000 in 2003/04 to 162,000 in 2016/17 (Bryant and Preston, 2017). While more effective fire safety and incident response systems have brought total casualties down, the costs associated with property loss due to fire have remained high since 1990 in the United States, residing at a value of $13.2 billion as of 2014 (Zhuang et al., 2017). Due to such cost associated with fire, municipal fire brigades, such as The City of Pittburgh's Bureau of Fire (Madaio et al., 2018) conduct inspections on properties to assess properties deemed to have a high risk of fire. Owing to the high number of potential properties it is not possible to carry out inspections on all buildings on a regular basis (Pringle and Welsh, 2015). Approaches to inspection allocation, employed throughout the world, rely on the analysis of the relationship between building fire incidents and potential determinant variables, ranging from physical building characteristics to sociodemographic factors, to focus efforts where they are needed. For this reason, numerous studies have focussed on finding effective means by which properties at high risk of fire can be more readily identified (Dang et al., 2019; Walia et al., 2018). By adopting such methods, fire brigade services can reduce their operational costs and more fires may be prevented. Although several studies have employed data regarding fire inspection, demography, and commercial information to classify non-residential building fire risk, the use of aerial imagery and digital surface models (DSMs) was rarely explored. Furthermore, while the datasets used are effective in their task, little research has investigated these variables as individual contributors to classification performance.

Risk prediction models have been implemented extensively in a range of subject areas, such as medicine (Preuschoff, Quartz and Bossaerts, 2008) and economics (Kuester et al., 2006). Although academic work has been dedicated to the task of predicting non-residential building fire risk, there has been a relative lack of studies that explore feature impact on classification performance and none have explored the use of aerial imagery and DSMs as features. In the context of machine learning ML, the term feature denotes a variable that is used to predict an output.

To focus fire risk inspection efforts to areas within a city where risk of fire a greater, some studies have grouped areas of higher building fire incidence together and treated them as a whole. In a randomized control trial with Surrey Fire Services, British Columbia, Claire et al. (2012) indicated zones of high residential building fire risk within the study area for a smoke alarm installation initiative. They then mapped residential structure fires as point features before a series of ellipses were drawn to capture points within zones containing high concentrations. A drawback of such an approach is the subjective nature of the methods. In contrast, DaCosta et al., (2015) joined American Housing Survey and American Community Survey datasets then used a Random Forest to model residential fire risk at census block level. Such a method, whereby regional open data is aggregated will also be adopted in this study. Increasingly, studies concerned with prediction of fire risk are looking to a more granular approach where individual building fire risk is considered. Garis and Clare (2014) developed a commercial building fire risk framework based on physical building characteristics in conjunction with history of fire regulation compliance. Although this system is heavily based on specifics of local fire safety regulation, their method builds a systematic approach to fire safety inspection resource allocation. To prioritise fire inspection to higher risk commercial properties, Madaio et al., (2015) joined historical fire incident and inspection records at the building-level with U.S. census data, such as age, ethnic population, and income, at a more regional level. They also used crime incident records and Google Place API data to get up to date business information. The area under curve (AUC) for true positive rate against false positive rate obtained from random forest and SVM models was 0.813 and 0.805, respectively, suggesting that both algorithms used produce a good separation of the classes. Although this research refers to data sources used, the features themselves are not described in detail. The current study draws some parallels from Madaio et al., (2015) in that Google Places API and demographic data will be used to gather information to aid fire risk predictions at the building scale.

Madaio et al., (2015) briefly touch on feature importance in their modelling. They find that for the random forest model, features relating to building size and its physical characteristics have the greatest impact on their output. In contrast, they then construct a logistic regression model and analyse feature coefficients which suggest that the Google Places feature has the greatest importance. Dang et al. (2019) used data provided by Humberside Fire and Rescue Service regarding property inspections and fire severity in conjunction with publicly available open data on food hygiene and Google Places ratings, among others, to build a commercial fire risk model. They experiment with several different learning algorithms.

While the multi-layer perceptron model they implemented attained an area under curve (AUC) score of 0.78, an AUC score of 0.89 for an XGBoost model was also achieved. Although this study represents the highest performing classification of commercial building fires in the literature, it does not give much consideration to analysis of features as individual contributors to classification performance.

Whilst there is some crossover in mutually used features in Madaio et al., (2015) and Dang et al., (2019), there is little rationale given for features used and attempts made to analyse the classification performance contributions of individual features are inconsistent within studies. In contrast to using a feature set that includes demographic variables, Hong and Jeong (2018) conducted a study whereby 16 features relating to physical building characteristics and fire history were used to make fire predictions of fire risk. Data was used to train support vector machine SVM, Naıve Bayes, Decision tree and artificial neutral networks ANN models. The SVM had highest test set accuracy overall at 63.54%. While many features used by Hong and Jeong (2018) relating to physical building characteristics were not available for this study, it is hypothesised that the use of aerial imagery and DSMs may supplement physical features that they used.

Although this study is only concerned with predicting fire risk for buildings, by using features based on aerial imagery, similarities are shared with studies in wildfire prediction and land use classification (Collins et al., 2020; Oliveira and Zêzere 2020), building assessment (Monfort et al., 2019), potential of rooftop solar energy (Schunder et al., 2020). Mitri et al., (2015) used pixel segmentation and subsequent object classification to determine presence of different wildfire fuels (e. g. grass, shrub, etc.) to discover areas susceptible to combustion.

In ML a hybrid model is a combination of two or more existing algorithms to produce a single output that can make use of different forms of data related to the same task. Such architectures have been implemented using CNN models for multiple image features (Wang et al., 2018) and ANN models for features of varying structure (Audebert et al., 2019). This method has advantages over running models separately as the error can be calculated over the entire network. Audebert et al., (2019) trained a hybrid model for classification of documents from image and textural input which attained a higher performance than the two separate models. A hybrid model will be implemented in this study to make use of multiple features of varying formats for each training example.

Convolutional neural networks (CNN) for risk-based classification of image data have been used more widely in the field of medical research than any other subject. Wang et al., (2018) uses three image features of lung scans to predict the malignancy risk of pulmonary nodules through use of a CNN. In their methodology they compare a hybrid CNN (whereby three image inputs per training example are inputted to a multi-branch

CNN) with a multi-channel fusion CNN (where three image inputs are layered on of one another to produce a single tensor then propagated through a single-branch CNN). They achieve multi-channel and fusion channel AUC or 0.93 and 0.97, respectively. Although Wang et al., (2018) achieve a greater AUC on the fusion-channel model, the multi-branch CNN approach makes more sense for this study as the inputs require merging with additional dimensionally dissimilar tabular data before a final classification.

Urban street view imagery has been successfully implemented in several studies to extract useful information about the built environment. Liu et al. (2017) used street view imagery to classify construction and maintenance quality of buildings in Beijing. The model they produced achieved an F1 score of 61.8%, suggesting that computer vision can be implemented to classify quality of buildings to some extent. Similarly, Law et al., (2018) produced a CNN model that classified street view imagery by aesthetic street frontage quality. Furthermore, Law et al., (2018) used aerial imagery in conjunction with street view imagery to estimate house prices. This research relates to these studies in that an attempt will be made to gain information regarding the built environment through computer vision.

Numerous studies have investigated classifying roof types of buildings. In a study concerned with building detection and roof type classification from aerial imagery Alidoost and Arefi (2018) used labelled instances of roof types to achieve a classification accuracy of 92% with a CNN model. Such research suggests that ML algorithms have the potential to classify areal characteristics of buildings.

Our study will address this gap in the literature by exploring the use of aerial imagery and DSMs, and assess the relative importance of these in addition to traditional variables for non-residential building fire risk classification using CNN. Moreover, we assess the ability of traditionally used features in building fire risk prediction. We do this using a feature attribution method, named Shapley values, to understand the importance of feature values towards the prediction; and conduct a feature ablation study to assess the impact of each individual feature.

2. Material and methods

2.1 Data

The following section details the procedure taken and rationale behind the implementation in the data pre-processing and methodological stages of the study. Scripts used in this research can be found in a Github repository and accessed via the URL in supplemental materials. The tabular data included in the predictive modelling of non-residential building fire risk was chosen due to them being included in previous studies. This allows better comparison of the results of this work with other models constructed in the literature. While some data sets like those used previously have been implemented here, data availability has been a constraining factor. Many variables used in this study have been attributed to buildings via spatial joins, summarised in Figure 1. An assumption of this method is that data collected for a building is representative of that building, the people who use it and the processes that occur within or around it. Furthermore, it is assumed that data of a more regional resolution that is attributed to individual buildings is also representative of the social characteristics at a local level.

Due to the investigation of the inclusion of imagery and surface models in prediction being a central aim of this study, the approach chosen has been influenced by methods where image data may be an input to the model. An ANN was chosen as the algorithm to be used as they have been successfully implemented in recent years in the field of image classification and can accept multiple data types as inputs to the model (e.g. tabular data and imagery) (Geiß et al., 2020; Kim et al., 2020a, 2020b).

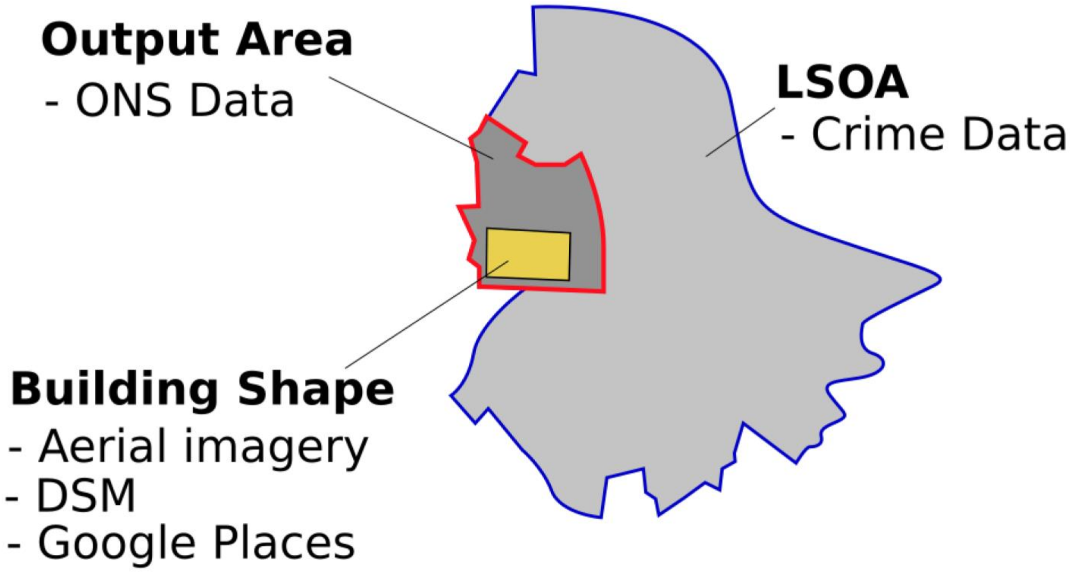


Figure 1: Diagram illustrating how data from varying levels of spatial attribution have been merged to represent individual buildings, LSOA = Lower super output area

2.1.1 London Fire Brigade Incident Data

The London Fire Brigade (LFB) provide information regarding all fire incidents reported from 2009 to present (London Fire Brigade, 2011). This was used in order to find locations of fire incidents. The accuracy to which the incident location is recorded varies within the dataset. This is recorded with an address qualifier variable that states, for example, whether the incident location is correct or accurate to the street on which it occurred. When the proportion of building category fires in the dataset is compared to the proportion of building category fires recorded at the correct incident location (figure 1) it can be seen that there is a bias towards collecting the correct location information for non-residential fires. Although only non-residential building fires are the subject of this study, this bias suggests that there are some inconsistencies within the ways that the data is collected. While this is only apparent in a contrast between how residential and non-residential buildings are handled, it is not clear whether there exists a bias within the way that non-residential building fire locations are recorded.

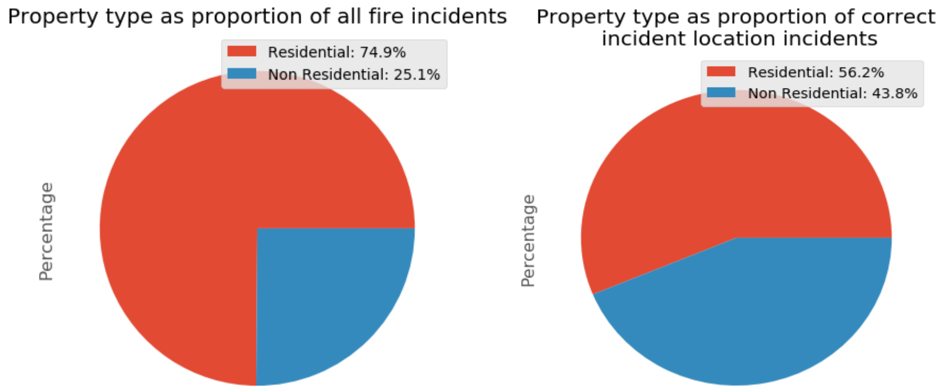


Figure 2: Proportions of property types for (a) all fire incidents in dataset and (b) fire incidents with correct location

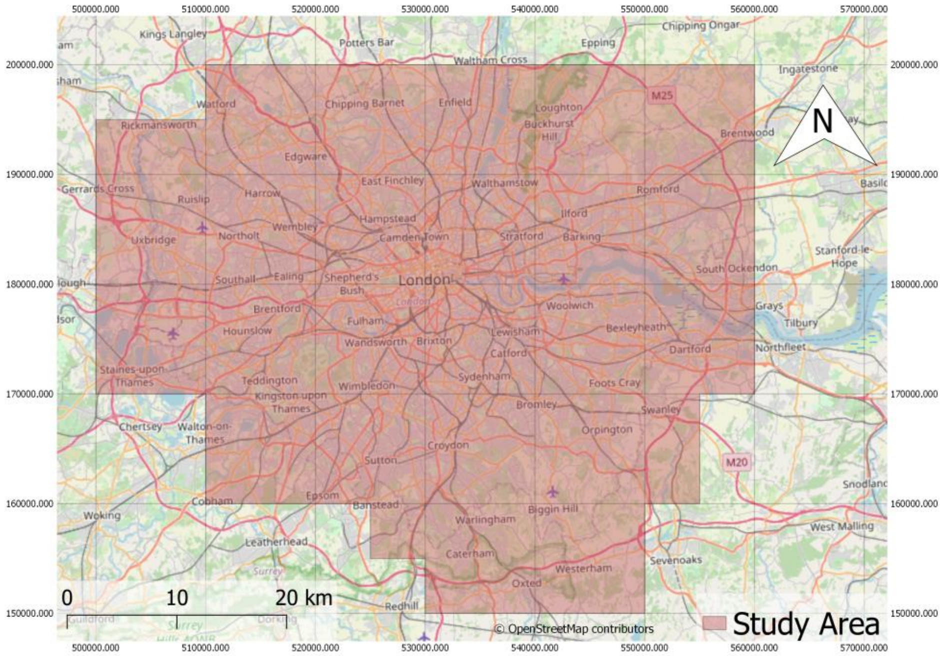


Figure 3: Study area used in this investigation, Greater London

The geographic study area is based on the spatial extent of the LFB dataset. This covers the area of Greater London and is presented in figure 3.

This study is concerned with fires whose severity may have been impacted by building design. For this reason, all fires in the dataset were filtered for those that were primary fires. Primary fires are generally more serious and caused more damage than other categories of fire (Home Office, 2020). The address qualifier variable, stating to what degree of accuracy the incident location is recorded was used to filter incidents for those that are recorded at an accuracy of being in the correct building or greater. The location of the incidents were recorded in British National Grid, which was used to convert the tabular data into point features.

2.1.2 Ordnance Survey MasterMap

Ordnance Survey MasterMap (MM) is a database of shapefiles recording every fixed feature larger than a few metres in Great Britain (Ordnance Survey, 2020). This data was used to obtain the shapes of buildings used for clipping aerial imagery and DSMs. This data was accessed and downloaded from Digimap (Edina, 2019). Multiple versions of MM were been downloaded to match the annual extents of the aerial imagery.

MasterMap shapefiles covering the area of Greater London were collected for all years available within the timespan of the London Fire Brigade (LFB) dataset. After being filtered for buildings, LFB incident points were grouped annually, and a spatial join was performed between the LFB points and the version of MM building polygons closest to the time of the LFB incidents. All points that did not fall within a building shapefile were eliminated from the data set.

2.1.3 Aerial Imagery and Digital Surface Model

The vertical aerial imagery used in this study was obtained using EDINA Aerial Digimap Service (Edina, 2019) and collected by Getmapping (Getmapping, 2019). Although satellite imagery may have been available for use in this study, the flight captured imagery had a resolution of 25cm ground sample distance (GSD), greater than any other open aerial imagery data available at the time this research was conducted.

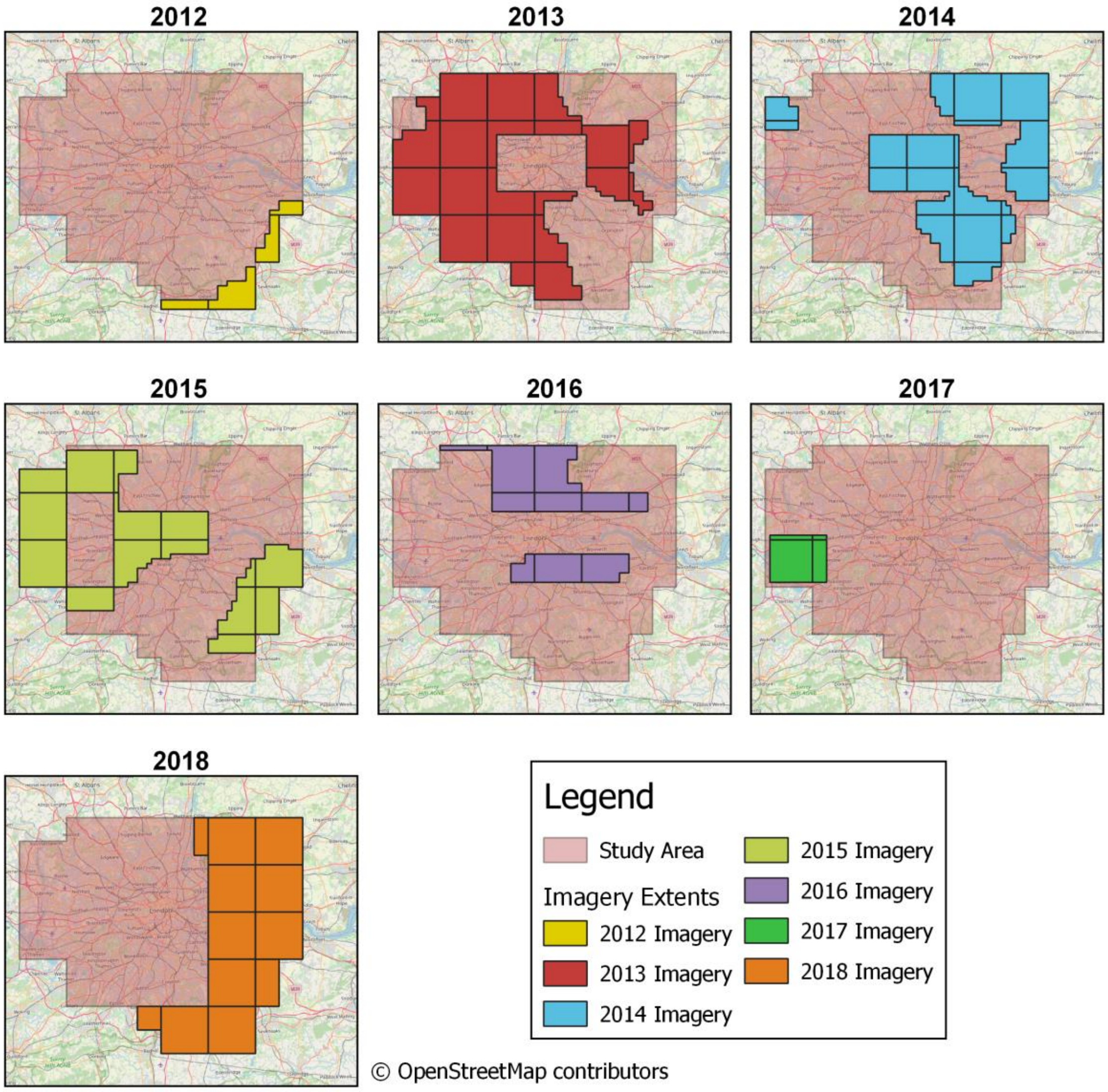


Figure 4: Annual spatial coverage of aerial imagery used in the study (10km tile grid)

Aerial imagery used was collected between 2012-2018, however different extents of the study area were collected each year (figure 4), with each point on the ground being revisited every 3-4 years. Georeferencing had already been performed on the imagery and this was provided in the form of JGW files. The imagery was not orthorectified so some building lean exists in the data. The digital surface models were accessed from EDINA DigiMap (Edina, 2019) and consist of a LiDAR composite digital surface model. The data was collected by the Environment Agency from 2006 onwards and is available under the Open Government Licence for public sector information (Environment Agency, 2020). The dataset has a vertical accuracy of ±15cm RMSE and a spatial accuracy of ±40cm RMSE (Environment Agency, 2020). As the resolution of the dataset is 1m, greater than the mean spatial error, the spatial error has negligible impact in this application.

Imagery metadata was used to determine the date that each image was taken. Because the appearance of, or even the buildings themselves, can change over time it was decided that only imagery that was taken within a year prior to a fire incident would be used. The recorded incident date was used to find relevant imagery. In contrast, the DSM was much more mixed in terms of when it was collected and processed, for instance some individual tiles have data collection spanning several years so it is difficult to determine the exact time in which the DSM for a building was collected. While building structures do change over time, they are not as variable as the appearance of buildings and so the time of DSM was not considered in this study.

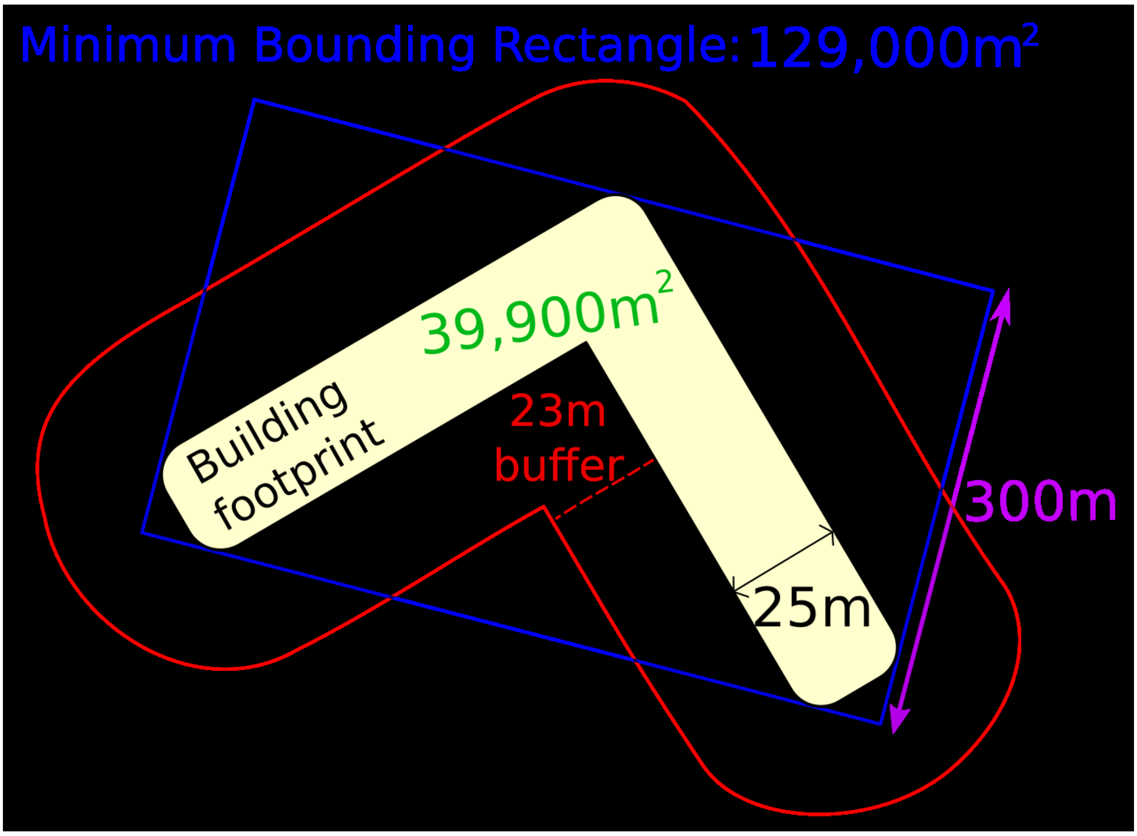


Figure 5: Diagram showing how data regarding building shape is used to calculate a scale and shape complexity variant buffer distance with equation (1), Pink=MRA shortest dimension, Green=Shape area, Blue=MRA, Red=Buffer distance.

Building footprint*0.25*300(39900/129900) =23

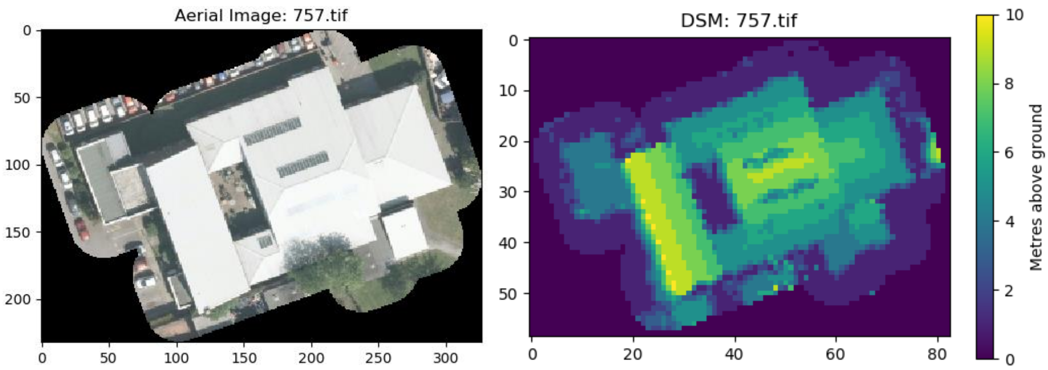
The area surrounding a building may also have an impact on its fire risk, so a buffer was made around the building before clipping the imagery. A building size dependent buffer amount was chosen due to variation in building size within the dataset. When using a buffer size proportionate to the area of the building or the bounding box of the shape, complex or narrow and branching building shapes became dominated by the buffered area (Figure 5). A method was required for the final images to be representative of the building, whilst also considering the surrounding features.

The following rule was applied.

$$BD = 0{:}25 *S\ (A/MRA) \quad (1)$$

Where BD is the buffer distance, A is the area of the shape, MRA is the area of the minimum bounding rectangle of the shape, and S is the shortest dimension of the minimum bounding rectangle. This method is summarised in figure 6.

When the building is a rectangle this results in a buffer size of a quarter of its shortest dimension, however, with more complex building shapes with more open space within their bounding boxes, the size of the buffer becomes moderated by the ratio between the area and the minimum bounding rectangle area.



(a) Aerial imagery, left panel (b) Digital surface model, right panel

Figure 6: Example of aerial imagery and DSM data after clipping (Getmapping, 2019; Environment Agency, 2020).

The building shapes were then buffered with a buffer size from the equation (4.1) and this was used to clip the aerial imagery and DSM to image files (figure 6). The minimum height of the DSM for each building was then subtracted from each file so that the DSM was relative to the ground level in each instance. The aerial imagery was greyscale sampled to produce a 2D feature. The regions of the aerial images and DSM that fell outside of the buffer were given a pixel value of 0.

2.1.4 Google Places

The Google Places API is a tool implemented in this study to obtain information regarding businesses and services that occupy a building (Google, 2020). In addition to finding information about the buildings in which incidents occurred, Google Places was also used to find commercial data for spatially random buildings where fires did not occur. Using the service involves sending a request with a query to the Places server before receiving a response of 20 results per request. Requests can be made by place name, address, or by location and can be filtered by type of place.

To get the business information about the buildings in the fire incidents dataset, a request was made for each MM building centroid linked to a fire incident. A radius option, whereby places returned whose location falls within were prioritised in the request, was set at 100m. Returned places results were searched for businesses whose location was within the building shapes. The building, incident and place data were then joined. The Places API was also used to find commercial buildings to be used for the negative fire incident dataset (i.e. those where a fire did not occur). To attempt to have the negative classes to have a similar spatial distribution to the positive class, the frequency of fire incidents in each 10km tile was calculated. These were then used as the basis for the quantity of negative classes to find in each 10km tile. For each tile, a series of random points was generated and used to make Places API requests. The results were then searched for places whose location fell within buildings where no fire incident was recorded. This served as the foundation for the negative fire examples for which aerial imagery and DSM was also clipped. Due to 95 categories of business being represented in the data, attempts had to be made to reduce the sparsity in the data to improve performance. These Places data were aggregated into 14 broad categories: amusement, car, drink, emergency, food, contractor, leisure, medical, office, public building, retail, service, transport, and storage. A summary of all type categories before and after reduction can be found in Table A.2 Supplemental materials (appendix A). The places type data was then converted into one-hot-vector variables for categorical data to be represented in the model.

2.1.5 Demographic and Crime Data

Demographic features were used in this study due to their inclusion in previous building fire prediction research (Madaio et al., 2016; Walia et al., 2018). Demographic data was collected and released by the ONS and accessed from EDINA DigiMap as data attributed to output area shapefiles (Office for National Statistics, 2016; Edina, 2019). The census data used was released in 2011 and is the most recent census data available (Office for National Statistics, 2011). Data relating to age, employment, education, social class, residency, tenancy, and ethnicity was collected for the study in order to evaluate their contribution to classification performance. Crime data was included in this study due to its usage in prior research (Madaio et al., 2016; Walia et al., 2018). Crime data was collected by the Metropolitan Police service and accessed from the London data store (London Data Store, 2019). Crime records represent numbers of crime incidents at the lower super output level (Office for National Statistics, 2016). Crime rates for 2016 were used as this interval represents the period in the middle of the incidents used. Population data was also collected and used to derive crime rates from the crime totals.

The demographic data came in the form of output area shapefiles of ONS output areas attributed with demographic variables. The attribute tables were filtered for series that were relevant to the application before spatial joins were performed to attribute building shapefiles with the desired variables. Crime total data was acquired in the form of tables of crime rate data organised by lower super output area (LSOA). LSOA crime values were divided by LSOA population totals to produce crime rates. A shapefile of London LSOAs was downloaded, and crime data was joined to produce geographic extents of crime rate values. This was then spatially joined to the MM buildings to pass on the desired attributes.

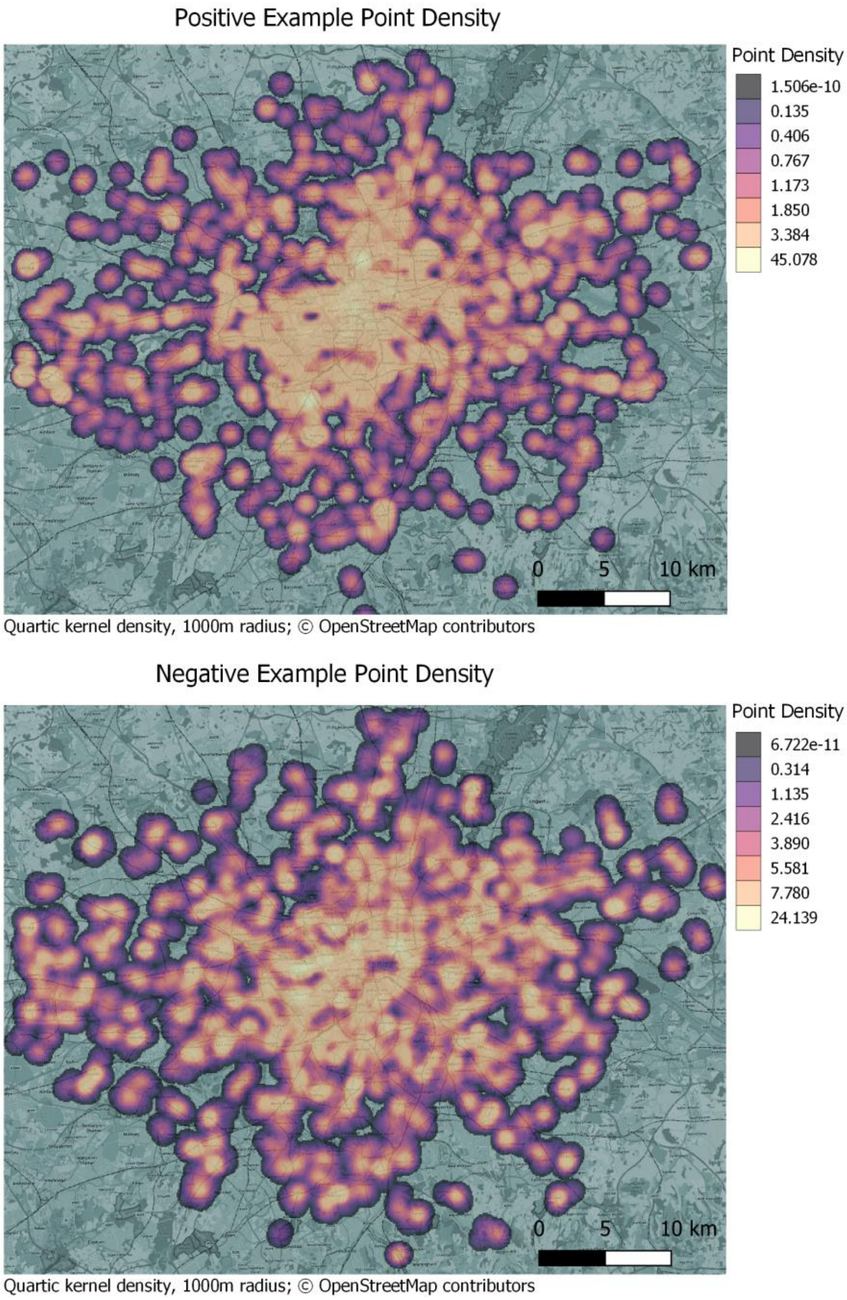


Figure 7: Spatial distribution of datasets for positive and negative classes

2.1.6 Final Data

After elimination of data entries that were missing features, a total of 6690 examples remained in the final dataset. Within the final data were 2087 positive examples and 4603 negative examples of building fire, giving a ratio of 0.312: 0.688 to be used for class weighting. The spatial distributions of the classes can be found in Figure 7. All data series apart from the aerial imagery and DSM were put into the same data table and normalised in preparation to be inputted to the model. The image and DSM files were put into their own directory. The image data was also normalised before being used in modelling. A training-validation-test split of 80:10:10 was used.

2.2 Modelling

The PyTorch package was chosen as the framework for using ANN in this study due to general usability, ease at which it can be implemented, GPU Capability and its wide range of features (PyTorch, 2019).

2.2.1 Architecture

An initial architecture compatible with the input data types had to be chosen upon which hyperparameters could be tested. Two CNNs (Figure 8, a,b) would each take an image input then a standard ANN would take the tabular data and run it through a series of fully connected layers before being concatenated to the outputs of the convolutional layers (Figure 8c). This would then go through a final fully connected layer before classification. The path of each data input through the model will be referred to as a branch (e.g. the image branch). This architecture is illustrated in Figure 8 and 9. The benefit of having a hybrid model is that the algorithm can make a classification based on multiple sources of data that are of different types (Audebert et al., 2019). This allows the model to optimise itself across the different data sources.

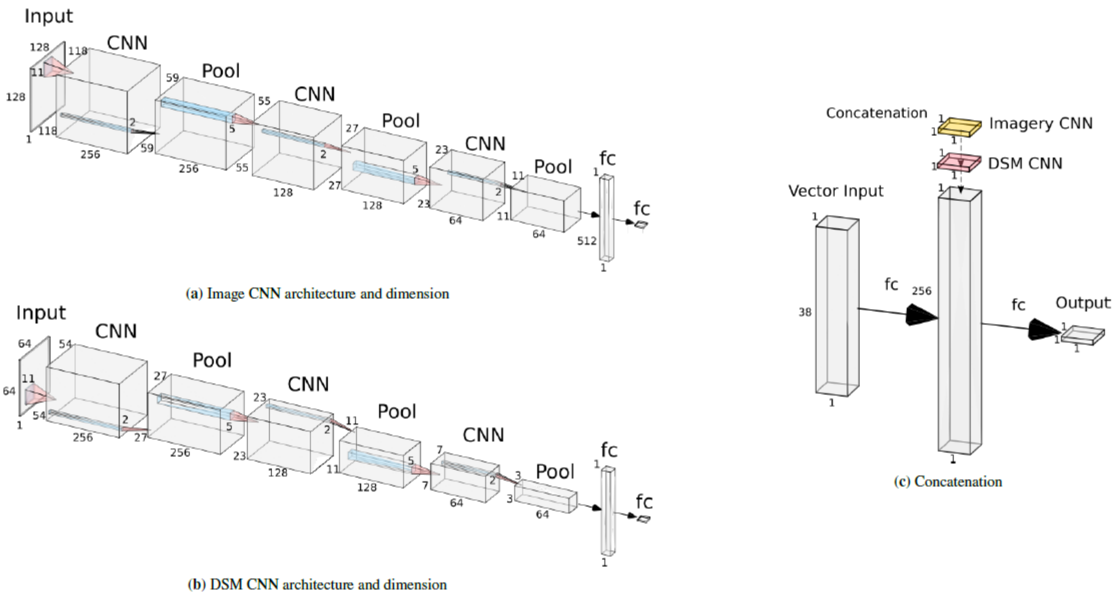


Figure 8: Diagrams summarising model architecture: The output of CNN image branches (a) and (b) are concatenated to the output of a fully connected layer of the vector branch before a final fully-connected layer (c), produced using NN-SVG (LeNail, 2019) rather than make separate models that each make an individual classification.

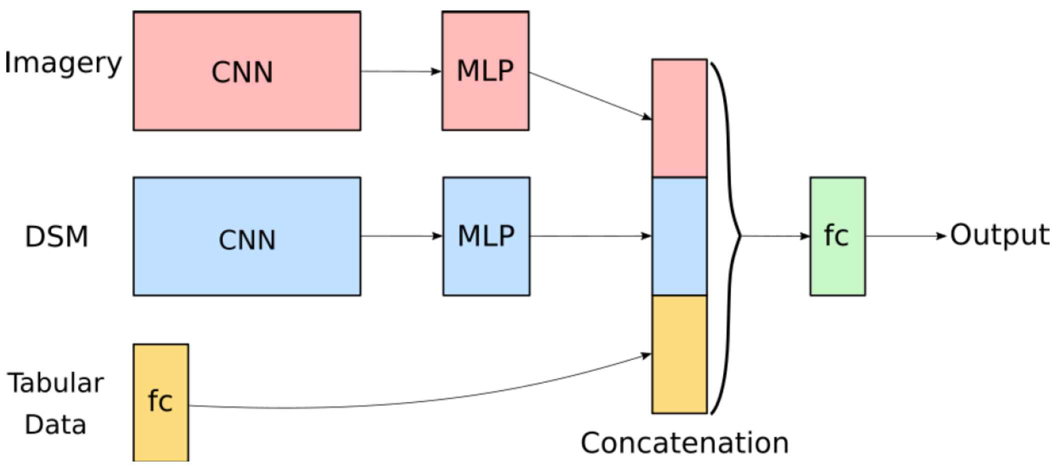


Figure 9: A schematic diagram of the model architecture chosen after hyperparameter testing. CNN=Convolutional neural network, MLP=Multi-layer perceptron, fc=Fully connected layer (a) Image CNN architecture and dimension, (b) DSM CNN architecture and dimension, (c) Concatenation

Preliminary testing of the CNN branches as individual models found that by having a 11x11 convolution kernel size in the first CNN layer led to an 13% reduction in average validation error when conducted over 10 sets of training. This was used as it may be that the larger kernel is able to recognise larger objects within the images. The other convolution kernel sizes were kept at 5x5. The increased size of the first convolution kernel is also seen in the architecture of AlexNet (Krizhevsky et al., 2012).

2.2.2 Loss Function

The classes of the dataset being used in this study are unbalanced and so a loss function that takes this into account is used so that each class is treated equally. The binary cross-entropy loss function is used as, along with the model prediction and target, this function also takes in a weighting of each example (Phan et al., 2006). The values 0.688 and 0.312 were used for the positive and negative classes, respectively. These are proportionate to the class imbalance so that the model will not just attempt to optimise with respect to the majority class.

2.2.3 Training Evaluation

Model training is achieved by inputting a dataset to the model and mapping this input to a desired target output. The error between the target output and prediction will usually improve with number of epochs taken however it is not a good indicator of model performance as the model has seen the training data before and may eventually learn the subtleties of the training data perfectly, thus overfitting the data. By testing performance against an unseen validation data set, the model's ability to generalise can be investigated. This gives a better indication as to whether there is a correlation in the data used to make predictions and the actual target the model is attempting to predict. Loss was tested against the validation data on every 32nd batch during training. This was done over 128 data samples, close to the maximum number that could be accommodated on the GPU.

All pre-processing stages were conducted on a Dell XPS 9550 laptop with an intel i7 processor. A computer running a Nvidia Titan RTX was connected to via SSH for running parallel processing in the methods used to train the ANNs.

Pre-processing was carried out on a Windows 10 operating system with Python 3.7 using PyCharm 2019.2 as the IDE, while Jupyter Lab was used for SSH. Python packages used in the study are summarised in Table A.1 in Supplemental materials, appendix A while scripts used can be found in appendix B.1.

2.2.4 Hyperparameter Optimisation

Overall, 162 combinations of hyperparameters were tested. The hyperparameters tested are summarised in Table 1. Due to the complexity of the architecture of the model used, simplifications have been made to the range of potential hyperparameters that could be tested. In each branch a starting number of nodes was specified for the first hidden layer then this value was halved with each succeeding hidden layer. Furthermore, the hyperparameters associated with the CNNs were both altered in unison to further reduce the number of required hyperparameter permutations. The rectified linear unit function was used for activation of layers as it offers better performance and generalization when compared to some counterparts (Chigozie et al., 2018). A sigmoid function was used for the output to produce a probability value.

Table 1.

A learning rate scheduler was implemented whereby an initial rate of 0.005 was halved whenever validation loss did not improve for 7 consecutive epochs. An early dropout was also implemented where if the model did not improve for 20 epochs training would cease. If validation loss was lower than any previous end-epoch value, the model weights were saved, potentially overwriting a previous epoch's weights. This allowed the highest performing weights to always be preserved. As training was run on a GPU to speed up processing, the batch size was constrained to a maximum of 128 so that there was sufficient memory available to hold all data and weights at any time.

Hyperparameter	**Description**	**Values tested**
Vector start nodes	The number of nodes in the first hidden layer of the vector branch	64, 128, 256
CNN start nodes	The number of nodes in the first hidden convolutional layers of the image branches	64, 128, 256
Vector layers	The number of hidden layers the vector data propagates through before concatenation	1, 2, 3
CNN layers	The number of convolutional layers the image data propagates through before concatenation	1, 2, 3
Batch size	The number of training examples to be used in backpropagation in each epoch	64, 128

2.2.5 Shapley Values

In order to calculate Shapley values, the SHapley Additive exPlanations (SHAP) package (Lundberg and Lee, 2017) is implemented where, for each example put through the model, feature values are altered to observe the impact they have on the model output. It is assumed that by giving a feature a low value the absence of a feature is simulated.

2.2.6 Ablation

Individual features and feature groups were aggregated within the tabular data, summarised in Table A.2 in supplemental materials, and were ablated by excluding the features from training. With the aerial image and DSM CNN branches, each CNN branch architecture had to be excluded along with the feature. For each feature ablated the model was trained 10 times and median validation ROC AUC was used to compare the performance of the models. The highest performing model from this analysis would then be used as the final model.

2.3 Cost-Benefit Analysis

Building fire incidents have the potential to inflict high cost of repair and even fatalities when they occur. For this reason, there is motivation in taking preventative measures towards reducing fire risk, such as inspections carried out to ensure that the building complies with fire regulation. The cost of an inspection in conjunction with the potential cost of a fire incident may be combined in order to assess the costs and benefits of mitigation measures, where the resources required to carry out the inspection are thought of as the cost value then the cost of fire that may be avoided will be the benefit value. The cost-benefit ratio can then be used as the acceptable FN/FP ratio in classification, defined below (Sheng and Ling, 2006).

$$\text{Cost/ benefit} = \text{FN/FP} \quad (4.2)$$

A cost-benefit analysis will be implemented in this study to suggest an optimal classification threshold of the final model in a fire brigade's operational scenario. While the LFB does not publish the cost of a building inspection, a value of £1875 was averaged over three commercial quotes for the building fire safety inspection of a 230m^2 property, the median size in the dataset used. The average area of fire damage of a non-residential building in the UK was 28.3m^2 in 2018/2019 (Home Office, 2018), while the average cost of fire damage for non-residential buildings is £1405/m^2 (Salter et al., 2013). This gives an average fire damage cost of £39,761. Thus, the desired FN/FP can be calculated.

$$1875/39761 = 0.047 \quad (4.3)$$

The threshold at which the FN/FP is closest to 0.047 will be used to demonstrate the performance of the final classifier in an operational setting.

3. Results

3.1 Hyperparameter Optimisation

Experimentation of model hyperparameters was conducted to find the best parameters for a model. As training was conducted on an unbalanced dataset the validation loss was used to measure model performance during training. The hyperparameters of the model used are presented in Table 2, while changes in error associated with varying hyperparameters are graphed in (Figure 1 supplemental materials).

Table 2: The hyperparameter combination used for the hybrid model in this study

Hyperparameter	Value chosen
Vector start nodes	256
CNN start nodes	128
Vector layers	1
CNNlayers	3
Batch size	128

The lowest validation loss achieved in hyperparameter testing was 0.225. The hyperparameter combination of this model is summarised in Table 2 while the losses and accuracy values are shown in Figure 10. The weights of the chosen model were saved after training over 47 epochs, shown by the green vertical line in Figure 10, before training was dropped out after 67 epochs.

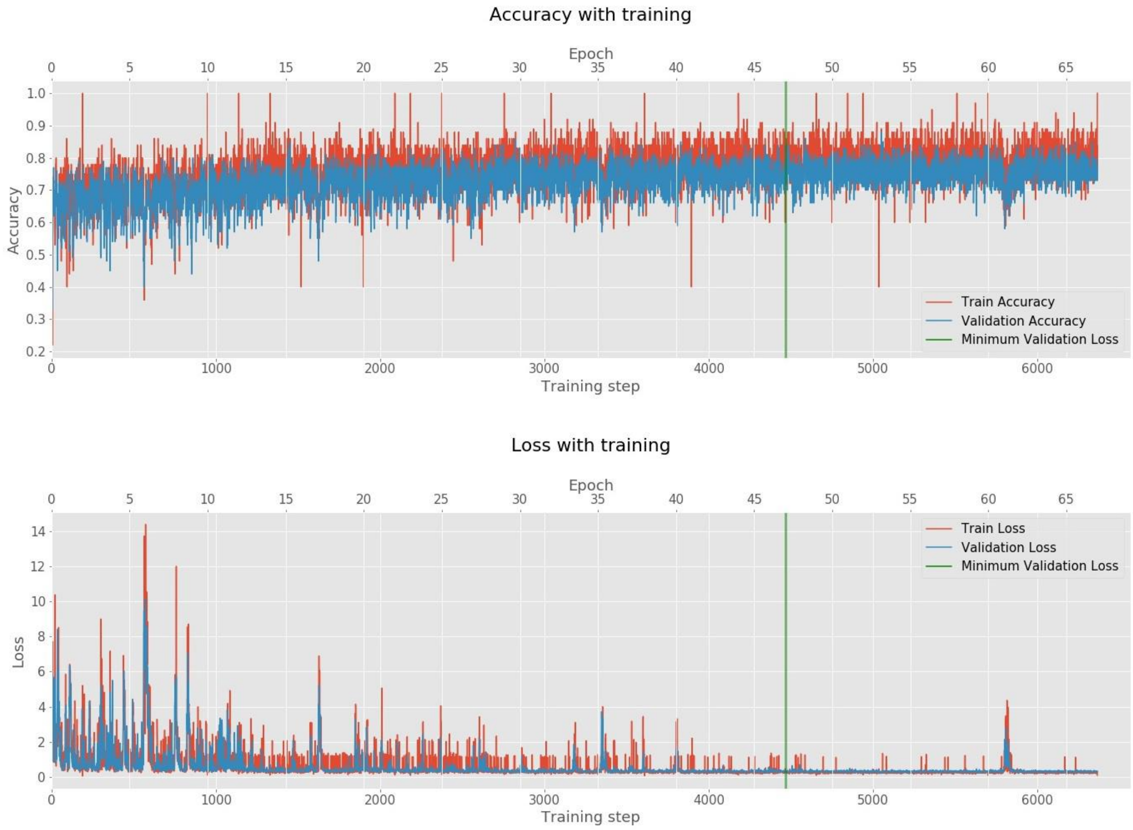


Figure 10. Graph showing changes in loss and accuracy over train and validation sets throughout training

It can be seen from Figure 10 that, overall, train accuracy increases and train loss decreases throughout training of the model, suggesting that there is a correlation between the training data and the target that the model is capable of learning. Furthermore, a similar pattern in the validation accuracy and loss suggests that the learned mapping is consistent across the dataset and train performance are not entirely due to overfitting. This model achieved a validation ROC AUC of 0.778.

3.2 Shapley values

Shapley values were calculated over the entire training set for all features from the tabular data using the SHAP package. In general, most features show some contribution to the output value. The summary plot in Figure 11 show the SHAP values. Each row on the plot represents a feature of the dataset and each training example is represented by a dot.

The colour of the dot represents the value of the feature while the SHAP value is shown by the dots position on the x-axis. The area feature is shown on its own separately on Figure 12 using a different scale while all other features are shown on Figure 11.

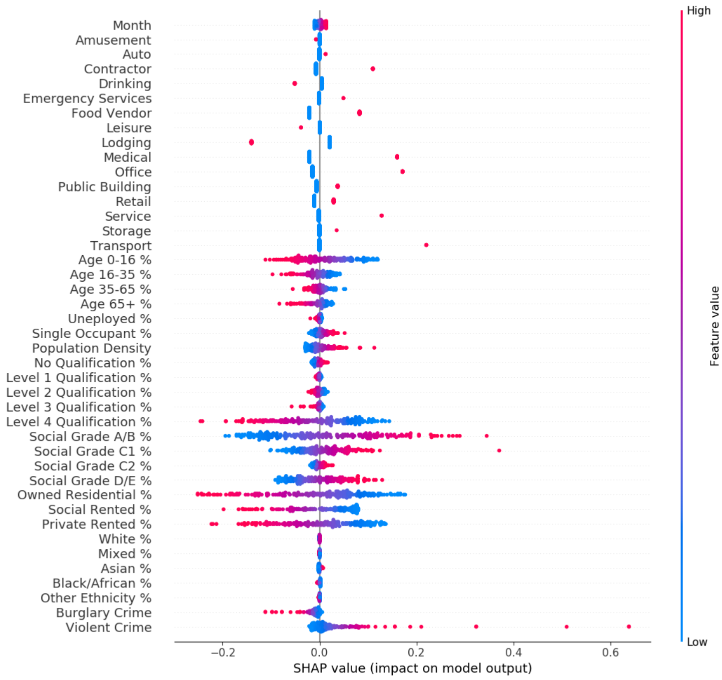


Figure 11. SHAP summary plot for all tabular features excluding building area

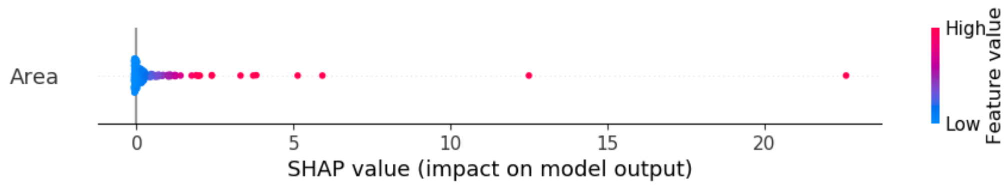


Figure 12. SHAP summary plot the building area feature

For the area feature, in very few examples, a high feature value heavily influences the target output. Compared to other features, the area feature has a vastly higher mean SHAP value at 5.7, 32 times greater than the second highest mean SHAP value, for violent crime at 0.018.

3.3 Ablation

In the feature ablation portion of this study features were excluded from the dataset, models were trained using the new feature set, and performance was evaluated to gauge the impact of the individual features on the classification performance of the model. Each model was trained 10 times to get a reliable estimate of performance. Figure 13 shows a boxplot of all ablation AUC results for each group removed. A lower position on the y axis represents a deterioration in classification performance when that feature was removed. The median statistic is chosen for comparison as several groups exhibit a skewed distribution of AUC values.

In Figure 13, while models trained excluding some feature groups, such as age, show a narrow range of AUC scores, for other features, such as social class, there is a wide range. Comparisons between median AUC is more easily made by observing differences between the ‘all’ feature model median AUC (greed dashed line on Figure 13) and the group removal medians (blue lines within boxes). By removing either the age, crime, ethnicity, month, qualification or residential features, the median AUC increases. Conversely, it is seen that the removal of either area, demographics as a whole, the image branch, places or social class lead to a decrease in the median AUC. When the DSM branch is removed there is little change in median AUC.

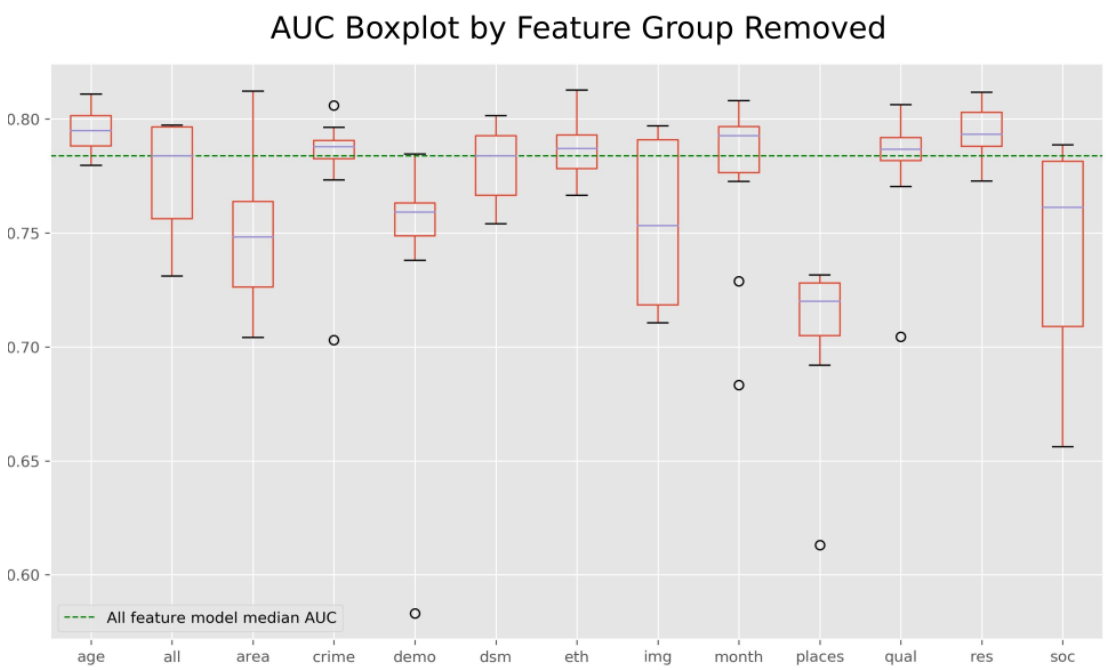


Figure 13. Box plot of AUC for models trained on datasets with features removed. The

x-axis shows the feature removed, except for 'all' which is the model trained with the original set of features.

3.4 Final Model

The highest AUC from the ablation study on the validation set was 0.8128 and was achieved by a model where the ethnicity feature was removed. The final model ROC for validation and test set of this model is presented in Figure 14. The threshold of highest classification performance, observed to be 0.654, was found by taking the threshold at which the sum of sensitivity and specificity (1-FPR+TPR) was at its maximum. A map of confusion matrix components at this threshold is presented in Figure 15. The threshold at which FN=FP = 0:047, the cost-benefit ratio, was 0.093. Tables 3 and 4 show the confusion matrices for highest class separation and cost-benefit thresholds, respectively. The lowest validation loss weights for this mode were saved after 83 epochs at 0.2113 validation loss before model training was dropped out after 103 epochs. Figure 15 summarises the accuracy and loss over the training and validation data sets throughout training. Diagrams summarising the CNN branches and network architecture are shown in Figure 9.

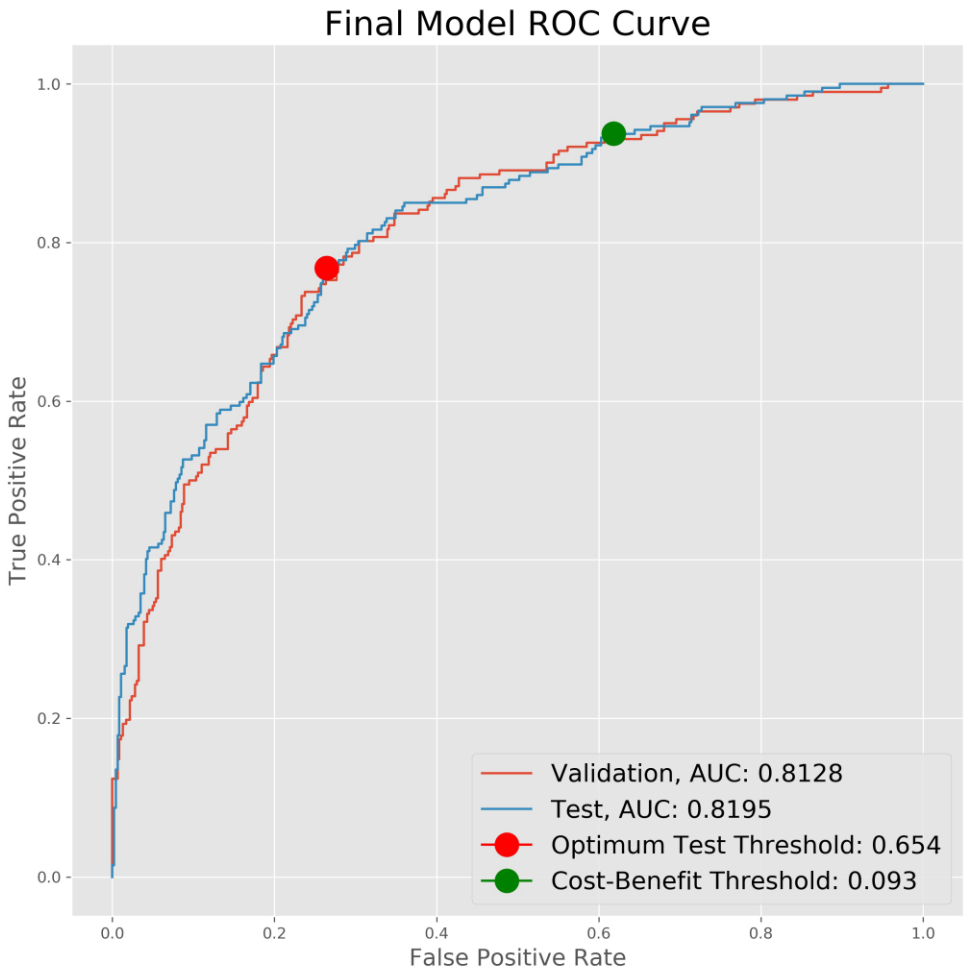


Figure 14. ROC curve for final model validation and test classification performance showing positions of optimum and cost-benefit thresholds

Table. 3 Test set confusion matrix for the threshold of max class separation

		predicted	
		positive	negative
actual	positive	129	59
	negative	106	375

Table. 4 Test set confusion matrix for the cost-benefit threshold

		predicted	
		positive	negative
actual	positive	194	14
	negative	285	176

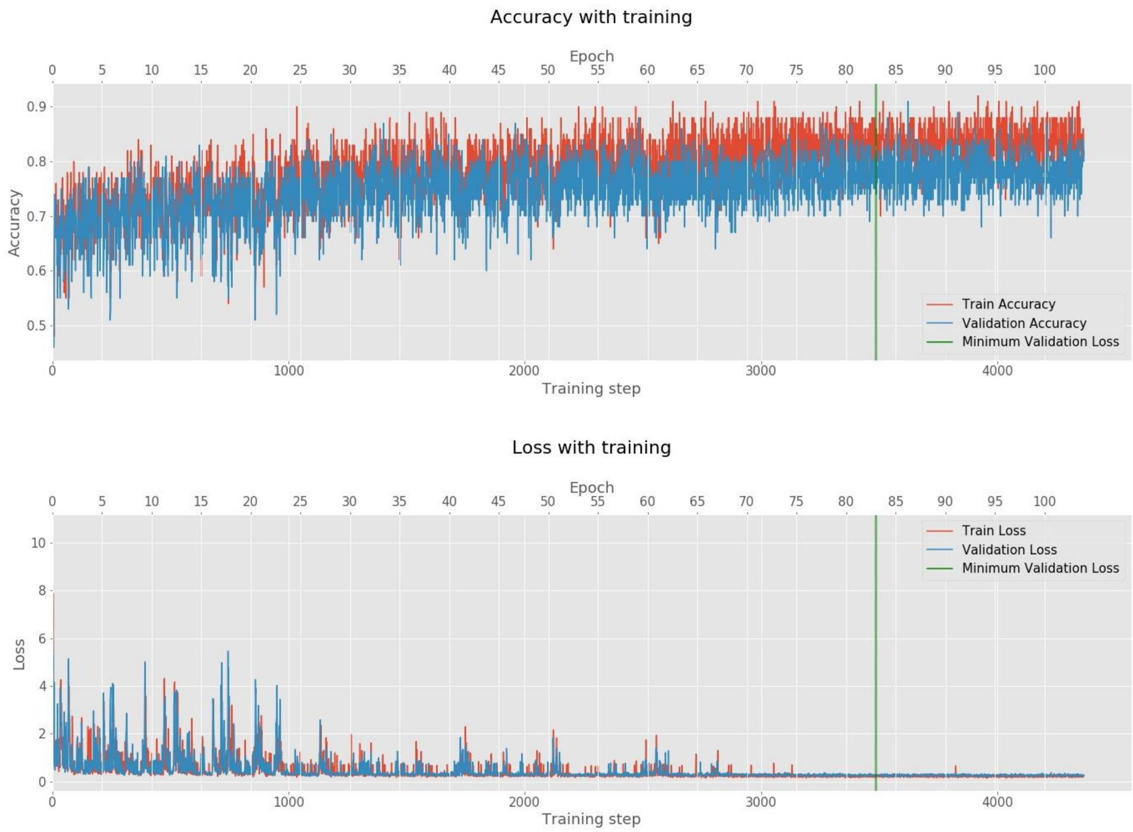


Figure 15. Training summary for the final model

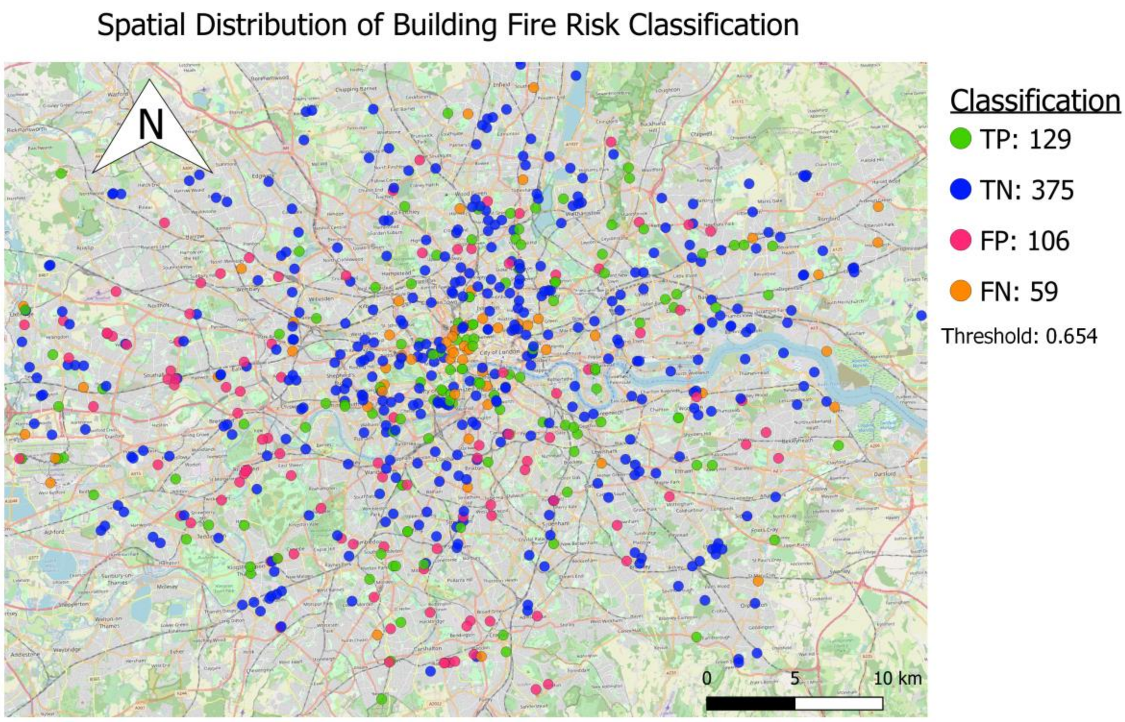


Figure 16. Spatial distribution of test set confusion matrix classification performance terms for a threshold of 0.654. TP represent the number of true positives in the dataset, TN represent the number of true negative in the dataset, FP represent the number of false positives in the dataset, FN represent the number of false negatives in the dataset

4. Discussion

4.1 Shapley Values

The results from the Shapley value investigation represent the impact of feature values on a model output.

4.1.1 Building Area

The building area feature had a higher impact on model output than any other feature from the tabular data. Similar results have been found by Madaio et al., (2016) who point out that many of the most important features from their modelling are related to building size for their Random Forest model. While one would expect larger buildings to be generally more susceptible to fire risk as they have potential for more activity to take place within them, it is difficult to conclusively say that this is the reason for the observed correlation. The subject is complicated further by the fact that incident location is not accurate for all building incidents in the LFB data. It may be that there is a greater driving force behind the accurate collection of data relating to larger buildings within which incidents with more casualties or a greater cost of damage may occur. This would lead to larger buildings being over-represented within the positive fire incident class and give a skewed representation of reality.

4.1.2 Places

Generally, the places features had a large impact on model outputs. This is a significant finding as it solidifies the idea that the types of activities that occur within a non-residential building have an impact on the buildings fire risk. This agrees with the regression coefficients described by Madaio, et al. (2016). A higher building fire risk from food establishments may be explained by the cooking activities that are carried out in these establishments as cooking equipment have been shown to be one of the major sources of building fire ignition (Shai, 2006). This is also supported by Manes and Rush (2018), who found that fire incident rate for food premises in the UK were 1.8 times higher than any other category. It is interesting to note that while the SHAP values for the presence of food vendors were positive, they were negative for the presence drinking (alcohol) establishments. These results suggest that useful fire risk predictors can be made through a distinction between these categories.

The presence of medical services within a building also leads to a higher SHAP value. This pattern is seen in the results of Manes and Rush (2018) whose statistical analysis of building fires in the UK from 2014-15 find that fires in hospitals occurred in 2% of all buildings of that category within the sample, the second highest incidence rate of the categories they observe. High SHAP values have been found to be associated with the presence of offices in buildings in this study. This contrasts with results of Manes and Rush (2018) who find that offices have a 0.3% incidence rate, second lowest to dwellings.

4.1.3 Age

For all age group features, generally higher feature values are associated with a decrease in the model output. This seems at face value rather counter-intuitive as these features are proportions, and it was expected that higher fire risk would be associated with higher proportions of some age group. There does appear to be some minor trends, however.

For the age group 0-16, representing children, the extremes of this feature were associated with a higher magnitude impact on the model output than the other age features. In particular, the absence of children in the population is associated with the highest model output of age features.

4.1.4 Unemployment, Qualification and Social Class, Ethnicity and Crime

The SHAP values for unemployment show a narrow spread suggesting that there is little impact of employment rates on model output and virtually no contribution to fire risk. This finding is concordant with the results found by Špatenková and Stein (2010), where no link was found between unemployment and fire incidence. A higher proportion of single occupants in the population of an area lead to an increase in SHAP value. A similar finding is also seen in population density.

Of the qualification features, level 4 qualification had the potential to have the greatest impact on SHAP values. Higher proportions of level 1, 2, 3 and 4 qualified populations led to a negative model impact, while higher proportions of population with no qualification lead to a slight positive impact on model output. These results may suggest that education levels in the population have an impact on non-residential fire incidence.

Higher social class feature values all have a positive impact on model output. The largest impact on model output is seen from the A/B social grade which reaches up to around 0.3 and as low as -0.2 SHAP value. Grades C1 and D/E feature values have a moderate impact on model output, while C2 has little impact.

All ethnic population features have a very low impact on model output when compared to other features analysed. This suggests that they are not a good predictor of non-residential building fire and may cast doubt onto their use in previous studies (Madaio et al., 2016), however the impact is likely to vary for different cities.

While high burglary crime values were associated with an increase in model output, the opposite was the case for violent crime where a few high feature value outliers lead to SHAP values around 0.6. This suggests both features may be effective contributors to model performance.

4.2 Ablation Study

4.2.1 Places

It was found that the places data was the largest contributor to the classification performance of the dataset. The median AUC decreased by 8% when the places features were removed, higher than any other feature group removal. This finding also reflects the result of the Places SHAP values. This suggests that, in this model, building use is the most crucial factor to consider when classifying non-residential fire risk and outweighs the importance of social factors. This also highlights a major difference in analysis of building fire risk factors between residential and non-residential properties.

4.2.2 Building Area

The second highest feature removal decrease in median AUC was from the area feature, with a 5% decrease in median AUC. While the area features removal tended to improve the model, it is worth noting that the second highest performing model from the ablation study had this feature removed at over 0.81 AUC. While it tends to improve the model there is a possibility that removing the area feature can improve classification performance. It may be the case that, while the area feature does provide a useful indicator of potential fire risk, the model has the potential to rely on it heavily as an individual feature. When this is related to the SHAP value results, where the area features impact on model output had the potential to be over 30 times greater than any other feature, it can be seen how the area feature could overshadow the values of other features when making a prediction.

4.2.3 Demographic Features

When all demographic features were removed the median AUC decreased by 3% suggesting that social factors contribute somewhat to classification performance and are useful indicators of non-residential fire risk. When the sub-groups are analysed, however, the removal of only the social classes group leads to a decrease in median AUC. Other demographic variables, when treated as their own feature groups, do not provide any increase in classification performance.

4.2.4 Aerial Imagery and DSM

Ablation was also performed on the aerial imagery and DSM CNN branches. It must be noted that while the other ablation results represent the removal of tabular features from the vector branch of the model, the imagery and DSM ablation results are collected by removing the entire CNN branch of that feature. While the architecture of the model changes slightly with the tabular feature ablation in that the dimension of the inputs change, the architecture is changed more profoundly by removing an entire CNN branch. When the imagery branch was removed there was a decrease in median AUC 6.3. Final Model Performance 47 by 4%, the third highest decrease seen in the ablation study. This suggests that the aerial imagery component of the model is an effective contributor to classification performance. Conversely, the median AUC increased by 0.002% when the DSM channel was removed, suggesting that it contributed little improvement to classification performance and was not as useful as the imagery branch. There are some fire risk factors of non-residential buildings that can be deciphered visually. These results do not suggest a mechanism for what characteristics of a building image the model uses to come to more accurate classifications. For instance, it may be that like Liu et al. (2017), the model is able to visually assess the construction and maintenance quality.

4.3 Final Model Performance

It was found that the removal of the ethnicity feature yielded the highest classification performance with a validation set AUC of 0.8128 and a test set AUC of 0.8195. An optimum threshold was found to be at 0.654 where a TPR of 0.768 and FPR of 0.264 was achieved (Figure 14). The final TPR vastly exceeding the FPR suggests that this classifier has good potential to focus inspection efforts to buildings of high risk.

4.4 Classification Spatial Distribution

Figure 16 shows the spatial distribution of confusion matrix terms from a classification at optimum class separation. Generally, the spatial distribution of the test set is like that of the entire dataset (figure 6), with a higher concentration of examples in central London. Although true positives are found in most regions of the study area, there is some clustering of these values around central London.

4.5 Cost-Benefit Analysis

This study has demonstrated how the classification threshold may be moved to meet a cost-benefit efficiency level for building inspections. It can be seen in Figure 14 and Table 3, that the cost-benefit threshold on the final model manages to correctly classify 93.2% of instances of building fire in the test set, an increase of 36% when compared to the optimal threshold. This does come at a cost, however, as the threshold would then incorrectly classify 65.7% of negative examples as being instances of building fire, an increase of 180% when compared to the optimal threshold. When the sum of sensitivity and specificity (1-FPR+TPR) for the two thresholds are compared it can be seen that the optimal threshold, at 1.466, is 12% higher than the cost benefit threshold, at 1.314. This suggests that, although such an approach enables certain operational requirements to be met, it is by no means the best performing classifier overall.

For instance, municipal fire departments may implement systems whereby, instead of an average area of building, the inspection cost value could be calculated using the building area of individual cases. Furthermore, the unit area cost of non-residential buildings varies between occupancy types so an occupancy dependent benefit value may be implemented (Salter et al., 2013).

5. Conclusion

This study has presented a non-residential building fire risk prediction methodology based on a hybrid CNN-MLP approach and assessed the effectiveness of some features commonly used in the literature along with novel image features that were previously unexplored. Three key conclusions have been made through this investigation: i) while classification performance may be improved by including an aerial imagery feature of the building to the model via a CNN branch, the inclusion of 1m GSD DSM data to the model showed no improvement, ii) data relating to building use had the greatest impact on classification performance, while demographic data, apart from that regarding social class, did not lend benefit to the model. Such a finding is significant as some existing studies have used a suite of demographic features, iii) spatial analysis of final model classifications suggest that models built over large regions may lead to areas of poor model performance.

As future perspectives, whilst the findings of this study answer some questions about use of specific features in building fire risk models, it raises many more about the future of building fire risk classification. As aerial imagery has been found to benefit building fire risk classification, future work should be concerned with investigating other novel features that hold contextual building information.

While the 1m DSM feature was not beneficial to the classification in this study, there is insufficient evidence to suggest that building geometry is completely irrelevant. Future work should experiment with DSMs at higher resolution to assess any classification potential before it is ruled out entirely. A more rigorous ablation study, potentially assessing a wider range of features, in all combinations would yield more conclusive evidence regarding which features are relevant. Furthermore, a comparison of classification models built with the same features for different locations should be explored. Future work should investigate whether temporal evolution of features can aid classification.

Acknowledgements: We thank two anonymous reviewers whose comments and suggestions helped improve on earlier drafts of the manuscript.

References

- Alidoost F., and Arefi H. A CNN-based approach for automatic building detection and recognition of roof types using a single aerial image. PFG – Journal of Photogrammetry, Remote Sensing and Geoinformation Science, 86:235–248, 12 2018.
- Audebert N., Herold C., Slimani k., and Vidal C. Multimodal deep networks for text and image-based document classification. arXiv:1907.06370 [cs], 07 2019.
- Bryant S., and Preston I. Focus on trends in fires and fire- related fatalities, 2017.
- Chigozie E., Nwankpa W., Ijomah A., Gachagan S., and Marshall. Activation functions: Comparison of trends in practice and research for deep learning, 2018.
- Clare J., Garis L., Plecas D., and Jennings C. Reduced frequency and severity of residential fires following delivery of fire prevention education by on-duty fire fighters: Cluster randomized controlled study. Journal of Safety Research, 43:123–128, 2012.
- Collins, L., McCarthy, G., Mellor, A., Newell, G., & Smith, L. (2020). Training data requirements for fire severity mapping using Landsat imagery and random forest. Remote sensing of environment, 245

- DaCosta, M., Krinsley, J., & Abelson, B. (2015). Optimizing Local Smoke Alarm Inspections with Federal Data. Bloomberg Data for Good Exchange (2015).
- Edina. Digimap. https://digimap.edina.ac.uk/, 2019.
- Environment Agency. Lidar composite DSM 2017 - 1m. https://data. gov.uk/dataset/80c522cc-e0bf-4466-8409-57a04c456197/ lidar-composite-dsm-2017-1m, 03 2020.
- Garis L., and Clare J. A dynamic risk–based framework for redesigning the scheduling of fire safety inspections, 2014.
- Getmapping. Getmapping — uk aerial photography. https://www. getmapping.com/, 2019.
- Geiß, C., Schrade, H., Aravena Pelizari, P. and Taubenböck, H.: Multistrategy ensemble regression for mapping of built-up density and height with Sentinel-2 data, ISPRS J. Photogramm. Remote Sens., 170, 57–71, doi:10.1016/j.isprsjprs.2020.10.004, 2020
- Google. Overview — places api. https://developers.google.com/places/web-service/overview, 2020.
- The Home Office. Detailed analysis of fires attended by fire and rescue services. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/831136/detailed-analysis-fires-attended-fire-rescue-england-1819-hosb1919.pdf, 2018.
- The Home Office. Fire statistics definitions incident recording system (irs). https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/610453/fire-statistics-definitions.pdf, 2020.

- Hong S.,and Jeong S.R. Development and comparison of data mining-based prediction models of building fire probability. Journal of Internet Computing and Services, 19:101–112, 12 2018.
- Kim, D., Liong, S.-Y., Gourbesville, P. and Liu, J.: An Innovative DEM Improvement Technique for Highly Dense Urban Cities, pp. 229–240, Springer, Singapore., 2020a.
- Kim, D. E., Liong, S.-Y., Gourbesville, P., Andres, L. and Liu, J.: Simple-Yet-Effective SRTM DEM Improvement Scheme for Dense Urban Cities Using ANN and Remote Sensing Data: Application to Flood Modeling, Water, 12(3), 816, doi:10.3390/w12030816, 2020b.
- Krizhevsky A., Sutskever I., and Hinton G.E. Imagenet classification with deep convolutional neural networks. 2012.
- Kuester K., Mittnik S., and Paolella M.S. Value-at-risk prediction: A comparison of alternative strategies. Journal of Financial Econometrics, 4:53–89,01 2006.
- Law S., Brooks Paige, and Chris Russell. Take a look around: Using street view and satellite images to estimate house prices. ACM Transactions on Intelligent Systems and Technology, 18, 2018.
- Law S., Seresinhe C.I, Shen Y., and Gutierrez-Roig M. Street-frontage-net: urban image classification using deep convolutional neural networks. International Journal of Geographical Information Science, 34:681–707, 12 2018.
- LeNail A., NN-SVG: Publication-ready neural network architecture schematics. Journal of Open Source Software, 4:747, 01 2019.
- Liu L., Silva E.A, Wu C., and Wang H., A machine learning based method for the large-scale evaluation of the qualities of the urban environment. Computers, Environment and Urban Systems, 65:113–125, 09 2017.

- London Data Store. Recorded crime: Geographic breakdown – London datastore. https://data.london.gov.uk/dataset/recorded_crime_summary, 2019.
- London Fire Brigade. London fire brigade incident records – London datastore. https://data.london.gov.uk/dataset/london-fire-brigade-incident-records, 2011.
- London Fire Brigade. Firefighters or officers visiting your building. https://www.london-fire.gov.uk/about-us/services-and-facilities/techniques-and-procedures/firefighters-visiting-your-building/, 2020.
- Lundberg S.M., and Lee S.I. A unified approach to interpreting model predictions, 2017.
- Madaio M., Chen S.T., Haimson O., Zhang W., Cheng X., Hinds-Aldrich M., Chau D., and Dilkina B. Firebird: Predicting fire risk and prioritizing fire inspections in Atlanta, 2016.
- Manes M., and Rush D. A critical evaluation of bs pd 7974-7 structural fire response data based on USA fire statistics. Fire Technology, 55:1243–1293, 10 2018.
- Mitri G., Jazi M., and McWethy D. Assessment of wildfire risk in Lebanon using geographic object-based image analysis, 06 2015.
- Monfort D., Negulescu C., Belvaux M., Remote sensing vs. field survey data in a post-earthquake context: Potentialities and limits of damaged building assessment datasets, Remote Sensing Applications: Society and Environment, Volume 14, 2019, Pages 46-59, ISSN 2352-9385,
- Office for National Statistics. 2011 census - office for national statistics. https: //www.ons.gov.uk/census/2011census, 2011.
- Office for National Statistics. Census geography - office for national statistics. https://www.ons.gov.uk/methodology/geography/

ukgeographies/censusgeography, 2016.

- Oliveira S, Zêzere JL. Assessing the biophysical and social drivers of burned area distribution at the local scale. J Environ Manage. 2020 Jun 15; 264:110449
- Ordnance Survey. Os mastermap topography layer — Great Britain's landscape vector map data. https://www.ordnancesurvey.co.uk/business-government/products/mastermap-topography, 2020.
- Phan H., Yamamoto K., Phan T., and Yamamoto K. Resolving class imbalance in object detection with weighted cross entropy losses, 2006.
- Preuschoff K., Quartz S.R., and Bossaerts P. Human insula activation reflects risk prediction errors as well as risk. Journal of Neuroscience, 28:2745–2752, 03 2008.
- Pringle P., and Welsh B. Times investigation: Thousands of large L.A. buildings are long overdue for fire inspections. Los Angeles Times, 06 2015. PyTorch. Pytorch. https://pytorch.org/, 2019.
- Salter C., Ramachandran G., Emmitt S., and Bouchlaghem N. Economic cost of fire: Exploring UK fire incident data to develop a design tool. Fire Safety Journal, 62:256–263, 11 2013.
- Shai D. Income, housing, and fire injuries: A census tract analysis. Public Health Reports, 121:149–154, 03 2006.
- Shapley L.S. Stochastic games. Proceedings of the National Academy of Sciences,
  39:1095–1100, 10 1953.
- Sheng V., and Ling C. Thresholding for making classifiers cost sensitive, 2006.
- ˇSpatenkov´a O., and Stein A. Identifying factors of influence in the spatial distribution of domestic fires. International Journal of Geographical Information Science, 24:841–858, 04 2010.

- Schunder T., Yin D., Bagchi-Sen S., Rajan K., A spatial analysis of the development potential of rooftop and community solar energy, Remote Sensing Applications: Society and Environment, Volume 19, 2020, 100355, ISSN 2352-9385,
- Wang H., Zhao T., Li L.C., Pan H., Liu W., Gao H., Han F., Wang Y., Qi Y, and Liang Z. A hybrid CNN feature model for pulmonary nodule malignancy risk differentiation. Journal of X-Ray Science and Technology, 26:171–187, 01 2018.
- Zhuang J., Payyappalli V., Behrendt A., and Lukasiewicz K. Total cost of fire in the united states final report by, 2017.