Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.39276

Applications of Explainable Machine Learning to Casualty Risk Prediction in Residential Building Fires: Current Status, Roles of Fire Protection Systems, and Future Directions

Zhongyi Niu

Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu, China
18223060338@163.com

Abstract. The conditions that lead to casualties in residential building fires are a combination of several factors: vulnerability of the occupants, fire origin and spread, response to fire alarms, fire-service response and the effectiveness of protection systems. In this paper, explainable machine learning for casualty and loss prediction, interpretation of predictions, and active fire protection are discussed. The mixed variables and nonlinear relationships of the incident records can be addressed by tree models and ensembles. Training data distribution may change with resampling, which may help determine minority casualty classes. Neural networks are able to learn and model spatio-temporal patterns, but require larger datasets and more rigorous external validation. Nor do feature importance, nor SHAP, give causality. Another study has also been linked to a reduction in the risk of deaths when smoke alarms are present. Occupants with limited mobility may be protected through automatic suppression such as detector activated systems that do not need an occupant's response. However, research is hampered by class imbalance, regional variation, limited information on the status of the system and reliance on counterfactual estimates. Issues that need attention are consistent casualty categories, consistent evaluation metrics, validation across regions, use of model explanations in fire-safety decisions, and system reliability and the needs of different population groups.

Keywords: Residential building fires, Casualty risk prediction, Explainable machine learning, Class imbalance, Automatic fire suppression systems

1. Introduction

Casualties in a residential fire are usually attributed to a single cause. A working alarm can not wake a sleeping person. After hearing it, a person with limited mobility might not be able to get away fast enough. The location of the ignition, combustibles, door and window locations affect the growth of the fire and the spread of smoke. The time of fire-service arrival in turn dictates the time of onset of external rescue. While they are recorded in separate fields on investigation forms, they must be taken in as a whole, so as to understand and appreciate the impact on the outcome in terms of the occupants, buildings, protection systems and response processes.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.39276

Early research relied mainly on statistical methods to relate individual factors to casualties. These methods can provide results that are easily interpreted, but they also rely on fairly strong assumptions when continuous and categorical variables, missing data, and nonlinear interactions all need to be taken into account. The prediction of property loss and casualties severity and occurrence, and fire spread in residential fires have been increasingly investigated using machine learning in recent years [1-4]. Another area of research is the use of a time series of sensor data obtained from an experiment or a simulation to predict the indoor temperature and flashover events so that more time can be used in evaluating the fire development [5, 6]. The incident records and experimental sensor data are used for different purposes in these studies. These numbers do not alone, however, give a true representation of their accuracy.

There are three interrelated strands of research. High risk incidents are identified using decision trees, random forests, gradient boosting and neural networks. To explain the resulting predictions, feature importance and SHAP are used. Research into smoke alarms and automatic suppression follows to determine if smoke alarm and other automatic suppression systems can save lives. In this work, emphasis has shifted from correctness of the overall result towards identification of minority casualty classes, interpretation of model output, validation on external data, and an assessment of the feasibility of intervention.

Immediate issue: class imbalance—incidents without casualties are common, incidents with severe casualties are uncommon. Many different types of non-casualty incidents can be classified correctly, resulting in high overall accuracy for the model despite a low number of correctly classified casualty cases. So, evaluation should include recall, precision, F1, macro-averaged F1, confusion matrix for casualty classes, a decision threshold and test-set class proportions. The discussion here follows predictive performance to model interpretation and fire-protection systems. It does not discuss the relative merits of each approach, but rather the contexts for their use, the strengths and weaknesses, and the limits of the evidence, without placing a hierarchy on the best reported values on different datasets.

2. A machine learning approach leads to the prediction of the risk of a fire in a residence

2.1. Model the behaviour of the original data

Decision Trees work with both continuous and categorical variables and produce fairly easy-to-interpret classification rules. However, with changing the sample, a single tree is easily overfitting. Random forests is a way of combining trees to mitigate variance. AdaBoost, XGBoost, and CatBoost, however, fix mistakes step-by-step, and they are typically applicable to structured fire-incident data. Liu and Zhuang made a comparison between the tree-based models for property loss and casualty predictions using the UK Home Office Residential Fire Dataset [1]. It relates to fire in ordinary dwellings, sheltered accommodation, caravans, houseboats and houses in multiple occupation in England. The samples prior to resampling are known as the original data. The results for these data were similar for all models: accuracy for property loss was around 86% and for casualties it was about 93%, while the maximum mean values of F1 did not exceed 61% and 51% respectively [1]. The casualty accuracy of 93% thus masked a F1 of only 51%. The more common ordinary incidents were primarily classified correctly while the models still had difficulties with the minority casualty class.

Incident records are typically from fire-department investigation forms. They can document the cause of the fire, where it started, and the times for the alarm and arrival, casualty outcomes, detector status, and suppression-system activation. Schmidt et al. analysed over 48, 000 building

DOI: 10.54254/2755-2721/2026.39276

fires that occurred in Oregon, United States, from January 2012 to August 2023, including 2, 136 fires resulting in civilian casualty and 317 confirmed fatalities. They attached the incident files with neighbourhood demographic, socioeconomic and infrastructure information [2]. Ahn et al. used 34 variables that included building characteristics, land characteristics, and demographic characteristics, and used 16 algorithms to create a five-level fire-risk model. Approximately 22% of the buildings were classified in the highest-risk category (54% of the fires that actually occurred [3]). Although these records reflect real events, such findings are still impacted by variations in investigation practice, missing values, and recording bias. The lack of an entry, for instance, does not prove that a protection system was not in place.

2.2. The changes made due to the re-sample

There are several strategies for dealing with class imbalance: undersampling, oversampling, class weighting, and tuning the decision threshold. Undersampling discards samples in the majority class, decreasing the training expense while sacrificing information on common occurrences. SMOTE generates artificial samples in the vicinity of the minority class to enable a model to learn the decision boundary. Those examples do not necessarily denote actual events. Liu and Zhuang used random undersampling along with SMOTE and achieved a mean F1 score of 75% for property loss and 77% for casualty, which is 14 and 26 percentage points higher than the original data [1]. Minority-class recognition was improved on that data set. The applicability of similar benefits in other cities has yet to be proven.

Resampling should be done within the training folds. Validation and test sets should also reflect the true class distribution or else the performance in deployment will be over estimated. Recall rates and false-positive rates should be reported for each treatment and comparisons should include class weighting, threshold shifting, and cost-sensitive learning. Liu et al. examined the fire spread characteristics in over 100,000 fire records from UK residential fires in the years 2015–2024. The proportion of severe cases was approximately 7% of the records. Class overlap was solved by using Tomek links and SHAP was used to investigate factors such as alarm delay and detection time [4]. The selection of imbalance treatment should thus be based on the task labels, number of samples, and real expense of misclassification.

2.3. Estimating the seriousness of injuries

Before creating an ensemble classifier with Bayesian-regularized neural networks, Schmidt et al. studied the variable importance with a random forest. The accuracy of casualty-severity classification was 92.5% when detailed incident information was used, and 87.6% when only neighborhood-scale spatial variables were considered [2]. This gap of 5.1 percentage points illustrates the added value of information about the scene post-ignition for severity assessment. Neural networks can learn complicated relationships between multidimensional variables. However, they have higher requirements for sample size, hyperparameter tuning, and external validation, and provide results that are less intuitive than the rules of a single decision tree.

Graph neural networks and deep learning can be applied to sensor time series, which can be used to model heat transfer in space and time. Zhang et al. conducted room scale experiments with a low number of sensors to predict fire evolution. Their model could provide about 20 seconds warning prior to flashover and had prediction errors of less than 10% in the temperature predictions [6]. Tam et al. implement a spatial-temporal graph neural network (FlashNet): rooms or sensors are represented as nodes, while spatial relationships are represented as graph edges. The study

DOI: 10.54254/2755-2721/2026.39276

condition, it predicted flashover approximately 30 seconds before with an overall accuracy of ~92.1% [5]. These models are appropriate for real-time evaluations of fire evolution. But experiments and simulations can only account for some of the variability in homes with doors and windows open, fuel configurations, sensor malfunctions and occupant behavior. Testing still needs to be done on a large scale prior to deployment.

2.4. Comparability across studies

The percentages reported of 93%, 92.5%, and 92.1% are for studies with different labels, region, sample composition, and error cost. The first two describe casualty occurrence and casualty severity and the third describes flashover prediction. The consequences of missing a fatal incident are also different than missing a few seconds of warning. Task definitions, minority-class proportions, data partitions, external validation and confidence intervals for performance metrics are all important considerations in comparing studies. Comparisons of the same study can often be more useful, such as differences in F1 between re-sample and original sampling, or differences in performance when the scene variables are removed. One maximum accuracy from each paper is not as telling. Tree models are appropriate for incident records in tabular format, and calculations of feature-importance are easy. Neural networks are better for complex spatial-temporal data and need more of them. Choice of model must be based on the data and intended use; there is no single best model.

3. Explainable machine learning and risk factors

3.1. The accuracy to risk interpretation

After an incident has been determined to be high risk by a model, fire-safety managers must know why. The measure of random-forest permutation importance is the decrease in performance when the variable is randomly permuted. Importance built in is nice to compute, but can be positively biased towards variables that have more possible values. SHAP is based on cooperative game theory to assign the difference between a prediction and a baseline value to the features. It offers worldwide rankings and explanations of the individual predictions [7]. In the case of residential fires, this may give clues to conditions that tend to make the outcome more risky and why a specific fire was rated as a high-risk fire. In order to interpret those explanations, you also need the background sample, output scale and correlations between the features.

3.2. Important factors in casualty models

Factors such as the victims age, fire-service response time and the proper use of smoke or fire detectors were significant determinants of fatal outcomes in Schmidt et al. analysis [2]. In Liu et al.'s work on residential fire spread in the UK alarm delay and detection time also played a prominent role [4]. The variables relate to each aspect of the problem – human vulnerability, external rescue and early discovery. They can direct Data Quality checks and Intervention Priorities conversations. They also illustrate that the structure of buildings is not in itself a predictor of casualty risk. However, there are complications when it comes to correlations. If age is highly correlated with mobility then a model can re-distribute contributions between age and mobility, or building age to system maintenance. Specific cases should be reviewed to assess these relationships.

DOI: 10.54254/2755-2721/2026.39276

3.3. Important features are not causes

Feature importance and SHAP describe the information a model uses. Both by themselves are not evidence of a causal relationship between a variable and casualties. Response time can be an indicator of population, road conditions, and fire-station distribution all in one. Home maintenance, income and building age may also affect detector status. To establish causality, there needs to be more rigorous study design, control of confounding, or a natural experiment. It is, therefore, more accurate to refer to a feature as associated with model predictions, rather than as an incident cause, or as contributing strongly within the model. If the permutation importance and SHAP align with the incident reviews then the explanation is more credible. When there are significant discrepancies between their rankings, this should be made clear.

4. Systems for identifying fires and automatic suppression

4.1. Smoke alarm study results

In their systematic review of smoke alarms and casualties, Rohde et al. found 13 studies. Four contributed to the fatality-rate meta-analysis and three to the meta-analysis of injury rate. The pooled rate of fire fatalities decreased by nearly 50% in homes with operative alarms when compared to homes with no-operative alarms, while there was no significant difference in injury rates [8]. The primary purpose of an alarm is to alert the building occupants earlier and provide more time for them to evacuate. However, getting to safety requires hearing, understanding, deciding and acting on the warning. Whether or not that sequence results in effective action depends on the acoustic context, sleep or cognitive condition, physical capabilities and immediate circumstances [9].

The introduction of smoke-alarm requirements in dwellings in German states varied. In 12 of 16 states there were statistically significant decreases in the risk of residential fire fatalities [10]. Official fatality reports, however, failed to indicate whether or not an alarm had been installed and had operated on each occasion. The effective date of the regulation would only be a rough indicator of device coverage. Therefore, the installation, operation and successful warning of the occupant should be documented separately.

4.2. Automatic suppression systems for potential effectiveness

Runefors et al. examined 144 unintentional household fires with fatalities that occurred in Sweden during the period 2011–2014 where each case has been individually analyzed. They estimated that heat-activated suppression systems might have prevented 68% of deaths, detector-activated suppression in bedrooms and living rooms 59%, and the ordinary smoke alarms 37% [11]. Among these measures, the systems directly controlling the fire had the highest estimated value. These are theoretical potential effectiveness figures for fatal cases, which already happened. They are NOT permanent death rate cuts that will result upon installation. A follow-on study of fatal residential fires in Sweden during 1999-2018 revealed that preventive measures effectiveness varied by age, living conditions and care setting [12]. Therefore, protection should be tailored to the population in question.

DOI: 10.54254/2755-2721/2026.39276

4.3. The same system is not suitable for everyone

Smoke alarms were estimated to be 14% effective and home suppression 31% effective for smokers receiving home care. Bedroom and living room detector controlled suppression was achieved at 88% [11]. The distinction is due to activation timing and dependency on the occupant. Ignition is frequently near a seat, bedding or the body. The local hazard may occur before sufficient heat reaches the conventional sprinkler to cause it to operate. A person with a mobility impairment may hear an alarm but not be able to move or extinguish a fire quickly. A linked system is capable of starting local suppression automatically upon the detection of smoke or fire, without having to rely upon the occupant to hear, understand and take action. This will help to reduce the time interval between discovery and intervention. Despite this, 88% is a hypothetical value of a specific sample. It cannot be used as a true indication of the protection rate of all home care recipients.

4.4. Interpretive limits of evidence pertaining to fire protection systems

Protection system evidence is provided by a variety of types of studies. Detectors play a major role in terms of casualty estimates [2]. Meta-analysis correlates the operative alarms with the reduction of fatality rates [8] and case-based counterfactual studies estimate the effectiveness of the different systems [11]. These findings are generally in the same direction, albeit they are answering different questions. One relates to the information provided to the prediction model. A third compares results of an operative alarm to a similar alarm not in use. The third asks how many people would have died if a certain system had not been in place. The results do not represent one, consistent system effect and an association do not imply causality. Reports should clearly indicate where their data comes from, their comparison group, their definition of effect and uncertainty.

5. Existing research problems

5.1. When the accuracy is high, the minority class may be hidden

The highest mean F1 for the casualty task was 51% but the accuracy for the original data was about 93%. When random undersampling was used in conjunction with SMOTE, the mean F1 score jumped to 77% [1]. The contrast shows the effect of the many non-casualty events affecting the overall accuracy. Residential fire models should be able to report recall, precision, F1 and the confusion matrix for casualty classes at a minimum. Resampling is allowed to be done within the training set, but not in validation and testing where the actual proportions of classes must also be kept. In the absence of these protections, researchers are unable to determine how many severe incidents a model would miss or estimate how many it would generate in use.

5.2. Importance of regional differences

With neighborhood spatial variables alone, the accuracy of the Oregon model was 87.6%, and the Oregon model's accuracy with incident-scene information was 92.5% [2]. This discrepancy is related to a region and a task and is not easily transferable to other cities. Nor does the estimated potential effectiveness in 144 fatal Swedish fires of 68% represent a general mortality reduction as a result of suppression-system installation in ordinary homes [11]. A second challenge with data is splitting it up. Similar incidents from the same region and contiguous years might be divided into both training and test sets, making the test too easy. The transferability of whole cities, years, or fire jurisdictions, provides a more rigorous evaluation of transferability.

DOI: 10.54254/2755-2721/2026.39276

5.3. It's not simply the presence of a system that matters

The installation, the status of an alarm at the time of a fire, the activation and the success in warning the occupants should be recorded individually in a database. The German regulatory study was unable to actually make a measurement of the alarm coverage, so instead it used the date on which the requirement took effect [10]. That restricted the conclusions that could be drawn as to the mechanism. Activation times and actual fire-control performance are also required for suppression records. Response time should be broken down into discovery, alarm, dispatch and arrival. An installed device may be ineffective due to missing batteries, power loss or inadequate maintenance. This simple "present or absent" field can thus cause a model to produce excessive estimates of protection, and can leave policymakers without a clear sense of whether extra protection is needed through increased installations or through improved maintenance.

6. Future research directions

Research requires standardized casualty categories, metrics for evaluation, and data dictionaries, with the latter being validated outside the city and year. System installation, its operating status, activation, and effective control should be differentiated in incident records. They should also contain the age, mobility, dwelling type, ignition time and room of origin where privacy regulations permit. Training should prevent spatial and temporal leakages. Class-imbalance treatments should only be used in the training data and reports should include decision thresholds, confidence intervals, and probability calibration. However, SHAP can still be applied to review predictive reasoning with a background sample defined. When used in conjunction with statistical analysis, sensitivity tests and individual case reviews, it can help to prevent the premature conversion of significant features into incident causes. Older people living alone, people with limited mobility and smokers receiving home care should have top priority for automatic suppression that does not require an occupant's response. Validation of real-time prediction systems is required in fully scale fires, with sensor failures or with anomalous data. The de-identified code and variable dictionaries and data-splitting rules could be made available to help others replicate results and use them in equipment maintenance, fire-service dispatch, and targeted interventions.

7. Conclusion

Fire risk models for the home need to address the combination of rare casualty events, complex relationships among variables, and regional variation. Incident records are suitable for tree models and ensembles. As long as the class distribution in the test is not changed, resampling can also increase the recognition rate of minority classes. Although spatial-temporal information can be incorporated in neural networks and graph models, it requires the real-fire data and region validation. Model explanations facilitate the easier investigation of predictions and are not necessarily causal explanations. Protection is also reliant on the system being functioning correctly and occupants having the ability to respond, making automatic intervention even more relevant to those who are not as mobile. Detailed records combined with an external validation, reliable interpretation and an awareness of the various population groups, would help translate predictions to actual decisions for equipment maintenance and fire prevention.

DOI: 10.54254/2755-2721/2026.39276

References

[1] Liu, Z., & Zhuang, Y. (2024). An investigation using resampling techniques and explainable machine learning to minimize fire losses in residential buildings. *Journal of Building Engineering, 95*, 110080. https://doi.org/10.1016/j.jobe.2024.110080.

[2] Schmidt, A., Gemmil, E., & Hoskins, R. (2025). Machine learning based risk analysis and predictive modeling of structure fire related casualties. *Machine Learning with Applications, 20*, 100645. https://doi.org/10.1016/j.mlwa.2025.100645.

[3] Ahn, S., Won, J., Lee, J., & Choi, C. (2024). Comprehensive building fire risk prediction using machine learning and stacking ensemble methods. *Fire, 7*(10), 336. https://doi.org/10.3390/fire7100336.

[4] Liu, Z., Xu, Y., Li, Z., et al. (2025). Towards evidence-based fire prevention policy: Uncovering drivers of urban residential fire spread via explainable machine learning. *Developments in the Built Environment, 24*, 100761. https://doi.org/10.1016/j.dibe.2025.100761.

[5] Tam, W. C., Fu, E. Y., Li, J., et al. (2022). A spatial temporal graph neural network model for predicting flashover in arbitrary building floorplans. *Engineering Applications of Artificial Intelligence, 115*, 105258. https://doi.org/10.1016/j.engappai.2022.105258.

[6] Zhang, T., Wang, Z., Wong, H. Y., et al. (2022). Real-time forecast of compartment fire and flashover based on deep learning. *Fire Safety Journal, 130*, 103579. https://doi.org/10.1016/j.firesaf.2022.103579.

[7] Lundberg, S. M., Erion, G., Chen, H., et al. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence, 2*: 56–67. https://doi.org/10.1038/s42256-019-0138-9.

[8] Rohde, D., Corcoran, J., Sydes, M., & Higginson, A. (2016). The association between smoke alarm presence and injury and death rates: A systematic review and meta-analysis. *Fire Safety Journal, 81*: 58–63. https://doi.org/10.1016/j.firesaf.2016.01.008.

[9] Motta-Mena, N. V., Cloninger, C., & Nauhaus, G. M. (2020). A framework for analyzing residential fire outcomes in homes with operative smoke alarms. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 64*(1): 1660–1664. https://doi.org/10.1177/1071181320641404.

[10] Festag, S. (2026). The effect of the smoke alarm obligation: Experiences from Germany based on the data from 1998–2022. *Fire Safety Journal, 159*: 104590. https://doi.org/10.1016/j.firesaf.2025.104590.

[11] Runefors, M., Johansson, N., & van Hees, P. (2016). How could the fire fatalities have been prevented? An analysis of 144 cases during 2011–2014 in Sweden. *Journal of Fire Sciences, 34*(6): 515–527. https://doi.org/10.1177/0734904116667962.

[12] Runefors, M., & Nilson, F. (2021). The influence of sociodemographic factors on the theoretical effectiveness of fire prevention interventions on fatal residential fires. *Fire Technology, 57*(5): 2433–2450. https://doi.org/10.1007/s10694-021-01125-x.