Machine Learning for Wildfire Classification: Exploring Blackbox, eXplainable, Symbolic, and SMOTE Methods

M. Al-Bashiti[1], M.Z. Naser[1,2]
[1]School of Civil & Environmental Engineering and Earth Sciences (SCEEES), Clemson University, USA
[2]AI Research Institute for Science and Engineering (AIRISE), Clemson University, Clemson, SC 29634, USA
E-mail: malbash@clemson.edu, E-mail: mznaser@clemson.edu, Website: www.mznaser.com

Abstract

Whether triggered by natural or human-made events, wildfires are considered one of the most traumatic events to our community and environment. Thus, properly predicting wildfires continues to be an active area of research. This work showcases a statistical overview of the problem of wildfires and then presents a dense data-driven (D^3) approach that leverages a variety of machine learning (ML) techniques, namely, *blackbox* and *eXplainable* ML (i.e., deep learning (DL), decision tree (DT), Stochastic Gradient Descent (SGD), Extreme Gradient Boosted Trees (ExGBT), Logistic regression (LR)), and *symbolic* ML via genetic algorithms (GA) to classify and predict wildfire breakouts. This approach was developed and validated using two databases comprising more than 1.04 million geo-referenced wildfires that burned over 359,000 km^2 (88.7 million acres) between 1992 and 2015 in North America and Europe. Despite the complex nature of wildfire formation and the interdependency of its governing factors, the findings of this D^3 analysis show the feasibility of utilizing ML in preciously classifying the expected size of wildfires and predicting the possibility of the breakout of wildfires.

Keywords: Wildfires; Forests; Machine learning; Big data; explainable ML, Symbolic ML.

Introduction

The start of the twenty-first century marks a clear transition in which the number and intensity of wildfires have exponentially risen [1]. While they can start naturally, wildfires are often caused by humans with devastating consequences. On average, wildfires burn up to 1.11 billion acres of land each year [2,3]. The united states wildfires have been significantly increasing from (140 to 250 wildfires) from (1980 to 2012)[3]. Although wildfires occur worldwide, they are most common in regions with intense droughts and frequent lightning/thunderstorms.

This rise in wildfire occurrences mirrors the recent changes to our environment in which the combination of dry conditions, extended high temperatures, and trapped emissions contribute to some of these changes [4,5]. More specifically, climate change effects (and increased global warming) generate heated conditions that draw moisture from the soil and dry out plants. Global warming has not only led to the rise in wildfire occurrences, along with their intensity but has also led to an increase in human and animal casualties, property losses, and environmental damage [6,7]. This has been duly noted in the recent wildfires that broke in North America and Europe over the past few years.

Wildfires require three components to breakout, known as the fire triangle. These include; a heat source, fuel, and oxygen, heat sources, such as lightning, can supply enough heat to ignite a fire that turns into flames when fuel or any flammable material is present [8]. Such ignition is bound to spread and transport given the presence of favorable wind conditions [9,10].

Recent years have noted a surge in the amount of works that developed different approaches with the power of data analytics to forecast the breakout of wildfires. Collectively, a number of researchers [1,11–13] noted that there are three super high-tech approaches often used to predict wildfire occurrences and stop them from surging. These approaches are grouped under physics-based methods, statistical methods, and machine learning methods.

The first class of approaches, those lumped under physics-based methods, predicts fire breakout by using a mathematical formulation that relies on fluid and heat transfer principles [14]. As such, these approaches use novel software such as FireStation [15] and LANDIS-II [16] to model and trace wildfire through geographical space and time. Due to the extensive use of software and the need for detailed parameters on various inputs (i.e., fuel mass, characteristics of trees, air humidity, soil moisture, etc.), predictions from such approaches heavily rely on assumptions used in the analysis, are complex to set-up and computationally expensive [17].

The second approach, statistical methods, complements physics-based methods as they can also be applied to model large/spatial areas while overcoming the simulation complexity. Further, statistical methods can benefit from modern technologies (i.e., geographic information system (GIS), etc.) and can be applied at different scales and resolution/roughness [18,19]. Some of the notable statistical approaches include Poisson regression [20], Monte Carlo simulations [21], weights of evidence [22], etc. Unfortunately, statistical methods could be sensitive to the type of analyzed data and may require numerical manipulation to satisfy convergence criteria – especially for those methods associated with nonlinear nature [23].

The third and most recent approach is one that leverages advancements in computer sciences. More effectively, machine learning rises as an attractive approach given its good handling of complex and high dimensional data, scalability, and affordability. Machine learning algorithms are applied, tweaked, or created to understand the complex interaction of multi-variables associated with wildfires [24–28]. While the open literature seems to favor the use of such algorithms (i.e., neural networks [29], gradient boosting [30], k-nearest neighbors [31], etc.) and despite the convenience of user-friendly and easy-to-use software that streamlines the development of machine learning by employing pre-defined algorithms and training/validation procedures [32,33], we continue to lack sufficient works on this front.

A recent look into some of the works in this area clearly shows the merits of applying machine learning to enable modern and accurate prediction of wildfires [34,35]. In fact, Fig. 1 reinforces this notion by presenting the publication trend in article publications between 2000-2020 as obtained from a scientometrics analysis from the open-source scholarly database, *Dimensions* [36,37]. As one can see, this search returns 8,716 papers. This trend of publication is expected to continue to rise in the coming years as it capitalizes on the continued advancements in computer science.

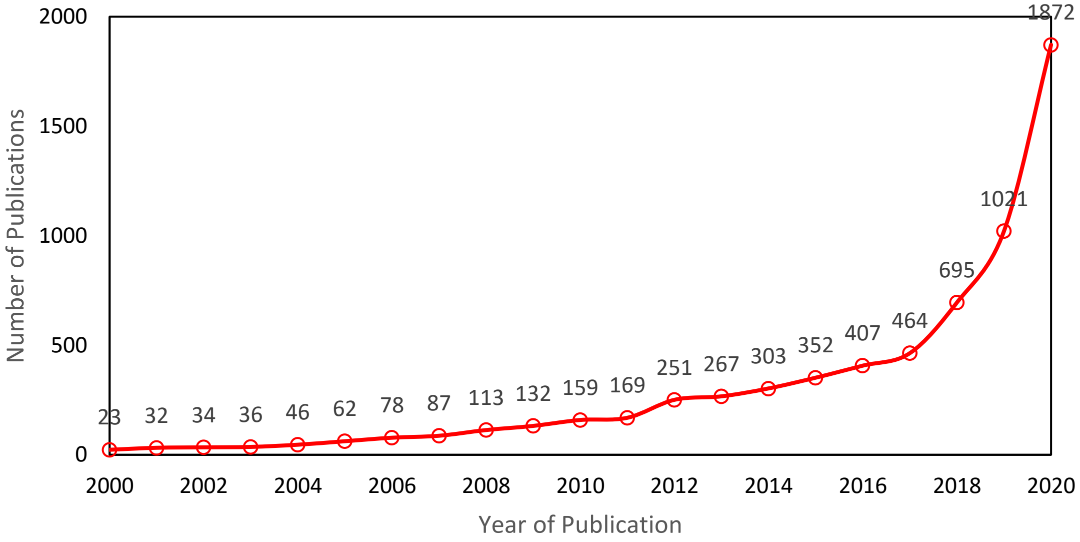


Fig. 1 Publication trend as obtained from the *Dimensions* academic database [36,37] [Note: Keywords "*Machine learning*" AND "*wildfire*"]

A deeper look into the majority of the noted publications displays that most works showcased the incorporation of one algorithm, often selected from researchers' familiarity with such algorithm (in a similar manner to opting to use a particular simulation software/package). However, our perspective is that reliance on a particular model, while it may produce favorable performance, can still generate biased models that could be overturned. In contrast, we would like to explore the use of multi-algorithmic search to identify suitable machine learning model candidates that can be used in parallel, thereby expanding a researcher's arsenal of predictive tools while adding an additional layer of redundancy.

In addition, the majority of the reviewed works adopt blackbox models that require the user to program and code the machine learning model. This may lead to dependence on computing stations and, most admittedly, a heavy reliance on the user's coding experience. To overcome these hurdles, we present the use of genetic algorithms as a means to augment the blackbox models and derive expressions that can substitute the need for algorithmic simulation. Simply put, the machine learning model will be run once to obtain the predictions, and then these predictions are fitted into an expression (or a form of a mapping function [38]) that can be substituted by hand or via a simple spreadsheet. The user would not need to code a new machine learning model to predict wildfire occurrence since the user can now use the derived expression directly.

In hopes of narrowing this knowledge gap and in pursuit of accelerating research efforts in this area, this work presents a statistical overview of the problem of wildfires and then deep dives to present a dense data-driven (D^3) approach that integrates different machine learning algorithms to realize modern wildfire assessment tools that have the capability to predict occurrence and size of wildfires. This approach was developed and validated using measured data points obtained from 1.04 million geo-referenced wildfires between 1992 and 2015 in North America and Europe. The consequence of this D^3 analysis demonstrates the suitability and feasibility of exploiting intelligent analysis tools to modernize wildfire disaster planning and optimal resource allocation.

Statistical Overview

Recent statistics have shown that the annual count of worldwide wildfires reaches 200,000 and that these fires burn 3.5–4.5 million km^2 (equivalent to 0.86–1.11 billion acres) [2,3]. It is also interesting to note that the average number of large wildfires occurring in the United States increased from 140 to 160 to 250 in the periods of 1980-1989, 1990-1990, and 2000-2012, respectively [3]. The United Nations Office for Disaster Risk Reduction (UNODRR) also supports these statistics and reports that insured losses arising from wildfires around the world in 2017 totaled $14 billion, the highest ever in a single year [4]. Trends were most substantial for southern and elevated regions, overlapping with tendencies toward amplified drought severity. The number of large fires increased as well as the total fire area increased per year [80].

In the United States, wildfires in the Western region are greater and burn more land than their counterparts in other regions. For example, nearly 26,000 wildfires burned approximately 9.5 million acres in the Western US, as compared with the over 33,000 fires that burned about 0.7 million acres in Eastern regions in 2020 [81]. This horrifying statistical information led to extending the average length of fire season from 5 months in the early 1970s to slightly over seven months nowadays (in the US) [3]. These prime conditions in forests for frequent and intense wildfires as opposed to those experienced in the past decades [7,8]. On the European front, the UNODRR also noted a similar observation and reported how the most damaging fires that occurred in June and October of 2017 fell outside of the traditional fire season (July to September), thus, indicating a shift towards a longer wildfire season [4].

The surge in the number of wildfires, along with their intensity, is expected to increase associated casualties, property losses, and environmental damage [9,10]. This has been noted in the recent wildfires that broke in North America and Europe over the past few years. For example, the last year was one of the most destructive fire season in California, in which over 7,600 km^2 burned, causing over $3.5 billion in damages. It was also in the same season that Mendocino Complex Fire (which burned over 1,860 km2) became the largest single fire in California's history [11]. Within the same timeframe, the Canadian province of British Columbia underwent its largest wildfire, which caused the burning of an area equivalent to 1.3% of the total territory. This fire also led to evacuating 40,000 people [11]. The past few years have also witnessed similar occurrences most notably in Greece [12], Portugal [13], and most recently in Australia. In a nutshell, forest wildfires pose a serious threat to our communities and need to be properly understood, predicted, and mitigated [14].

To date, over 46 million homes in 70,000 urban, suburban and native communities are at risk of wildfires in the US alone [39]. One should also be cognizant that in a severe fire season, wildfires can burn thousands of structures (e.g., 10,488 buildings in the 2020 California wildfires [40] and 5,900 buildings in the 2020 Australian bushfires [41]), and such numbers are expected to rise given the recent inertia for urban development and construction. While statistics on human losses are often accessible [42,43], statistics on animal losses may not be as easily obtained. According to the World Wide Fund for Nature [44] at least 1.25 billion animals were killed (and about 2.75

billion were harmed) during the 2020 Australian bushfires alone. At the time of this proposal, we were not able to find a reliable source to report the number of animal losses to US wildfires.

Methods

Development of Databases

In order to effectively apply the D^3 approach, there is a need to compile observations on wildfires in order to develop a proper wildfire database. As such, a literature survey was carried out and resulted in identifying two publicly available databases comprising more than 1.04 million geo-referenced wildfires that burned over 359,000 km^2 (88.7 million acres) between 1992-2015 in the United States [45] and Portugal [46]. These databases cover well documented wildfires with varying aspects and characteristics. These databases will be used for separate machine learning analyses.

The first analysis aims to use the first database (to be referred to as the US database here) to create machine learning classifiers that can predict the occurrence and expected size of a given wildfire as a function of a set of variables (outlined below). In the second analysis, the second database (aka. Portugal database) is used to create mathematical expressions that can identify the expected size of a wildfire pending environmental features. Both databases, along with their variables, are described below.

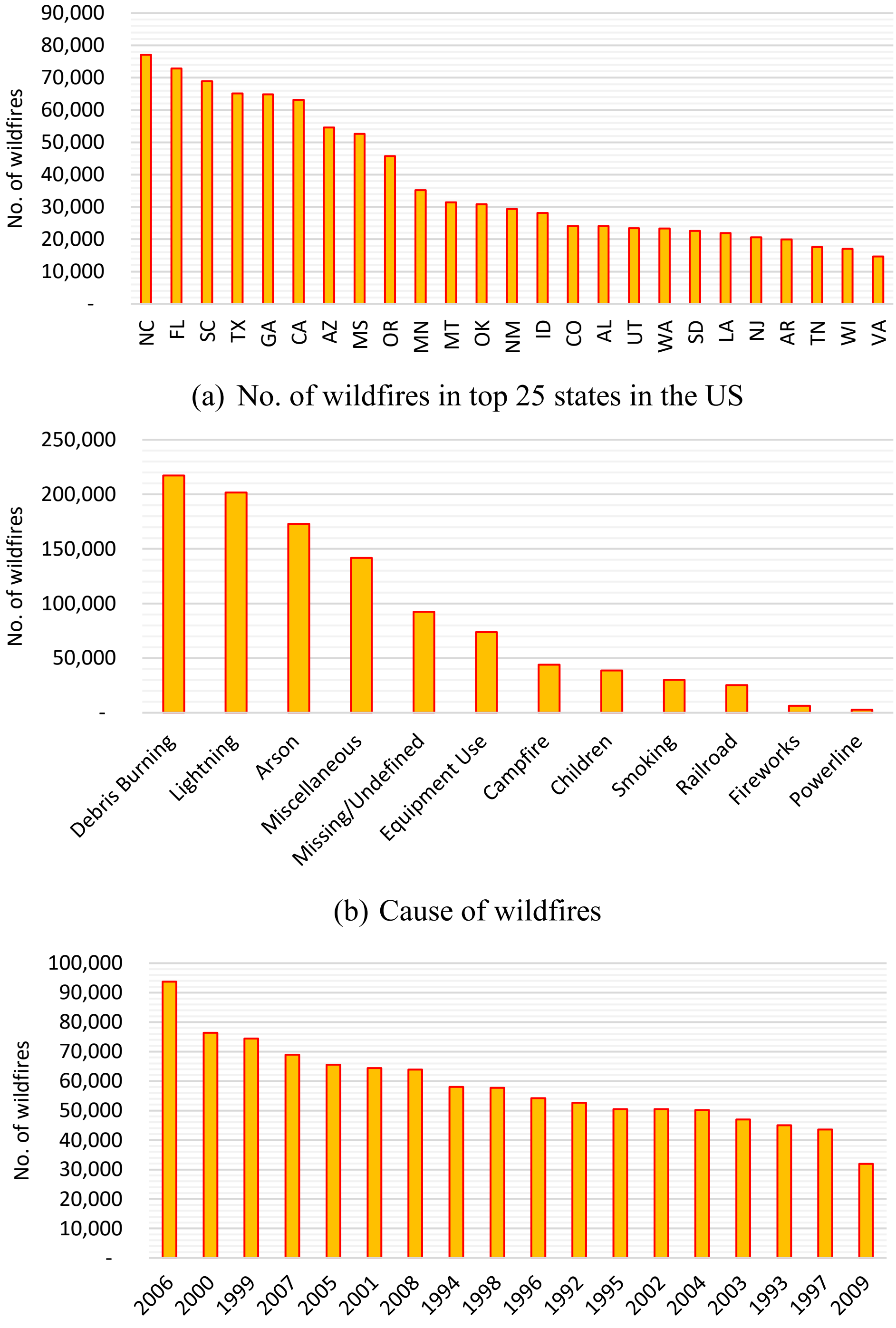
Database on wildfires occurring in the US

The first database covers a spatial description of major wildfires that broke out within the United States (US) from 1992 to 2015. The US area covers approximately 9,830,000 km^2. These fires were obtained from the reports published by federal, state, and local fire organizations. The observations were transformed to comply with the standards of the National Wildfire Coordinating Group (NWCG) [47]. It is worth noting that this database was initially pre-processed to remove redundant and incomplete observations. After this cleansing procedure, a total of 1.04 million (out of 1.88 million) geo-referenced wildfire records that burned through 88.7 million acres during the aforementioned 24-year period were arrived at.

The same database contains 50 parameters (ranging from geographical location to fire breakout cause and size etc.) and can be freely accessed at [45] or [48]. The database also contains six variables: discovery day of wildfire (a numerical value ranging between 1-365), year of wildfire (a numerical value ranging between 1992-2015), latitude and longitude of wildfire occurrence, wildfire cause (in thirteen categories[*]), and state at which wildfire took place. Further, this database has one predictor as "wildfire size," and this was divided into seven classes that are arranged alphabetically; (A[†]=greater than 0 but less than or equal to 0.25 acres, B=0.26-9.9 acres, C=10.0-99.9 acres, D=100-299 acres, E=300 to 999 acres, F=1,000 to 4,999 acres, and G=5,000+ acres). Figure 2 shows further statistics and the geographic location of wildfires from this database.

[*] Categories include: arson, camp fire, debris burning, equipment use, fireworks, lightning, miscellaneous, powerline, railroad, smoking, structures, caused by children, and undefined.

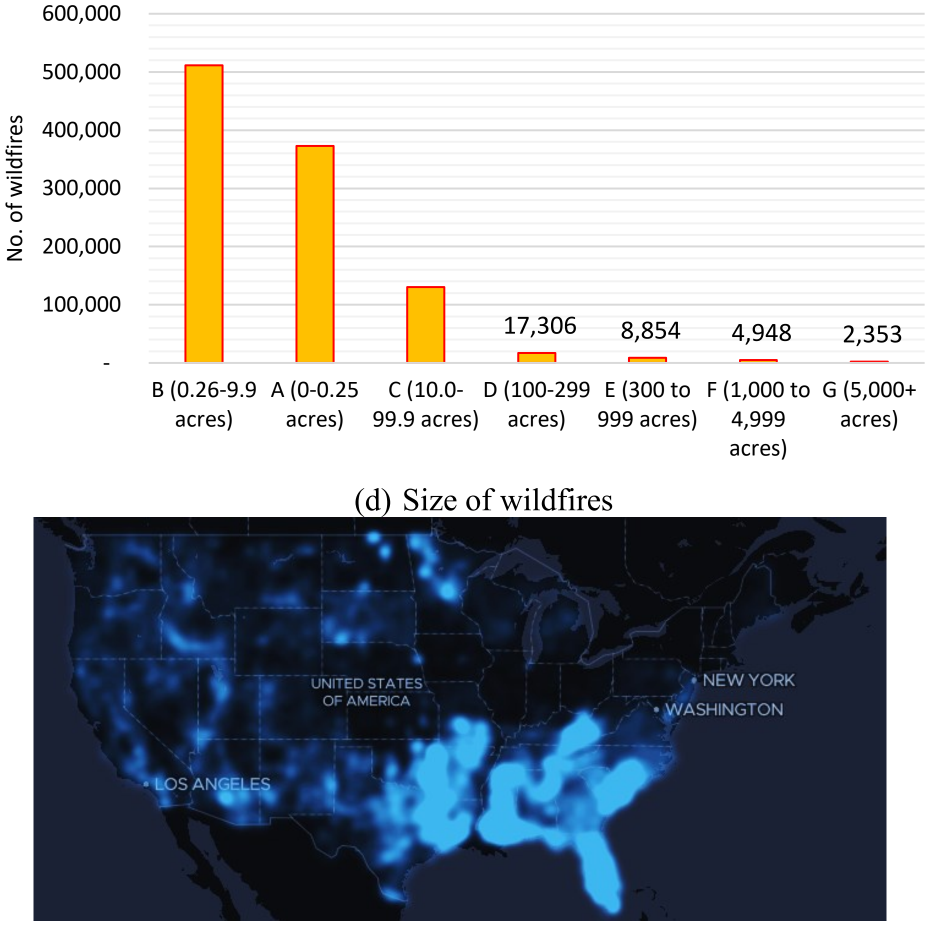
[†] Due to its small area, this class was not examined further herein.



(a) No. of wildfires in top 25 states in the US

(b) Cause of wildfires

(c) No. of wildfires per year



(d) Size of wildfires

(e) Spatial distribution of wildfires

Fig. 2 Statistics obtained from the first database (US)

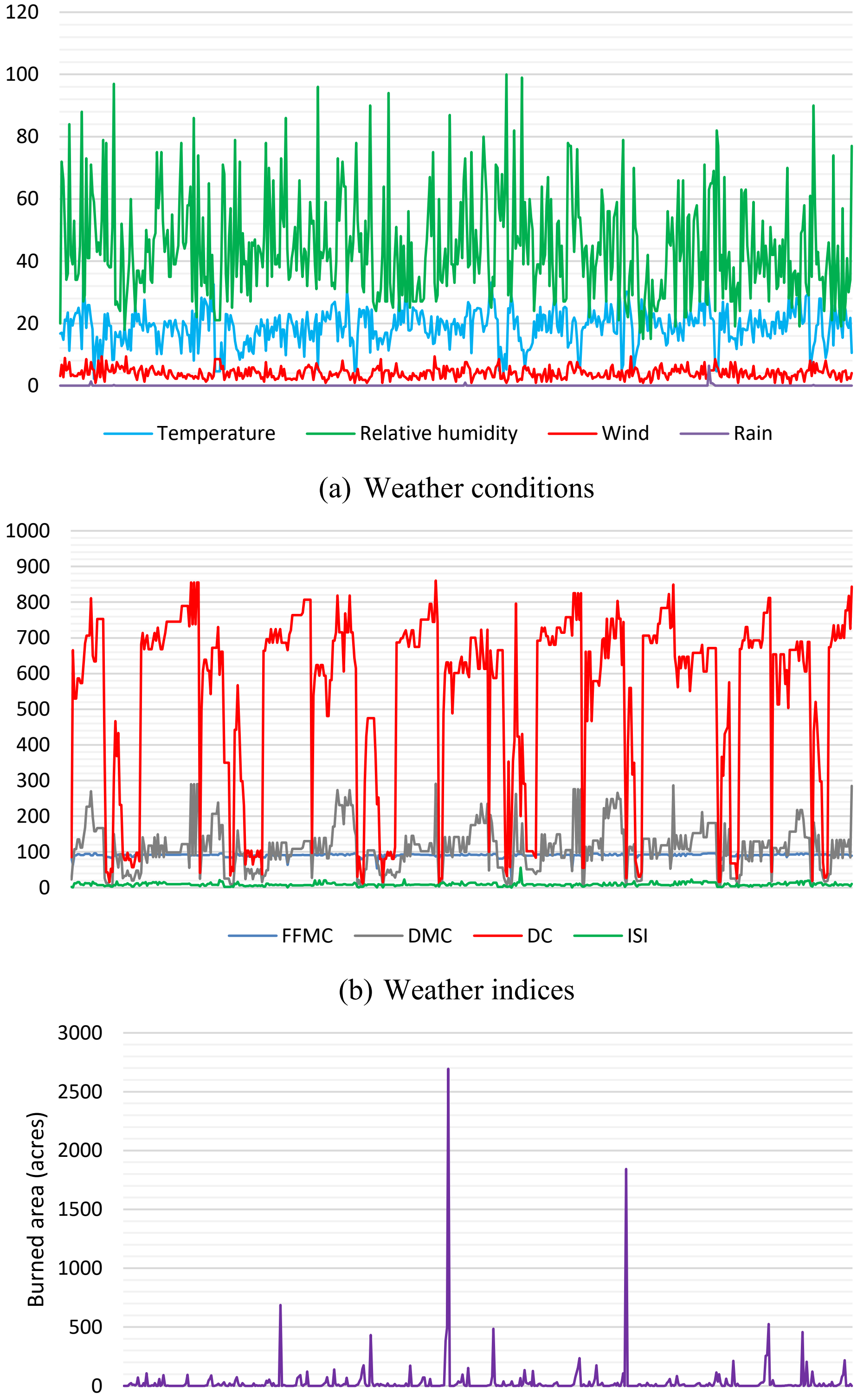
Database on wildfires occurring in Portugal

The second database was prepared by Cortez and Morais [49,50], and this database was collected from the burned areas of Montesinho natural park, located in the northeast region of Portugal. The database contains 517 wildfires that occurred between January 2000 to December 2003.

The database comprises the following attributes: geographic features, temporal variables, average monthly weather settings (e.g., temperature, relative humidity, wind speed, rain), as well as distinct weather-based indices. These indices include Fine Fuel Moisture Code (FFMC) which influences ignition and fire spread, Duff Moisture Code (DMC), Drought Code (DC), Initial Spread Index (ISI), which correlates with fire velocity spread, Buildup Index (BUI), and Fire Weather Index (FWI) following the Canadian system for rating fire danger[‡] [51]. As per suggestions laid out by Cortez and Morais (2008, 2007), the following four weather index were used as attributes from this database: FFMC, DMC, DC, and ISI (as the rest of the attributes make up these indices). The

[‡] FFMC: moisture content surface litter. DMC and DC represent the moisture content of shallow and deep organic layers (and hence affect fire intensity). BUI reflects upon the availability of fuel. FWI combines ISI with BUI and indicates the magnitude of fire intensity.

predictor in this database was the size of burned areas and ranged from 0-11 km^2 (0-2695.5 acres) in a similar breakdown to classes that used in the first database is also followed herein. Figure 3 shows further statistics on this database.



(a) Weather conditions

(b) Weather indices

(c) Size of burned areas

Fig. 3 Statistics obtained from the second database (Montesinho natural park, Portugal) [Note: The horizontal represents each individual wildfire]

Description of Machine Learning Algorithms

A dense data-driven (D^3) approach that leverages machine learning to uncover hidden patterns within the two datasets described above is presented herein.

The overarching goal of the D^3 approach is to draw conclusions that could be mapped into a solution (or set of solutions) to the wildfire occurrence phenomenon being investigated as part of this study. To attain at such a solution, *key features* governing a wildfire breakout and spread in addition to the *governing relation* that connects these features, are to be identified first. As researchers, our domain knowledge alludes to the fact that a wildfire can breakout once/if several conditions converge. Such conditions may include weather and climate factors (i.e., temperature, humidity, etc.), spatial factors (topology, ignition agents, etc.), and fuel conditions (i.e., vegetation type, heterogeneity of landscape, etc.), among others. The interaction of these features determines how a wildfire can break out and how it will spread, intensify, and potentially be controlled.

Hence, the rationale behind adopting D^3 is that since wildfires behavior can be observed (say in the databases from actual fires as collected in Sec. 3.0), then a governing relation connecting the cycle of a wildfire to its key features can be obtained through D^3. Such a relation can be arrived at via machine learning models, as well as could be converted into a mathematical expression via symbolic ML. A systematic analysis of such a magnitude will require special computational treatment, and this is where D^3 shines. The following algorithms are used herein; deep learning (DL), decision tree (DT), Stochastic Gradient Descent (SGD), Extreme Gradient Boosted Trees (ExGBT), Logistic regression (LR), and genetic algorithms (GA), and these are further described below.

Deep learning (DL)

The architecture of a DL algorithm follows that of the brain and consists of a similar topology or layout (see Fig. 4). Such topology is characterized by layers. The outermost layer receives the data (representing attributes, say wildfire cause, metrological conditions, etc.) to be analyzed. For this reason, this layer is denoted as the *input layer*. The inputs are then fed into the next set of layers, the *hidden layers*. These layers, or in some cases one layer, house processing units called neurons. The neurons analyze input data via a series of generated weightages (connections). It is through this analysis that the algorithm learns and recognizes any relevant patterns impeded by input data points. This recognition is then mapped into patterns using transformative operations and functions. This aforenoted process sums up the training process of a typical DL algorithm. Once this process passes fitness requirements (whether a pre-defined number of iterations and/or until satisfying a set of fitness metrics), the training is completed, and the algorithm is set into the testing stage.

The most frequently adopted optimization technique in DL is called Leveberg-Marquard. This technique assesses the error by evaluating the mean squared error (MSE) [52]. In this optimization method, if z is the experimental dataset, then MSE is evaluated using Eq. 1:

$$MSE = \frac{1}{z}\sum_{i=0}^{z}(e_i)^2 = \frac{1}{z}\sum_{i=0}^{z}(m_i - p_i)^2 \quad (1)$$

where, z = the total number of datasets, e_i = the error for each input set, m_i = the measured output, and p_i = the estimated output.

In this development, a pre-sensitivity analysis inferred that adopting a *ReLu* activation function for DL with an initial learning rate of 0.001, 3 hidden layers (with 256, 128, and 64 units) led to achieving an optimal DL architecture. The final outcome within the hidden layers is then forwarded to the output layer for visualization.

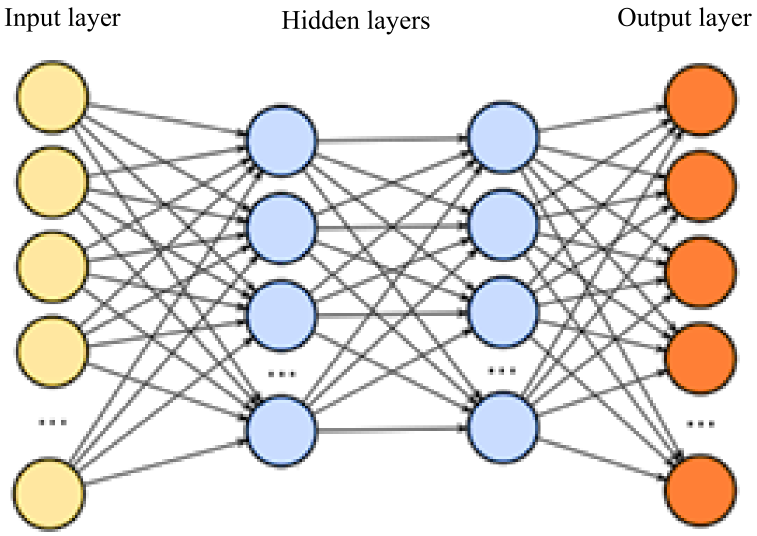


Fig. 4 A typical DL topology

Decision tree (DT)

A DT algorithm is attractive in a classification-like problem similar to that tackled herein to classify the expected size of a wildfire. A key advantage to DT is its ability to create a diagram-like depiction of all likely decisions [53]. The DT analysis starts by separating the database into branch-like shapes. Then, a random decision tree is created at a root node and then grows into other tree-like components (i.e., leaves, etc.). The created DT was optimally designed to have a maximum depth of 45, with a confidence level = 0.1, minimum leaf size, and maximum size for split equals 2 and 4, respectively [54,55]. A DT analysis may utilize additional measures such as Gini impurity to facilitate the analysis and processing of data points. For example, for a node *t*, Gini index *g*(*t*) is defined as [56]:

$$g(t) = \sum_{j \neq i} p(j|t)p(i|t) \quad (2)$$

where *i* and *j* are target field categories.

$$p(j,t) = \frac{p(j,t)}{p(t)}; p(j,t) = \frac{\pi(j)N_j(t)}{N_j}; \text{ and } p(t) = \sum_j p(j,t) \quad (3)$$

Stochastic Gradient Descent (SGD)

SGD regularizes linear models such as support vector machines and logistic regression with stochastic gradient descent (SGD) learning in classification problems [57]. SGD adopts a plain stochastic gradient descent learning process with a loss penalty function as shown in Eq (4).

$$E(w,b) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i)) + \alpha R(w) \tag{4}$$

Where, L is a loss function that measures model, R is a regularization term; $\alpha>0$ is a non-negative hyperparameter that controls the regularization strength. The developed algorithm was incorporated with *LogLoss* as a loss function, *ElasticNet* as a regularization function, and $\alpha = 2.2\times10^{-5}$.

Extreme Gradient Boosted Trees (ExGBT)

This algorithm [58] re-samples data points into a series of tree, where each tree boostraps a sample some data points in each iteration. ExGBT fits each successive tree to the residual errors from all the previous trees and focuses on the most difficult cases to predict to increase its prediction accuracy (see Eq. 5). The developed algorithm incorporated a learning rate of 0.05, a maximum tree depth of 5, a subsample feature of 0.8, and a minimum interval for early stopping of 200.

$$Y = \sum_{k=1}^{M} f_k(x_i)\,, f_k \in F = \{f_x = w_{q(x)}, q: R^p \to T, w \in R^T\} \tag{5}$$

where, M is additive functions, T is the number of leaves in the tree, w is a leaf weights vector, w_i is a score on i-th leaf, and $q(x)$ represents the structure of each tree that maps an observation to the corresponding leaf index [59].

Logistic regression (LR)

The regularized LR algorithm aims to maximize the likelihood of observing a phenomenon through its capability to estimate coefficients for identified features to measure the comparative influence of each feature on the phenomenon [60]. Therefore, LR is noted to be a successful algorithm for classification problems [61]. LR, and just like other algorithms, can suffer from overfitting. To avoid this, LR's loss function can be modified with a penalty term to shrink/penalize the estimates of the coefficients. L2 penalty is used herein as it is proven effective during a pre-sensitivity study [62]. The used algorithm has a true fit intercept and approximates the multi-linear regression function:

$$logit(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n \tag{6}$$

where, p is the probability of the presence of a phenomenon. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} \tag{7}$$

and,

$$logit(p) = \ln\left(\frac{p}{1-p}\right) + L2_{(penalty)} \qquad (8)$$

The developed algorithm incorporated a learning rate of 0.05, a maximum tree depth of 5, subsample feature of 0.8, and a minimum interval for early stopping of 200.

Genetic Algorithms (GA)

This algorithm is an evolutionary method that was initially presented by Holland [63] and Koza [64]. GA leverages the concept of the natural selection process to arrive at hidden relations between attributes and expected outcomes in a symbolic format. In GA, a set of expressions are numerically derived from mapping to mathematical expressions that can be used to represent the size of wildfires [65,66].

The GA analysis starts by creating a populace of arbitrary expressions. These expressions consist of a tree-like formation that houses mathematical operations (addition, multiplication, etc.) and/or mathematical functions (power, log, etc.). In some scenarios, a GA-based expression may also contain conditional and logic functions. The GA-based expression is configured into a tree with hierarchical form, which can then be transformed into a *Karva-expression* as shown in Fig. 5. Once a set of the suitable formula is generated, the algorithm then assesses the fitness (i.e., accuracy) of each expression. Only the fittest expression is then selected for the next stage of analysis. In this stage, the expression is then manipulated by bio-inspired transformative operations, i.e., *reproduction*, *crossover,* and *mutation* [64,67].

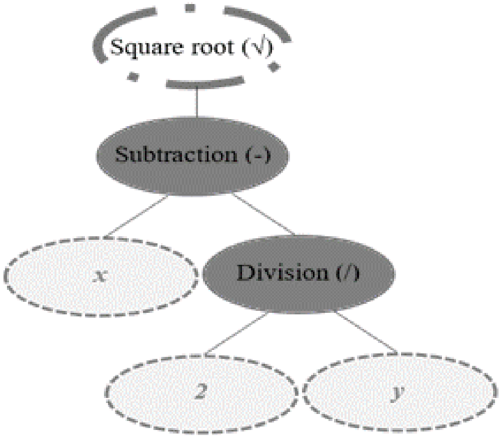


Fig. 5 Representation of a typical GA

The first, reproduction, the operation ensures that fittest expressions have higher primality of selection to the following stages of analysis. The second, crossover, operation allows the exchange of genetic code (i.e., mathematical functions) between evolved expressions. The third mutation, an operation, can randomly select a function from an expression to mutate into another function [68]. Similar to other algorithms, the GA analysis also terminates once the fitness of a fit expression is achieved or by satisfying a convergence condition. Figure 6 demonstrates a typical flow of GA analysis.

This is a preprint draft. The published article can be found at: https://doi.org/10.1016/j.nhres.2022.08.001.
Please cite this paper as:
Al-Bashiti, M., Naser, M.Z. (2022). Machine Learning for Wildfire Classification: Exploring Blackbox, eXplainable, Symbolic, and SMOTE Methods. *Natural Hazards Research*. https://doi.org/10.1016/j.nhres.2022.08.001.

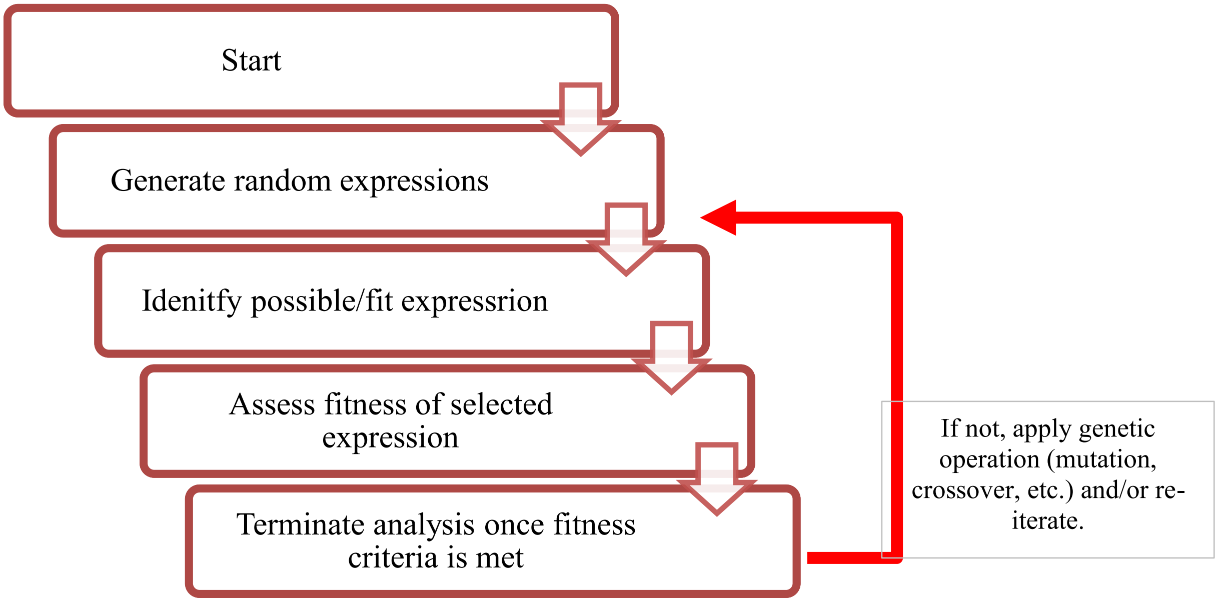

Start
Generate random expressions
Idenitfy possible/fit expressrion
Assess fitness of selected expression
Terminate analysis once fitness criteria is met
If not, apply genetic operation (mutation, crossover, etc.) and/or re-iterate.

Fig. 6 A flowchart of GA analysis

Results and Discussion

Now that the databases are compiled, these databases can be analyzed using the selected algorithms. This analysis resembles a classification problem where each machine learning model is expected to correctly label the examined fires (given each fire's set of variables). To start this analysis, first, each dataset was first randomly shuffled to minimize biases arising from a specific wildfire attribute. Then, the ML algorithms are trained using 10 k-fold cross validation [69,70]. The analysis was conducted by using the aforenoted algorithms in Matlab [71], Python [72], and GMDH environments [73,74][§].

The outcome of each machine learning model is then cross checked against that of the ground truth. In this pursuit, specific classification metrics are used. The first is a composite metric known as the confusion matrix, and the second is the *LogLoss* error [75].

The outcome of the D^3 analysis is listed in Table 1 by means of the confusion matrix. This matrix lists the fitness of the applied algorithms in classifying the wildfires as well as two fitness metrics (Accuracy and *LogLoss* error). The Accuracy (ACC) metric evaluates the ratio of a number of correct predictions to the total number of samples.

$$ACC = \frac{TP+TN}{P+N} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (9)$$

where, P (denotes the number of real positives), N (denotes the number of real negatives), TP (denotes the true positives), TN (denotes t the rue negatives), FP (denotes the false positives), and

[§] In this study, the databases were kept in their original datapoints without any preprocessing to minimize their imbalanced-nature to examine the raw effectiveness of the selected algorithms when applied "as is". A future study will explore different treatment techniques for imbalanced data for the same algorithms. Incorporating such techniques and results can significantly push the size of this paper beyond the limitation of a standard article.

FN (denotes false negatives) – and hence the composite nature of the matrix. And, LogLoss error metric measures where the prediction input is a probability value.

$$LLE= -\sum_{c=1}^{M} A_i logP \qquad (10)$$

where, M: number of classes, c: class label, y: binary indicator (0 or 1) if c is the correct classification for a given observation. It is worth noting that an accuracy closer to unity and a LogLoss error close to zero imply favorable predictive performance.

Blackbox ML

The first analysis adopts six *blackbox* machine learning algorithms and six variables: discovery day of wildfire, year of wildfire, latitude, and longitude of wildfire occurrence, wildfire cause, and state at which wildfire took place to predict the expected size of the wildfire. A closer look at Table 1 shows that all models achieved a comparable accuracy that centers around 80% and LogLoss error ranging between (0.42-0.61).

These results show a couple of interesting observations. For a start, regardless of the machine learning model type, or search mechanism, it is clear that the adopted models have a good grasp on predicting wildfire occurrences (with minimal tuning, as noted in Sec. 4.). Secondly, the DL, DT, SGD, ExGBT, and LR algorithms achieved comparable performance in accuracy, with DL, DT, and ExGBT ranking top three. Recent works on the front of wildfires have also noted the predictive capacity of such algorithms [76–78]. Thus, we can comfortably say that adopting these three models as independent and redundant models to identify wildfire breakouts can be of merit.

In all cases, whenever a wildfire class is mistakenly classified with an error larger than 20%, this error is highlighted in red. In addition, it is clear that SGD and LR suffered in predicting individual wildfire sizes. It is clear that the imbalanced nature of the used database on US wildfires adversely affected these algorithms.

Table 1 Performance of ML algorithms

DL	True C	True D	True E	True F	True G	Accuracy	*LogLoss*
Pred. C	84.3	23.8	51.7	30	54.3	0.825	0.418
Pred. D	7.5	76.0	0.0	0.0	0.0		
Pred. E	5.4	0.0	47.8	3.9	0.0		
Pred. F	1.7	0.0	0.4	49.4	0.2		
Pred. G	0	0.0	0.0	16.5	45.4		
DT	True C	True D	True E	True F	True G		
Pred. C	84.4	23.6	52.6	28.7	31.8	0.822	0.433
Pred. D	8	76.3	0.0	0.0	0.0		
Pred. E	5.4	0.0	47.3	3.8	4.5		
Pred. F	1.9	0.0	0.0	50.6	40.9		
Pred. G	1.1	0.0	0.0	16.7	22.7		
SGD	True C	True D	True E	True F	True G		
Pred. C	79.5	99.4	100	44.4	0.0	0.794	0.544
Pred. D	10.5	0.5	0.0	0.0	0.0		
Pred. E	5.4	0.0	0.0	0.0	0.0		
Pred. F	3	0.0	0.0	44.4	0.0		
Pred. G	1.4	0.0	0.0	11.1	0.0		
ExGBT	True C	True D	True E	True F	True G		
Pred. C	82.8	25	37.9	29.8	54.5	0.818	0.451
Pred. D	8.7	74.9	0.0	0.0	0.0		
Pred. E	5.6	0.0	62.1	0.0	0.0		
Pred. F	1.7	0.0	0.0	49.7	1.8		
Pred. G	1	0.0	0.0	16.3	43.6		
LR	True C	True D	True E	True F	True G		
Pred. C	79.5	0.0	0.0	83.3	0.0	0.797	0.611
Pred. D	10.5	0.0	0.0	0.0	0.0		
Pred. E	5.4	0.0	0.0	0.0	0.0		
Pred. F	3	0.0	0.0	0.0	0.0		
Pred. G	1.4	0.0	0.0	16.6	0.0		

Symbolic ML

While the above algorithms are used as more of a standard assessment “tool” for the first database, the second analysis uses GA to arrive at *symbolic* expressions that can be substituted into to estimate wildfire class in the second database, given the availability of information regarding four weather-based indices (FFMC, DMC, DC, and ISI). This decision can be rationalized by the notion that GA, unlike the other blackbox models, can output an expression that a user can apply/substitute directly instead of running a coded model.

For practicality, and since the second database included a number of wildfires that were of low size and intensity, GA-expressions were derived for wildfire classes C, D and E (and greater). Table 2 lists these expressions along with their performance. Table 2 shows that GA managed to properly derive simple expressions that can be used to predict the size of a given wildfire. The predictivity of these expressions was established through the correlation coefficient[**], *R* – see Eq. 11. As one can see, these equations are highly nonlinear and represent the complex nature of the phenomenon on hand.

$$R = \frac{\sum_{i=1}^{n}(A_i-\bar{A}_i)(P_i-\bar{P}_i)}{\sqrt{\sum_{i=1}^{n}(A_i-\bar{A}_i)^2\sum_{i=1}^{n}(P_i-\bar{P}_i)^2}} \tag{11}$$

where, *A* is actual data points, and *P* is for predicted data points.

The derived expressions can come in handy in assessing the projected size of a wildfire knowing the magnitude of the previously identified weather indices. Having such tools can come in handy in a variety of scenarios, especially those associated with abrupt wildfire breakout and those that may require a quick judgment call to allocate proper resources to fight the wildfire. For transparency and completion, we expect future works to be able to devise improved expressions with higher accuracy – especially once the dataset is massaged for imbalanced data.

Table 2 GA-derived expressions to predict wildfire class via weather indices.

Class	Expression	*R*
C	$Class\ C = Step\ (0.0815FFMC + 0.0208DC + 7.047\tan(1.522DMC) + 3.852\tan(193.6DMC) + \tan(136.5DMC) + \tan(\tan(0.2517DMC)) - \tan(5.203FFMC \times DMC) - 0.02121ISI - 0.0647DMC)$	0.82
D	$Class\ D = Logistic\ (0.0326DMC + 0.00656FFMC) + (-1.352 - \frac{\tan(8.933DMC)}{0.02ISI+3.96\times10^{-10}DC\times DMC^3} - \tan(1.524DMC) - \tan(1.522DMC) - \tan\left(5.446\times10^{-5}\ \times maximum(0.005DMC, -0.7548\tan(1.522DMC))\right) - 0.01207DC - 0.04806ISI)$	0.86
E	$Class\ E = Logistic\ \left(0.04039DMC + \tan(\tan(3.442DC)) - \cosh\left(sin\left(tan(0.005DMC)\right)\times\cos(0.161DC^2)\right)\right) - 0.00117DC - 0.1261ISI - 0.1525FFMC)$	0.87

In each case, a value of 1.0 indicates that the outcome of a given expression agrees with the identified class.

Explainable ML

To supplement the GA analysis and to combat the blackbox nature of the traditional algorithm wherein, for example, the models listed in Table 1 do not articulate how the correct or poor predictions were arrived at, we apply the *explainability* method SHapley Additive exPlanations (SHAP) [79] to the ExGBT to better analyze its performance when applied to the second database. In parallel, the initial phase of analysis noted a need to improve the model given the imbalances

[**] We also recommend the adoption of other companion metrics.

of the data in the categories (E, F). Thus, the Synthetic Minority Oversampling TEchnique (SMOTE) was applied [80]. SMOTE copies data from the small classes with the lower data point and adds it to the dataset to create a balanced dataset and better resemble or match the number of examples in most classes. Note that such a technique does not affect the model accuracy and only provides the model with different copies of the samples from the same category.

We start by re-validating this model against the second database. The confusion matrix (see Fig. 7) is also used to validate the model. This matrix shows exactly how many errors have been made by the model by comparing the testing set class with the predicted results and the training set class with the predicted results. It is clear that the model was able to classify over 90% of the samples correctly on the training set only; however, for the testing set, the model achieved 84%. Please note that our Python code is provided in the Appendix.

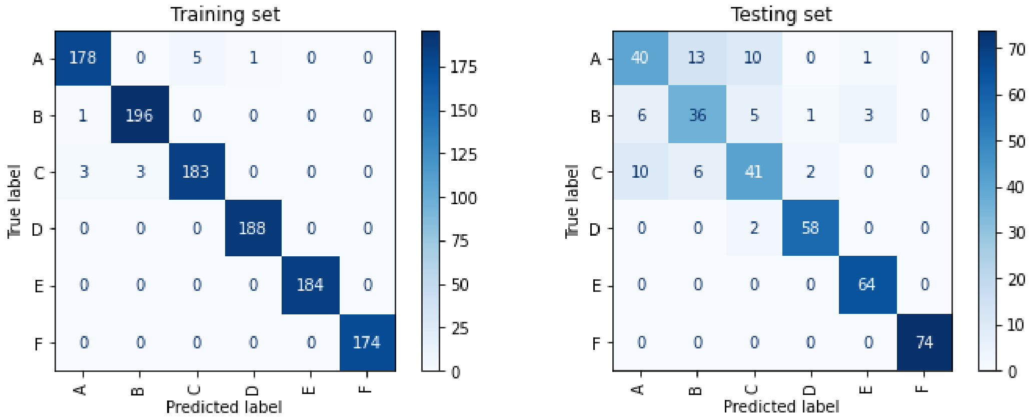


Fig. 7 Confusion matrix on the second database

Now that the model is validated, let us examine the results of our analysis use SHAP *feature importance plot,* which shows all the features stacked in horizontal lines representing the effect of each feature on the predicted class of the occurred wildfire (see Fig. 8). For instance, the temperature was found to be the most influential overall. However, for a specific class, that is not true. Taking class E as an example, the most important factor that affects class E wildfire is ISI. Similarly, the temperature was not seen to be o high importance for classes E and B. Also, temperature and wind is the most important factor in predicting the occurrence of class A wildfires but not for class C. Another example is that by looking at the DMC, we can conclude that it highly influences classes E and D. An inherent issue with such a plot is that it does not explain how each feature affects the model, positively or negatively.

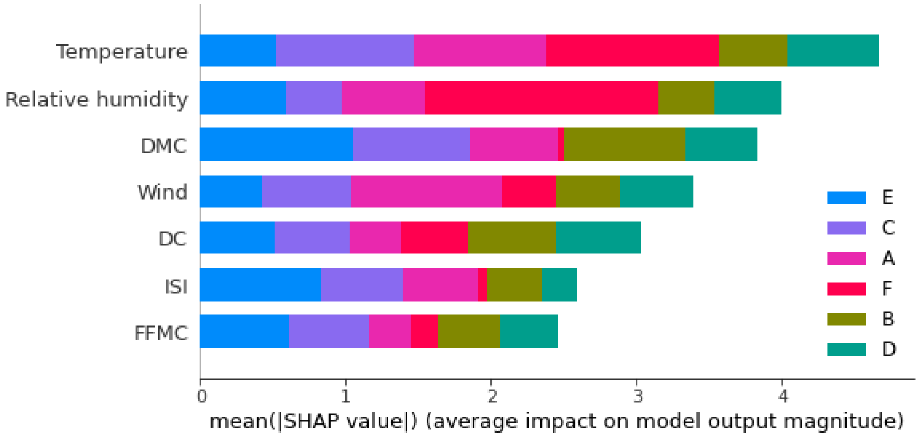
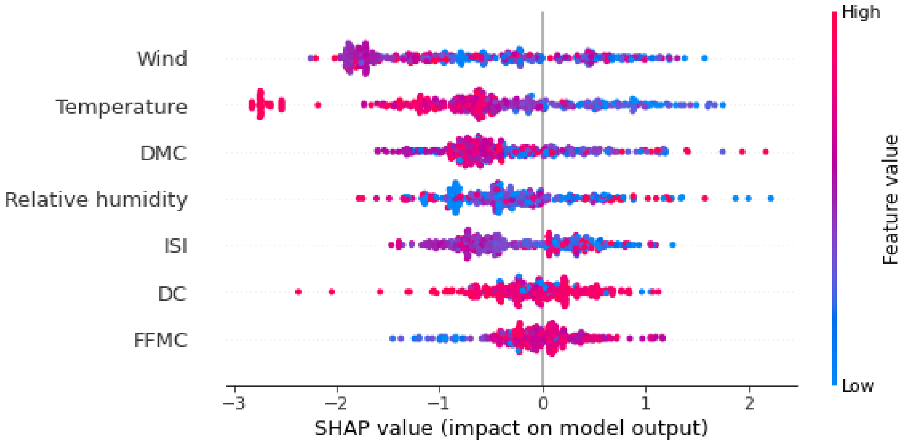
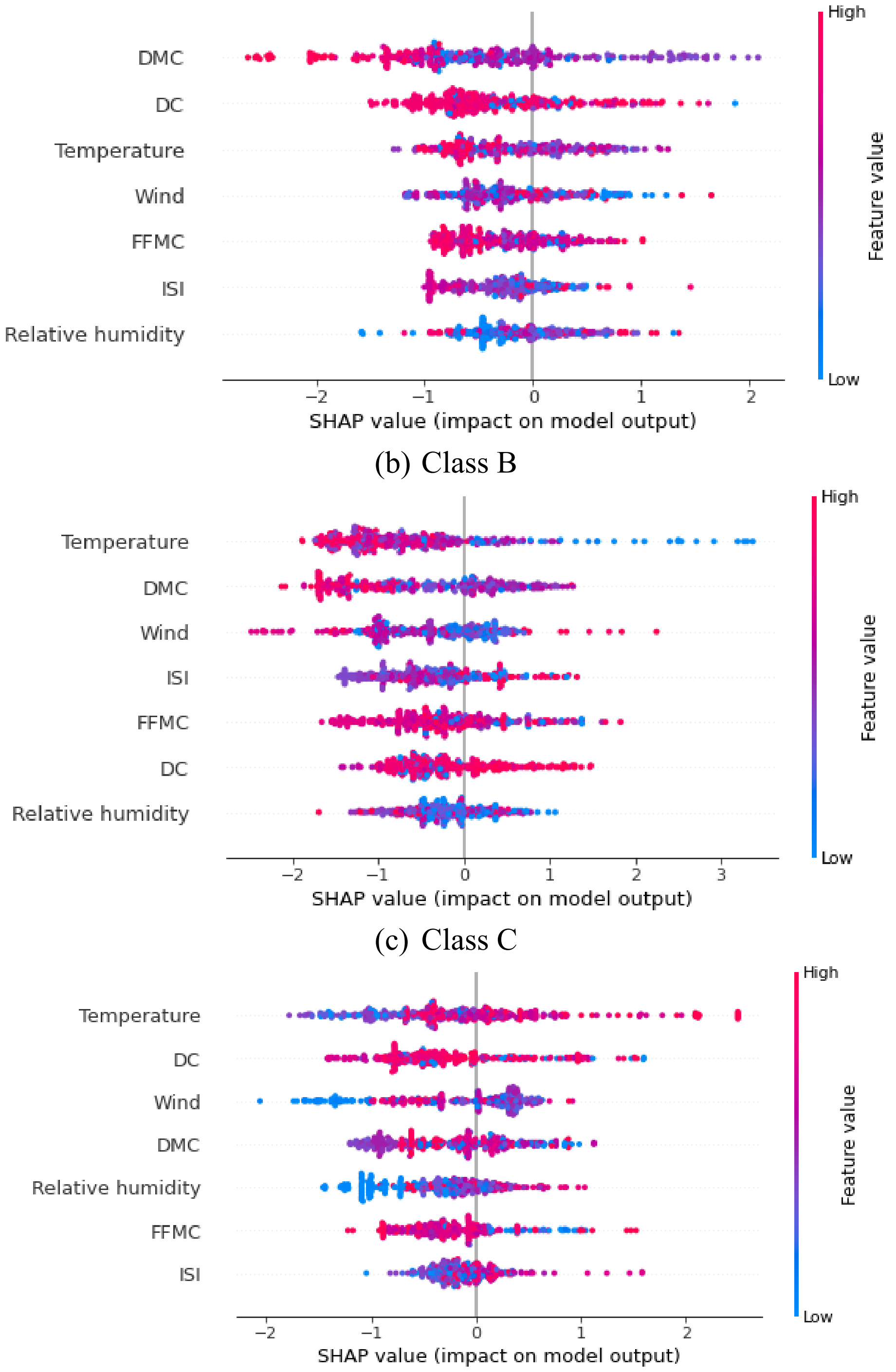


Fig. 8 Feature importance plot for all the classes (A-F).

To explain multi-class models with two or more classes, one needs to generate new features that are uniquely dedicated to these classes. This can be seen in terms of the *SHAP Summary Plot*, which represents both the feature importance and the direction each feature affects the model's class of wildfire. For instance, Fig. 9a represents class A's feature importance and the direction of each feature's effect in a specific class. One can see those high values of wind temperature and DMC negatively affects the occurrence of a wildfire in class A. the sub-figures shown in Fig. 9 represent the other classes B, C, D, E, and F, respectively.



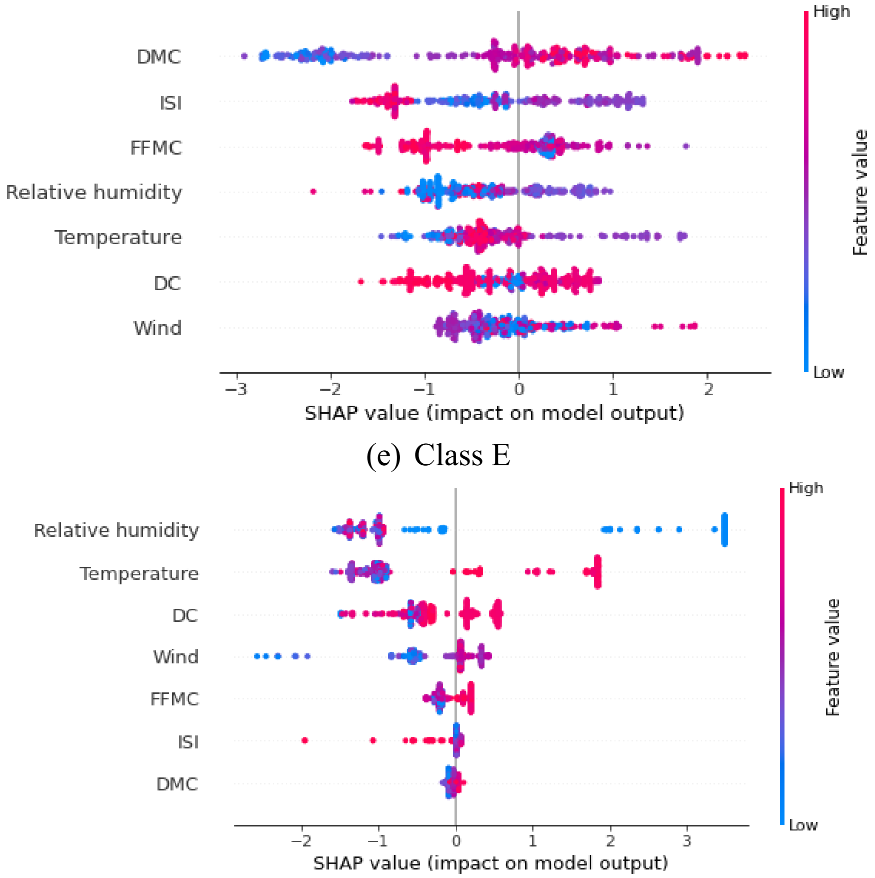
(a) Class A



(b) Class B

(c) Class C

(d) Class D



(e) Class E

(f) Class F

Fig. 9 SHAP summary plot for all the classes (A-F).

Closing Remarks on Wildfires Predictive and Classification Tools

It goes without saying that the accuracy in predictions obtained applying the dense data-driven (D^3) approach relies on the presence of information on correctly identified wildfires as well as properly documented parameters such as weather indices. While this study presents results obtained on two databases, one in the US and another one from Portugal, the reader should keep in mind that the presented approach can also be extended to other parts of the world as well as to encompasses a variety of input parameters [81]. This work infers that D^3 approach can lead to developing *support tools* that can aid the human-heavy decision making process, and we hope to explore such aspects in future work.

For example, if authorities are preparing for a wildfire season in the state of California, then they could possibly use the DL or DT tool to gauge the expected size of a wildfire, given that they input attributes comprising of: the discovery day of wildfire, latitude, and longitude of expected incident occurrence, and wildfire causes. Based on the outcome of the developed tools, the authorities will be able to estimate how many resources are expected to be allocated and deployed for such wildfires. One instance is given here as an illustration. In this scenario, a wildfire is expected to breakout on the 201st day of a given year in California in a location with longitude and latitude of -123.0 and 40.0, respectively. Based on the analysis from DL and DT tools, these tools show how that such wildfire is expected to be primarily of “B” size fire (based on observations from 1992-1994 and 2001-2009) with the potential to grow into a size “C” and beyond (based on observations collected between 1994-2001). While this estimation heavily relies on previous wildfire incidents still, it can be helpful to gauge the size and intensity of future wildfires with ease and in combination with currently used methods that utilize qualitative metrics and methods such as that shown in Fig. 10 [82,83].

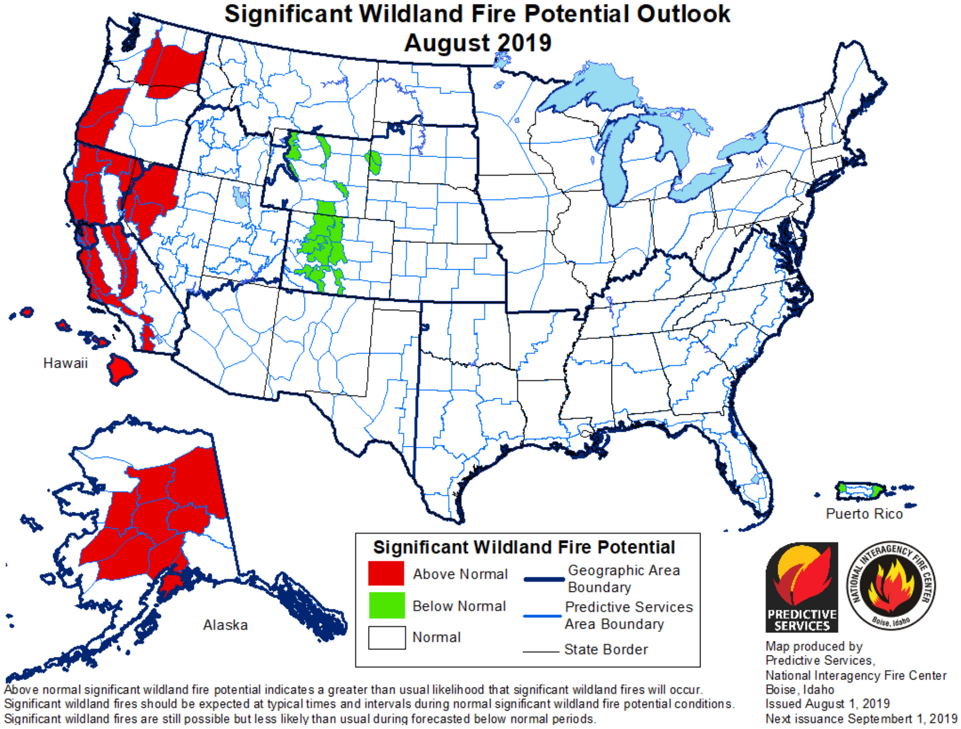

Significant Wildland Fire Potential Outlook
August 2019
Hawaii
Alaska
Puerto Rico
Significant Wildland Fire Potential
🟥 Above Normal
🟩 Below Normal
□ Normal
Geographic Area Boundary
Predictive Services Area Boundary
State Border
PREDICTIVE SERVICES
NATIONAL INTERAGENCY FIRE CENTER
Boise, Idaho
Above normal significant wildland fire potential indicates a greater than usual likelihood that significant wildland fires will occur.
Significant wildland fires should be expected at typical times and intervals during normal significant wildland fire potential conditions.
Significant wildland fires are still possible but less likely than usual during forecasted below normal periods.
Map produced by
Predictive Services,
National Interagency Fire Center
Boise, Idaho
Issued August 1, 2019
Next issuance September 1, 2019

Fig. 10 Predictions for wildland fire potential as obtained by the National Interagency Fire Center for the Month of August 2019

Similarly, the GA-derived expressions and explainable model for the case of wildfires occurring in Portugal can also be used to estimate the size of wildfires, given insights into weather indices. In this scenario, these expressions can be used to alarm authorities, occupants, and commuters in areas with high vulnerability to wildfire breakouts. This can also turn handy for preparedness and ensuring awareness in particular regions prone to wildfires. In all cases, the developed tools can be used as either predictive methods (i.e., to evaluate if a wildfire is expected to break out) or as classification methods (i.e., to estimate the size of an ongoing wildfire).

Finally, one should note that machine learning algorithms are adaptable and can improve by collecting new observations for analysis [84–86]. The proposed expressions/tools can also be designed to account for other attributes than those applied here. For example, future works are invited to explore adding attributes covering weather conditions, the magnitude of resource allocations (i.e., number of first responders, evacuation crews, etc.), expected damage to the environment (i.e., air quality, smoke/toxicity levels, fire spread, etc.) as well as to infrastructure (number of collapsed structures or bridges, etc.).

Conclusions

This work shows the merit of leveraging computational intelligence in order to develop predictive tools that are able to accurately predict the breakout and size of wildfires. More specifically, this paper explores the integration of deep learning (DL), decision tree (DT), Stochastic Gradient Descent (SGD), Extreme Gradient Boosted Trees (ExGBT), Logistic regression (LR), and genetic algorithms (GA) to gauge expected size of a wildfire given knowledge on existing geographical and environmental condition as well as human-based factors. In lieu of the above, the following conclusions can also be drawn from the findings of this study:

- Recent incidents have noted the increasing frequency and intensity of modern wildfires. As such, there is a need to properly predict the occurrence and size of such wildfires.
- Deep learning and decision tree algorithms seem to properly capture the wildfire phenomenon with accuracy exceeding 80%. On the other hand, genetic algorithms can also derive appropriate expressions that can be easily implemented into spreadsheets to predict the expected size of wildfires with good accuracy (*R* exceeding 80%). All these tools can potentially be implemented in practice to predict and classify wildfire sizes
- The use of explainable and symbolic ML can lead to realizing different types of transparent and equation-like tools to predict wildfires.
- The performance of the utilized algorithms herein (together with those to be developed in the near future) can be further enhanced with further training against properly documented wildfire observations, as well as historical information, etc.

Conflict of Interest: The authors declare that he has no conflict of interest.

Acknowledgment:

This work was performed under the following financial assistance award 70NANB17H213 from the U.S. Department of Commerce, National Institute of Standards and Technology. The authors are thankful for Prof. Nigel Kaye, the recipient of the above mentioned fund.

References

[1] A. Jaafari, E.K. Zenner, M. Panahi, H. Shahabi, Hybrid artificial intelligence models based on a neuro-fuzzy system and metaheuristic optimization algorithms for spatial prediction of wildfire probability, Agric. For. Meteorol. (2019). https://doi.org/10.1016/j.agrformet.2018.12.015.

[2] Insurance Information Institute, Facts + Statistics: Wildfires | III, (2020). https://www.iii.org/fact-statistic/facts-statistics-wildfires (accessed August 13, 2019).

[3] Union of Concerned Scientists, Infographic: Western Wildfires and Climate Change | Union of Concerned Scientists, (2013). https://www.ucsusa.org/global-warming/science-and-impacts/impacts/infographic-wildfires-climate-change.html (accessed August 2, 2020).

[4] J. Sherry, T. Neale, T.K. McGee, M. Sharpe, Rethinking the maps: A case study of knowledge incorporation in Canadian wildfire risk management and planning, J. Environ. Manage. (2019). https://doi.org/10.1016/j.jenvman.2018.12.116.

[5] T.B. Paveglio, C.M. Edgeley, A.M. Stasiewicz, Assessing influences on social vulnerability to wildfire using surveys, spatial data and wildfire simulations, J. Environ.

Manage. (2018). https://doi.org/10.1016/j.jenvman.2018.02.068.

[6] J. Handmer, R. Betts, Estimating the economic, social and environmental impacts of wildfires in Australia Catherine Stephenson1, Environ. Hazards. (2013). https://doi.org/10.1080/17477891.2012.703490.

[7] D. Paton, P.T. Buergelt, S. McCaffrey, F. Tedim, Wildfire Hazards, Risks, and Disasters, 2014. https://doi.org/10.1016/C2012-0-03331-5.

[8] K.-M. Hung, L.-M. Chen, T.-W. Chen, A Novel Hierarchical Wildfire Alarm System Based on Vegetation Features, J. Comput. 32 (2021) 137–151. https://doi.org/10.53106/199115992021083204011.

[9] A. Tohidi, N.B. Kaye, Stochastic modeling of firebrand shower scenarios, Fire Saf. J. (2017). https://doi.org/10.1016/j.firesaf.2017.04.039.

[10] R.A. Anthenien, S.D. Tse, A. Carlos Fernandez-Pello, On the trajectories of embers initially elevated or lofted by small scale ground fire plumes in high winds, Fire Saf. J. (2006). https://doi.org/10.1016/j.firesaf.2006.01.005.

[11] F. Hejazi, I. Toloue, M.S. Jaafar, J. Noorzaei, Optimization of earthquake energy dissipation system by genetic algorithm, Comput. Civ. Infrastruct. Eng. 28 (2013) 796–810. https://doi.org/10.1111/mice.12047.

[12] M.A. Haq, Planetscope Nanosatellites Image Classification Using Machine Learning, Comput. Syst. Sci. Eng. (2022). https://doi.org/10.32604/csse.2022.023221.

[13] M.A. Haq, P. Baral, S. Yaragal, B. Pradhan, Bulk processing of multi-temporal modis data, statistical analyses and machine learning algorithms to understand climate variables in the indian himalayan region, Sensors. (2021). https://doi.org/10.3390/s21217416.

[14] W. Mell, M.A. Jenkins, J. Gould, P. Cheney, A physics-based approach to modelling grassland fires, Int. J. Wildl. Fire. (2007). https://doi.org/10.1071/WF06002.

[15] A.M.G. Lopes, M.G. Cruz, D.X. Viegas, Firestation - An integrated software system for the numerical simulation of fire spread on complex topography, Environ. Model. Softw. (2002). https://doi.org/10.1016/S1364-8152(01)00072-X.

[16] B.R. Sturtevant, R.M. Scheller, B.R. Miranda, D. Shinneman, A. Syphard, Simulating dynamic and mixed-severity fire regimes: A process-based fire extension for LANDIS-II, Ecol. Modell. (2009). https://doi.org/10.1016/j.ecolmodel.2009.07.030.

[17] A. Bar Massada, A.D. Syphard, T.J. Hawbaker, S.I. Stewart, V.C. Radeloff, Effects of ignition location models on the burn patterns of simulated wildfires, Environ. Model. Softw. (2011). https://doi.org/10.1016/j.envsoft.2010.11.016.

[18] M. Van Kreveld, Geographic information systems, in: Handb. Discret. Comput. Geom. Third Ed., 2017. https://doi.org/10.1201/9781315119601.

[19] M. Finney, I.C. Grenfell, C.W. McHugh, Modeling containment of large wildfires using generalized linear mixed-model analysis, For. Sci. (2009). https://doi.org/10.1093/forestscience/55.3.249.

[20] I. Kochi, P.A. Champ, J.B. Loomis, G.H. Donovan, Valuing mortality impacts of smoke exposure from major southern California wildfires, J. For. Econ. (2012). https://doi.org/10.1016/j.jfe.2011.10.002.

[21] H. Xue, F. Gu, X. Hu, Data assimilation using sequential monte carlo methods in wildfire spread simulation, ACM Trans. Model. Comput. Simul. (2012). https://doi.org/10.1145/2379810.2379816.

[22] T.E. Dilts, J.S. Sibold, F. Biondi, A weights-of-evidence model for mapping the

probability of fire occurrence in lincoln county, Nevada, Ann. Assoc. Am. Geogr. (2009). https://doi.org/10.1080/00045600903066540.

[23] J.T. Abatzoglou, T.J. Brown, A comparison of statistical downscaling methods suited for wildfire applications, Int. J. Climatol. (2012). https://doi.org/10.1002/joc.2312.

[24] M. Castelli, L. Vanneschi, A. Popovič, Predicting burned areas of forest fires: An artificial intelligence approach, Fire Ecol. (2015). https://doi.org/10.4996/fireecology.1101106.

[25] M. Rodrigues, J. De la Riva, An insight into machine-learning algorithms to model human-caused wildfire occurrence, Environ. Model. Softw. (2014). https://doi.org/10.1016/j.envsoft.2014.03.003.

[26] D. Tien Bui, Q.T. Bui, Q.P. Nguyen, B. Pradhan, H. Nampak, P.T. Trinh, A hybrid artificial intelligence approach using GIS-based neural-fuzzy inference system and particle swarm optimization for forest fire susceptibility modeling at a tropical area, Agric. For. Meteorol. (2017). https://doi.org/10.1016/j.agrformet.2016.11.002.

[27] F. Zhang, P. Zhao, J. Thiyagalingam, T. Kirubarajan, Terrain-influenced incremental watchtower expansion for wildfire detection, Sci. Total Environ. (2019). https://doi.org/10.1016/j.scitotenv.2018.11.038.

[28] Q. Zhao, S. Yu, F. Zhao, L. Tian, Z. Zhao, Comparison of machine learning algorithms for forest parameter estimations and application for forest quality assessments, For. Ecol. Manage. (2019). https://doi.org/10.1016/j.foreco.2018.12.019.

[29] J.A. Blackard, D.J. Dean, Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables, Comput. Electron. Agric. (1999). https://doi.org/10.1016/S0168-1699(99)00046-0.

[30] S. Sachdeva, T. Bhatia, A.K. Verma, GIS-based evolutionary optimized Gradient Boosted Decision Trees for forest fire susceptibility mapping, Nat. Hazards. (2018). https://doi.org/10.1007/s11069-018-3256-5.

[31] D.T. Bui, K.T.T. Le, V.C. Nguyen, H.D. Le, I. Revhaug, Tropical forest fire susceptibility mapping at the Cat Ba National Park area, Hai Phong City, Vietnam, using GIS-based Kernel logistic regression, Remote Sens. (2016). https://doi.org/10.3390/rs8040347.

[32] D.E. King, Dlibml: A Machine Learning Toolkit, J. Mach. Learn. Res. (2009).

[33] R. Collobert, K. Kavukcuoglu, C. Farabet, Torch7: A Matlab-like Environment for Machine Learning, 2011.

[34] J. Yao, M. Brauer, S. Raffuse, S.B. Henderson, Machine Learning Approach to Estimate Hourly Exposure to Fine Particulate Matter for Urban, Rural, and Remote Populations during Wildfire Seasons, Environ. Sci. Technol. (2018). https://doi.org/10.1021/acs.est.8b01921.

[35] J. Rogan, J. Franklin, D. Stow, J. Miller, C. Woodcock, D. Roberts, Mapping land-cover modifications over large areas: A comparison of machine learning algorithms, Remote Sens. Environ. (2008). https://doi.org/10.1016/j.rse.2007.10.004.

[36] Dimensions, Dimensions.ai, (2021). https://www.dimensions.ai/.

[37] M. Thelwall, Dimensions: A competitor to Scopus and the Web of Science?, J. Informetr. (2018). https://doi.org/10.1016/j.joi.2018.03.006.

[38] M.Z. Naser, Mapping functions: A physics-guided, data-driven and algorithm-agnostic machine learning approach to discover causal and descriptive expressions of engineering phenomena, Measurement. 185 (2021) 110098. https://doi.org/10.1016/J.MEASUREMENT.2021.110098.

[39] D. Bailey, WUI Fact Sheet, 2013. http://www.iawfonline.org/pdf/WUI_Fact_Sheet_08012013.pdf.

[40] CALFIRE, 2020 Fire Season - CALFIRE, (2020). https://www.fire.ca.gov/incidents/2020/ (accessed November 16, 2020).

[41] D. Tin, A.J. Hertelendy, G.R. Ciottone, What we learned from the 2019–2020 Australian Bushfire disaster: Making counter-terrorism medicine a strategic preparedness priority, Am. J. Emerg. Med. (2020). https://doi.org/10.1016/j.ajem.2020.09.069.

[42] National Interagency Fire Center, Wildland Fire Fatalities by Year, 2017. https://www.nifc.gov/safety/safety_documents/Fatalities-by-Year.pdf (accessed November 16, 2020).

[43] BBC, Australia fires: A visual guide to the bushfire crisis - BBC News, BBC News. (2020). https://www.bbc.com/news/world-australia-50951043 (accessed November 16, 2020).

[44] WWF, Australian Bushfires - WWF-Australia - WWF-Australia, (2020). https://www.wwf.org.au/what-we-do/bushfire-recovery/bushfires#gs.lm0sfk (accessed November 16, 2020).

[45] K. Short, Spatial wildfire occurrence data for the United States, 1992-2015, 2017. https://doi.org/https://doi.org/10.2737/RDS-2013-0009.4.

[46] P. Cortez, A. Morais, A Data Mining Approach to Predict Forest Fires using Meteorological Data, in: Proc. 13th Port. Conf. Artif. Intell., 2007.

[47] The National Wildfire Coordinating Group, NWCG | NWCG is an operational group designed to coordinate programs of the participating wildfire management agencies., (n.d.). https://www.nwcg.gov/ (accessed August 13, 2019).

[48] R. Tatman, 1.88 Million US Wildfires, Kaggle.Com. (2017). https://www.kaggle.com/rtatman/188-million-us-wildfires (accessed August 6, 2019).

[49] P. Cortez, A. Morais, Forest Fires Data Set, UCI Mach. Learn. Repos. (2008). https://archive.ics.uci.edu/ml/datasets/forest+fires (accessed August 6, 2019).

[50] P. Cortez, A. Morais, Forest Fires Data Set Portugal | Kaggle, (2007). https://www.kaggle.com/datasets/ishandutta/forest-fires-data-set-portugal (accessed July 11, 2022).

[51] B.J. Stocks, T.J. Lynham, B.D. Lawson, M.E. Alexander, C.E. Van Wagner, R.S. McAlpine, D.E. Dubé, The Canadian Forest Fire Danger Rating System: An Overview, For. Chron. (1989). https://doi.org/10.5558/tfc65450-6.

[52] M. Naser, G. Abu-Lebdeh, R. Hawileh, Analysis of RC T-beams strengthened with CFRP plates under fire loading using ANN, Constr. Build. Mater. 37 (2012) 301–309. https://doi.org/10.1016/j.conbuildmat.2012.07.001.

[53] S.R. Safavian, D. Landgrebe, A survey of decision tree classifier methodology, IEEE Trans. Syst. Man. Cybern. 21 (1991) 660–674. https://doi.org/10.1109/21.97458.

[54] SciKit, Decision Tree, (2020). https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.ExtraTreesClassifier.html?highlight=decision tree#sklearn.ensemble.ExtraTreesClassifier.decision_path (accessed January 22, 2021).

[55] D. Che, Q. Liu, K. Rasheed, X. Tao, Decision Tree and Ensemble Learning Algorithms with Their Applications in Bioinformatics, in: Springer, New York, NY, 2011: pp. 191–199. https://doi.org/10.1007/978-1-4419-7046-6_19.

[56] J.-S.S. Chou, C.-F.F. Tsai, A.-D.D. Pham, Y.-H.H. Lu, Machine learning in concrete strength simulations: Multi-nation data analytics, Constr. Build. Mater. 73 (2014) 771–780. https://doi.org/10.1016/j.conbuildmat.2014.09.054.

[57] SciKit, SGD Classifier , (2020). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDClassifier.html (accessed January 22, 2021).

[58] Y. Freund, R.E. Schapire, A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, J. Comput. Syst. Sci. (1997). https://doi.org/10.1006/jcss.1997.1504.

[59] Gradient boosted tree (GBT), (2019). https://software.intel.com/en-us/daal-programming-guide-details-24 (accessed April 9, 2019).

[60] SciKit, Logistic Regression , (2020). https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html?highlight=logistic regression#sklearn.linear_model.LogisticRegression (accessed January 22, 2021).

[61] Y. Jafari Goldarag, A. Mohammadzadeh, A.S. Ardakani, Fire Risk Assessment Using Neural Network and Logistic Regression, J. Indian Soc. Remote Sens. (2016). https://doi.org/10.1007/s12524-016-0557-6.

[62] F. Bunea, Honest variable selection in linear and logistic regression models via l1 and l1+l2 penalization, Electron. J. Stat. (2008). https://doi.org/10.1214/08-EJS287.

[63] D.E. Goldberg, J.H. Holland, Genetic Algorithms and Machine Learning, Mach. Learn. (1988). https://doi.org/10.1023/A:1022602019183.

[64] J.R. Koza, A genetic approach to finding a controller to back up a tractor-trailer truck, in: Proc. 1992 Am. Control Conf., 1992.

[65] M.Z. Naser, Heuristic machine cognition to predict fire-induced spalling and fire resistance of concrete structures, Autom. Constr. 106 (2019) 102916. https://doi.org/10.1016/J.AUTCON.2019.102916.

[66] M.Z. Naser, AI-based cognitive framework for evaluating response of concrete structures in extreme conditions, Eng. Appl. Artif. Intell. 81 (2019) 437–449. https://www.sciencedirect.com/science/article/pii/S0952197619300466 (accessed April 1, 2019).

[67] A.H. Alavi, A.H. Gandomi, M.G. Sahab, M. Gandomi, Multi expression programming: A new approach to formulation of soil classification, Eng. Comput. 26 (2010) 111–118. https://doi.org/10.1007/s00366-009-0140-7.

[68] C. Ferreira, Gene Expression Programming: a New Adaptive Algorithm for Solving Problems, Ferreira, C. (2001). Gene Expr. Program. a New Adapt. Algorithm Solving Probl. Complex Syst. 13. (2001). https://www.semanticscholar.org/paper/Gene-Expression-Programming%3A-a-New-Adaptive-for-Ferreira/3232b2a24c2584ca8e81cb5bf6f55aef34f0aefe (accessed March 16, 2019).

[69] S.L. Oh, V. Jahmunah, C.P. Ooi, R.S. Tan, E.J. Ciaccio, T. Yamakawa, M. Tanabe, M. Kobayashi, U. Rajendra Acharya, Classification of heart sound signals using a novel deep WaveNet model, Comput. Methods Programs Biomed. (2020). https://doi.org/10.1016/j.cmpb.2020.105604.

[70] J. Abawajy, A. Kelarev, X. Yi, H.F. Jelinek, Minimal ensemble based on subset selection using ECG to diagnose categories of CAN, Comput. Methods Programs Biomed. (2018).

https://doi.org/10.1016/j.cmpb.2018.01.019.

[71] D. Searson, D. Searson, GPTIPS Genetic Programming &amp; Symbolic Regression for MATLAB User Guide, (2009). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.177.494 (accessed January 22, 2019).

[72] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, E. Duchesnay, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[73] GMDH, GMDH Shell DS, (2019). https://gmdhsoftware.com/.

[74] M. Sheikholeslami, F. Bani Sheykholeslami, S. Khoshhal, H. Mola-Abasia, D.D. Ganji, H.B. Rokni, Effect of magnetic field on Cu-water nanofluid heat transfer using GMDH-type neural network, Neural Comput. Appl. (2014). https://doi.org/10.1007/s00521-013-1459-y.

[75] M.Z. Naser, A.H. Alavi, · Amir, H. Alavi, Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences, Archit. Struct. Constr. 1 (2021). https://doi.org/https://doi.org/10.1007/s44150-021-00015-8.

[76] D. Radke, A. Hessler, D. Ellsworth, Firecast: Leveraging deep learning to predict wildfire spread, in: IJCAI Int. Jt. Conf. Artif. Intell., 2019. https://doi.org/10.24963/ijcai.2019/636.

[77] Z. Langford, J. Kumar, F. Hoffman, Wildfire mapping in interior alaska using deep neural networks on imbalanced datasets, in: IEEE Int. Conf. Data Min. Work. ICDMW, 2019. https://doi.org/10.1109/ICDMW.2018.00116.

[78] P. Jain, S.C.P. Coogan, S.G. Subramanian, M. Crowley, S.W. Taylor, M.D. Flannigan, A review of machine learning applications in wildfire science and management, Environ. Rev. (2020). https://doi.org/10.1139/er-2020-0019.

[79] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: Adv. Neural Inf. Process. Syst., 2017.

[80] N. V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. (2002). https://doi.org/10.1613/jair.953.

[81] Y.O. Sayad, H. Mousannif, H. Al Moatassime, Predictive modeling of wildfires: A new dataset and machine learning approach, Fire Saf. J. (2019). https://doi.org/10.1016/J.FIRESAF.2019.01.006.

[82] G.A. Trunfio, Predicting Wildfire Spreading Through a Hexagonal Cellular Automata Model, in: 2004. https://doi.org/10.1007/978-3-540-30479-1_40.

[83] C. Vega-García, E. Chuvieco, Applying local measures of spatial heterogeneity to Landsat-TM images for predicting wildfire occurrence in Mediterranean landscapes, Landsc. Ecol. (2006). https://doi.org/10.1007/s10980-005-4119-5.

[84] Y. Xie, M. Peng, Forest fire forecasting using ensemble learning approaches, Neural Comput. Appl. (2019). https://doi.org/10.1007/s00521-018-3515-0.

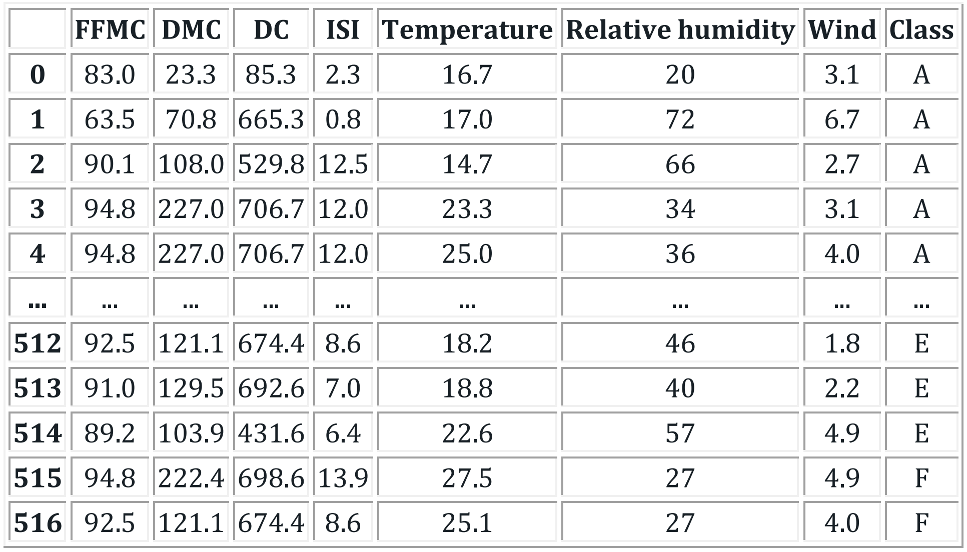
[85] A. Erdil, E. Arcaklioglu, The prediction of meteorological variables using artificial neural network, Neural Comput. Appl. (2013). https://doi.org/10.1007/s00521-012-1210-0.

[86] M. Liu, S.M. Lo, B.Q. Hu, C.M. Zhao, On the use of fuzzy synthetic evaluation and optimal classification for computing fire risk ranking of buildings, Neural Comput. Appl. (2009). https://doi.org/10.1007/s00521-009-0244-4.

Appendix

Here is our code. The database can be found at [46] and [50].

```
import sklearn
from sklearn.model_selection import train_test_split
import pandas as pd
import numpy as np
import shap
import xgboost as xgb
from matplotlib import pyplot
from sklearn.metrics import accuracy_score
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import accuracy_score
from imblearn.over_sampling import SMOTE
from imblearn.over_sampling import BorderlineSMOTEIn
wildfire=pd.read_excel('Portugal ABCDEF.xlsx')
wildfire
```



	FFMC	DMC	DC	ISI	Temperature	Relative humidity	Wind	Class
0	83.0	23.3	85.3	2.3	16.7	20	3.1	A
1	63.5	70.8	665.3	0.8	17.0	72	6.7	A
2	90.1	108.0	529.8	12.5	14.7	66	2.7	A
3	94.8	227.0	706.7	12.0	23.3	34	3.1	A
4	94.8	227.0	706.7	12.0	25.0	36	4.0	A
...	...	...	...	...	...	...	...	...
512	92.5	121.1	674.4	8.6	18.2	46	1.8	E
513	91.0	129.5	692.6	7.0	18.8	40	2.2	E
514	89.2	103.9	431.6	6.4	22.6	57	4.9	E
515	94.8	222.4	698.6	13.9	27.5	27	4.9	F
516	92.5	121.1	674.4	8.6	25.1	27	4.0	F

517 rows × 8 columns

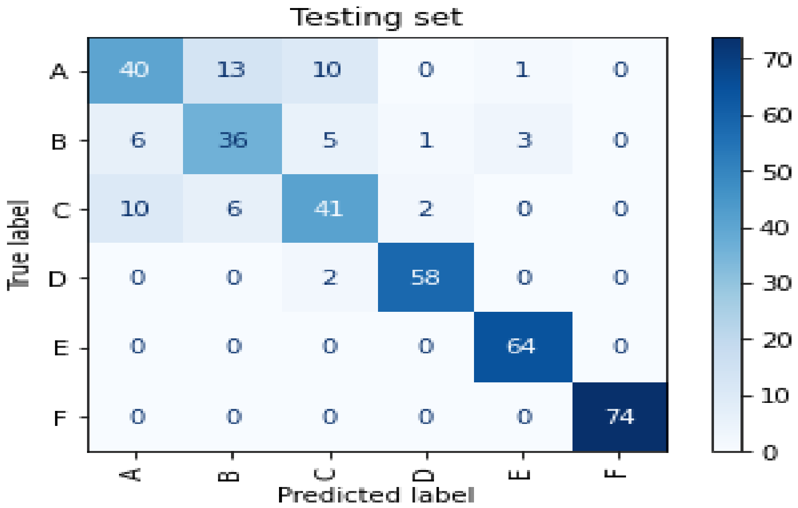
```
x=wildfire.drop(['Class'],axis=1)
y=wildfire['Class']
oversampled = SMOTE(sampling_strategy='auto',
            random_state=5,k_neighbors = 1
            )
x, y = oversampled.fit_resample(x, y)
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.250,random_state=10)

y_train.value_counts()
wildfire.info()
wildfire.isnull().any()
```

```
FFMC              False
DMC               False
DC               False
ISI             False
Temperature          False
Relative humidity    False
Wind              False
Class             False
dtype: bool
```

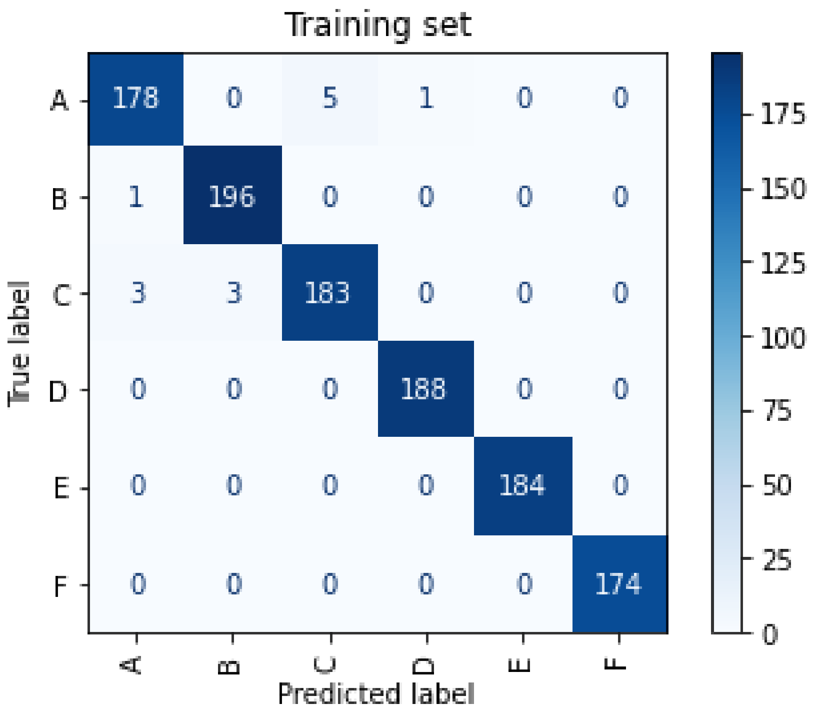
```
xgbc=xgb.XGBClassifier(objective='multi:softprob',
                      learning_rate =0.6,
                       n_estimators=800,
                       max_depth=6,
                       min_child_weight=0,
                       gamma=0.2,
                       subsample=0.9,
                       colsample_bytree=0.7,
                       nthread=40,
                        seed=230)
xgbc.fit(x_train,y_train)
predictions = xgbc.predict(x_test)
accuracy = accuracy_score(y_test, predictions)
print("Accuracy: %.2f%%" % (accuracy * 100.0))
```

```
Accuracy: 84.14%
```

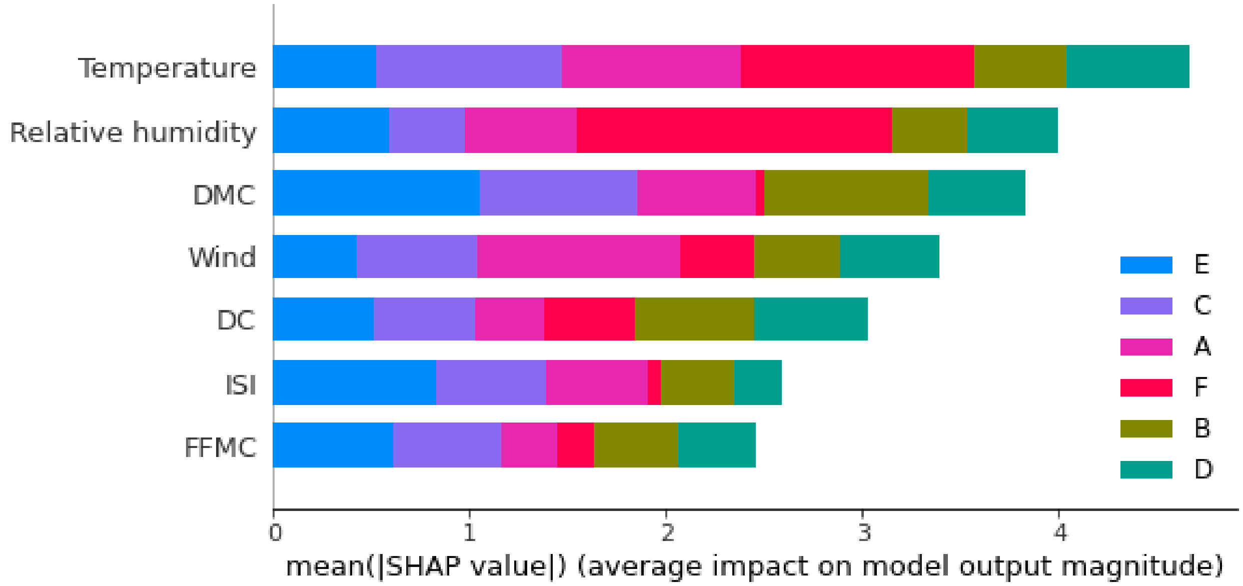
```
class_names = ['A', 'B', 'C','D','E','F']
disp = plot_confusion_matrix(xgbc, x_test, y_test, display_labels=class_names, cmap=pyplot.cm.Blues,
xticks_rotation='vertical')
pyplot.title('Testing set')
```



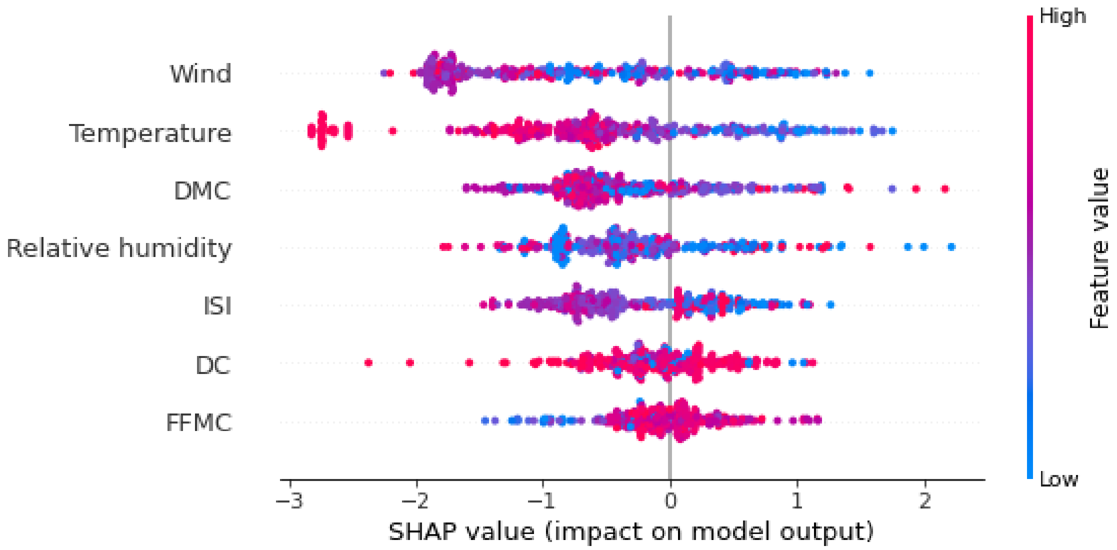
```
disp = plot_confusion_matrix(xgbc, x_train, y_train, display_labels=class_names, cmap=pyplot.cm.Blues,
xticks_rotation='vertical')
pyplot.title('Training set')
```



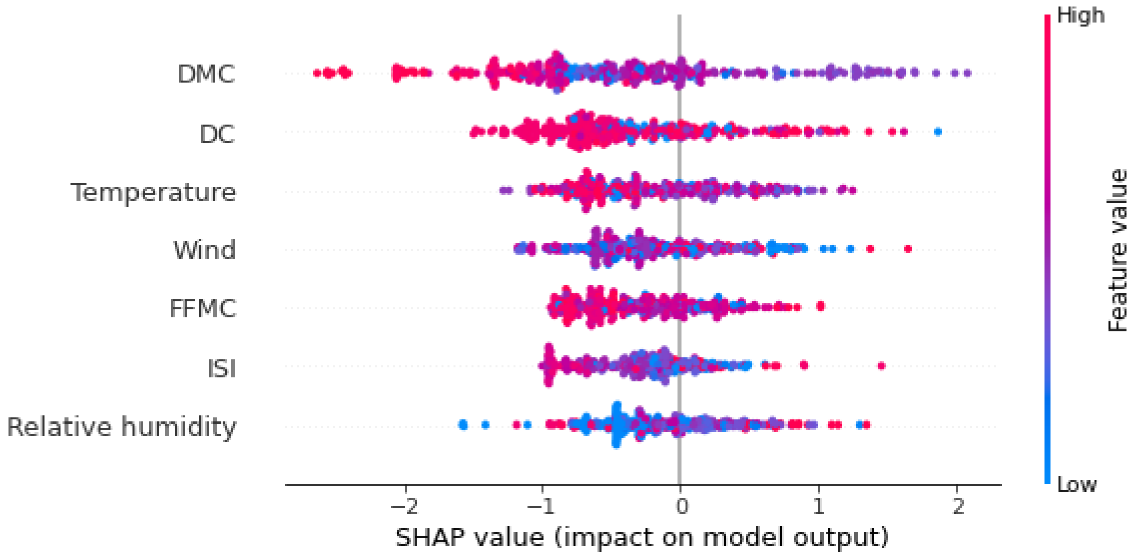
```
shap_values = shap.TreeExplainer(xgbc).shap_values(x_test)
                shap.summary_plot(shap_values, x_test,class_names = class_names, plot_type='bar')
```



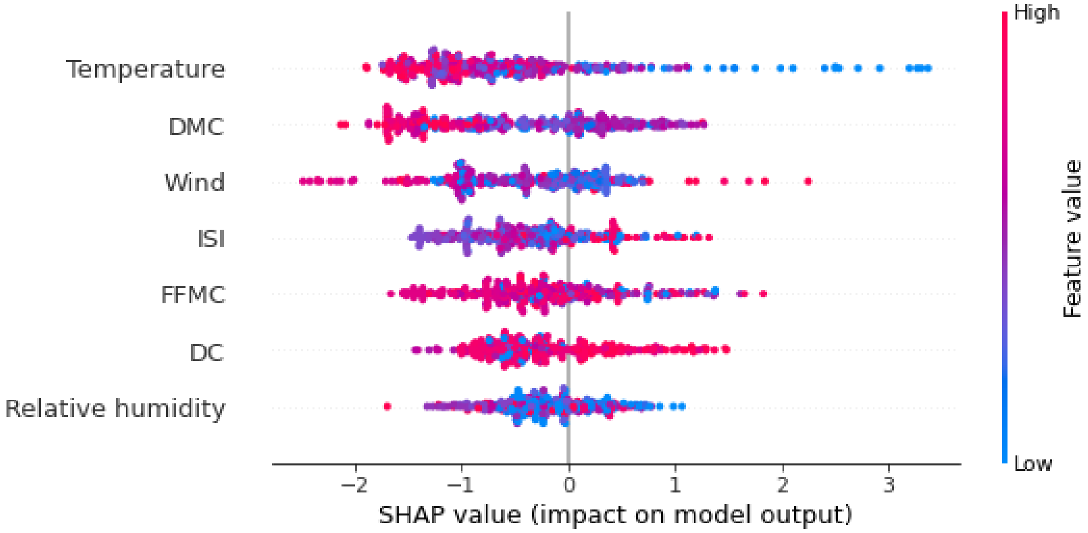
```
shap.summary_plot(shap_values[0], x_test, class_names=class_names,show=False)
pyplot.gcf().axes[-1].set_box_aspect(50)
pyplot.gcf().axes[-1].set_aspect(100)
pyplot.gcf().axes[-1].set_box_aspect(100)
```



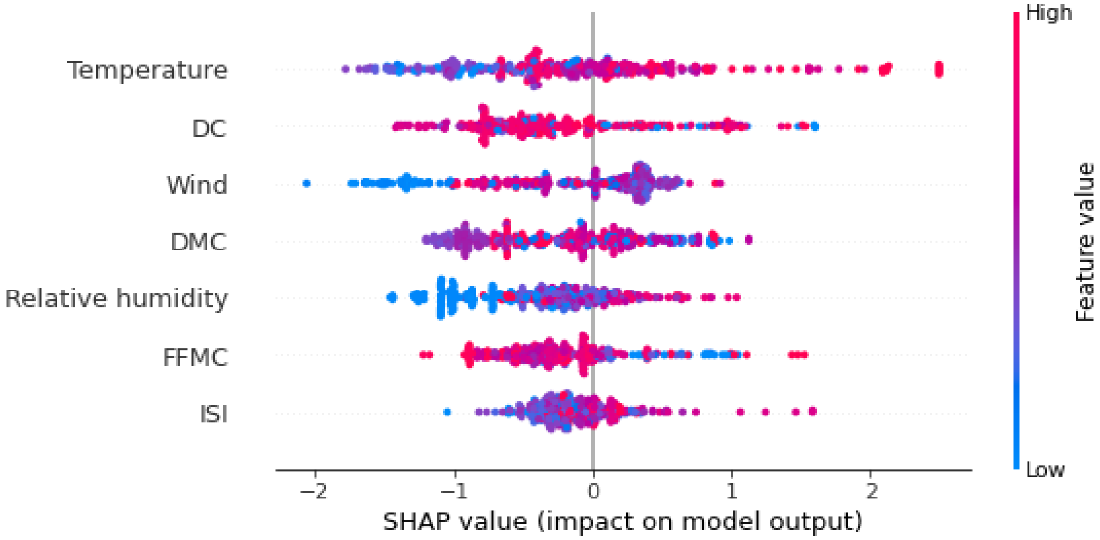
```
shap.summary_plot(shap_values[1], x_test, class_names=class_names,show=False)
pyplot.gcf().axes[-1].set_box_aspect(50)
pyplot.gcf().axes[-1].set_aspect(100)
pyplot.gcf().axes[-1].set_box_aspect(100)
```



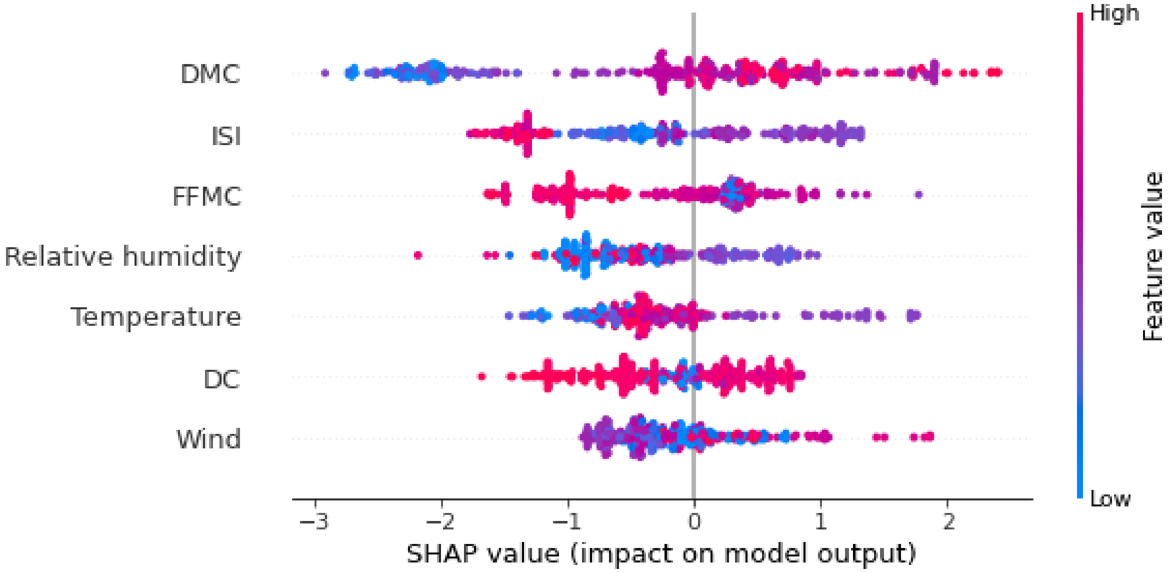
```
shap.summary_plot(shap_values[2], x_test, class_names=class_names,show=False)
pyplot.gcf().axes[-1].set_box_aspect(50)
pyplot.gcf().axes[-1].set_aspect(100)
pyplot.gcf().axes[-1].set_box_aspect(100)
```



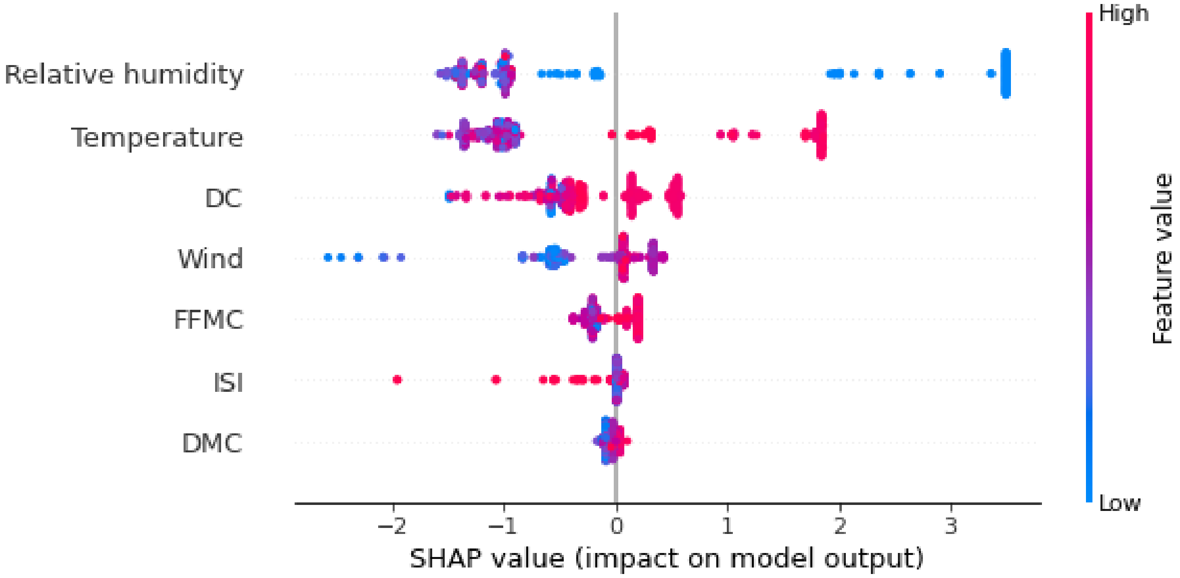
```
shap.summary_plot(shap_values[3], x_test, class_names=class_names,show=False)
pyplot.gcf().axes[-1].set_box_aspect(50)
pyplot.gcf().axes[-1].set_aspect(100)
pyplot.gcf().axes[-1].set_box_aspect(100)
```



```
shap.summary_plot(shap_values[4], x_test, class_names=class_names,show=False)
pyplot.gcf().axes[-1].set_box_aspect(50)
pyplot.gcf().axes[-1].set_aspect(100)
pyplot.gcf().axes[-1].set_box_aspect(100)
```



```
shap.summary_plot(shap_values[5], x_test, class_names=class_names,show=False)
pyplot.gcf().axes[-1].set_box_aspect(50)
pyplot.gcf().axes[-1].set_aspect(100)
                                       pyplot.gcf().axes[-1].set_box_aspect(100)
```


Relative humidity
Temperature
DC
Wind
FFMC
ISI
DMC
SHAP value (impact on model output)
High
Feature value
Low