Abstract: Wildfires generate substantial smoke containing fine particulate matter (PM$_{2.5}$) that adversely impacts health. This study develops machine learning models integrating pre-wildfire factors like weather and fuel conditions with post-wildfire health impacts to provide a holistic understanding of smoke exposure risks. Various data-driven models including Support Vector Regression, Multi-layer Perceptron, and three tree-based ensemble algorithms (Random Forest, Extreme Gradient Boosting (XGBoost), and Natural Gradient Boosting (NGBoost)) are evaluated in this study. Ensemble models effectively predict PM$_{2.5}$ levels based on temperature, humidity, wind, and fuel moisture, revealing the significant roles of radiation, temperature, and moisture. Further modelling links smoke exposure to deaths from chronic obstructive pulmonary disease (COPD) and lung cancer using age, sex, and pollution type as inputs. Ambient pollution is the primary driver of COPD mortality, while age has a greater influence on lung cancer deaths. This research advances atmospheric and health impact understanding, aiding forest fire prevention and management.

Keywords: wildfires; machine learning; air pollution; artificial intelligence; chronic obstructive pulmonary disease (COPD); tracheal; bronchus; lung cancer (TB&L)

1. Introduction

Wildfires rank among the primary global natural hazards, presenting an ever-mounting worldwide peril to human lives, safety, and the environment [1]. The hot and dry weather conditions can impact wildfires directly and indirectly by providing fuels with low moisture content that is suitable for wildfires [2]. The evidence shows that climate change will increase the severity and frequency of wildfires [3,4]. Amidst the persistent trend of climate warming, it is anticipated that fuel moisture content will decrease, exacerbating fire weather conditions and elevating both the severity and frequency of wildfires [5,6].

As an example of wildfire negative impacts, the direct impact of the 2019–2020 wildfires in Australia burned over 19.4 million hectares, causing 34 direct deaths [7], 429 indirect deaths, and 3230 hospital admissions due to respiratory issues [8].

To efficiently oversee complex systems such as wildfires and minimize their direct and indirect negative consequences, recognizing the interconnections is crucial. Previous studies tried to individually evaluate the consequences of the wildfires on society (i.e., wildfires’ health impact due to the air quality [9], monitoring indoor, outdoor, and personal PM$_{2.5}$ exposure during a wildfire season [10], wildfire smoke impacts on asthma [11], fire smoke detection from satellite imagery [12–14], drones wildfire detection [15], wildfire air quality on Australian schools [16], and developing machine learning models to predict air quality [17]). Despite growing research on wildfires, studies examining the interconnectedness between pre-fire conditions and post-fire impacts remain limited. This study helps address this gap by integrating diverse factors related to wildfires, including pre-fire fuel...
and weather conditions and post-fire health effects. It is important to recognize that this matter is highly intricate, and there is a multitude of potential variables associated with wildfires. Within the scope of this study, we have considered specific factors, including fuel and weather variables, indicators of smoke, as well as the number and characteristics of deaths linked to smoke exposure. Using an interdisciplinary approach, we analyzed the complex interrelationships between these multiple spheres linked to wildfires. This novel perspective strengthens the understanding of cascading wildfire effects, supporting improved prediction and management under a changing climate. By anticipating weather and fuel variables sooner, timely public health interventions, including proactive evacuations for susceptible populations, can effectively reduce the risk of mortality.

This study integrates multiple factors related to wildfires, encompassing fuel characteristics, weather attributes, indicators of wildfire smoke exposure, and the associated health impacts. The health impact of wildfires is related to thick and heavy smoke that adversely impacts people and wildlife [18–20]. The cloud of smoke contains wildfire-related fine particulate matter with a diameter $\leq 2.5 \mu g/m^3$ (PM$_{2.5}$) [21]. Past studies demonstrated that PM$_{2.5}$ is responsible for exacerbating chronic obstructive pulmonary disease (COPD) [22] and lung cancer [23]. Moreover, in comparison with other sources of air pollution such as traffic, short-term exposure to wildfire-related PM$_{2.5}$ can be responsible for health impact as well [24]. In this study, firstly, we constructed five data-driven regression models to predict the wildfire smoke indicator, specifically PM$_{2.5}$, based on a comprehensive set of fuel and weather indicators. These indicators encompass fuel moisture, air temperature, wind speed, rainfall, solar radiation, and relative humidity. The modelling results highlight the contribution of each input variable to the PM$_{2.5}$ levels and provide actionable recommendations for considering air pollution resulting from future wildfires by effectively managing pre-disaster parameters. Second, we conducted an analysis of the impact of air pollution, specifically PM$_{2.5}$, on the number of fatalities associated with COPD and tracheal, bronchus, and lung cancer (TB&L). PM$_{2.5}$ exacerbates lung cancer by altering genes, causing inflammation, and impacting immune responses. It triggers tumor growth and affects cell survival through oxidative stress and changes in gene activity. This analysis takes into account additional factors such as sex, age, year, and type of PM$_{2.5}$ exposure (ambient or household). Additionally, our study estimates the relative importance of each input variable in determining the number of fatalities. By simultaneously considering both the pre- and post-disaster factors, our research enables a comprehensive exploration of the interdependencies between these factors. This approach allows for a deeper understanding of the complex relationship between pre-disaster conditions, such as fuel and weather indicators, and the subsequent health impacts and fatalities resulting from air pollution. In this study, pre-exposure refers to the time period before the wildfire starts, and post-exposure is the time period after the wildfire occurs. Both the short- and long-term health impacts of wildfires are considered as a whole in this study based on the dataset. Pre-disaster parameters are mostly naturally occurring ambient conditions that may not be manipulated. However, understanding their impact on PM$_{2.5}$ helps experts predict and be prepared for air pollution resulting from wildfires. The connection between pre-fire and post-fire variables in this study is air pollution, quantified by using PM$_{2.5}$ levels. Pre-disaster variables impact PM$_{2.5}$ concentrations, and elevated PM$_{2.5}$ levels subsequently affect people’s health. This thorough method is essential for understanding and mitigating the health impacts of wildfire-related air pollution. The most important novelty of this study lies in its comprehensive consideration of parameters both before and after wildfires to establish connections between weather, fuel variables, and health impacts. Additionally, the study introduces a newly collected dataset tailored for this purpose. Novel methodologies include the application of state-of-the-art machine learning models such as ensemble methods and SHAP for model interpretation. These innovations contribute significantly to the advancement of understanding and prediction capabilities in wildfire-related health outcomes.
2. Methods and Materials

2.1. Dataset Description

Wagga Wagga (35.10° S, 147.38° E) is selected for this section of this study due to the fact that Wagga Wagga holds significance in the context of wildfires in grasslands [25,26]. Located in New South Wales, Australia, Wagga Wagga is situated within a region prone to bushfires and wildfires. The area is characterized by a combination of factors that contribute to the increased risk and severity of wildfire events. These factors include a dry and arid climate, abundant vegetation, and strong winds, creating an environment conducive to the rapid spread of fires. The proximity of Wagga Wagga to national parks, forests, and rural regions significantly enhances the susceptibility to wildfires due to the availability of suitable fuels. Given that grassfires loss is a significant factor contributing to wildfires in Australia [27], and considering the abundance of grassland in Wagga Wagga, which can produce substantial smoke [28], particularly during grassfires, this location has been chosen for the study.

Weather variables for the Wagga Wagga region in Australia are gathered and evaluated to analyze their impact on the occurrence and behavior of smoke. Data from the Bureau of Meteorology (BoM) website in Wagga Wagga were gathered, encompassing daily records of air temperature, rainfall, and solar radiation spanning from 1 June 2019 to 31 May 2020. Monthly records of relative humidity and wind speed from the BoM during this timeframe are also incorporated into the dataset. In light of the presence of grasslands in this region, fuel moisture is estimated using operational methods reliant on environmental factors. The selection of this particular year is attributable to the occurrence of significant wildfires during that period [8]. It is essential to emphasize that the first research question’s time frame is limited to one year, which may not provide a sufficient window to observe lung-cancer-related deaths but rather pertains to cancer exacerbations. Therefore, for the second research question, we extended the timeline to encompass lung cancer considerations in connection with smoke exposure.

Fuel variables hold significant importance due to their crucial role in the generation and behavior of wildfire smoke [29]. One crucial fuel variable extensively considered is fuel moisture content [30]. Fuel moisture content represents the amount of water present in the fuel, such as vegetation, trees, and organic matter, that can potentially serve as fuel for wildfires. Higher moisture content can inhibit the ignition and spread of fires [31], while lower moisture content increases the likelihood and severity of wildfire events. Fuel moisture content for grasslands near the Wagga Wagga region is estimated based on the following equation [32]:

\[ M_{dl} = 9.58 - 0.205T + 0.138RH \]  

where \( M_{dl} \) is fuel moisture; \( T \) is temperature; and \( RH \) is relative humidity. PM\(_{2.5}\) is a particulate matter in air pollution with a diameter of 2.5 micrometers or smaller. PM\(_{2.5}\) measurement serves as an important air quality index for evaluating the smoke generated by wildfires. PM\(_{2.5}\) particles are tiny and can easily penetrate deep into the respiratory system when inhaled, posing significant health risks to humans [33]. The NSW Department of Planning and Environment web page is used for collecting daily PM\(_{2.5}\) in Wagga Wagga for the mentioned period. It is important to emphasize that air pollution is influenced by various factors, including traffic, not solely limited to wildfires. Therefore, the correlation between weather and fuel variables with PM\(_{2.5}\) is not exclusive to wildfire smoke but encompasses a broader context. The objective of developing the ML model for the first research question is not centered solely on achieving accurate predictions. Instead, it aims to discern the influence of each variable and identify the most crucial one, with the ultimate goal of facilitating practical integration into decision-making and management processes in the future.

There are some missing points (several days) for PM\(_{2.5}\), and they are estimated based on linear interpolation which is popular for environmental studies [34,35]. To address the second research question, a distinct dataset sourced from the Global Health Data Exchange (GBD) is employed, featuring a different time frame and encompassing diverse
locations, including multiple countries, rather than being limited to Wagga Wagga. GBD is an extensive database, managed by the Institute for Health Metrics and Evaluation, that compiles health-related data from a broad spectrum of sources. It includes information from 1990 to 2019 for 23 age groups on more than 350 human health conditions, gathered from 204 different countries and territories. The data are sourced from surveys, surveillance records, hospital admission and outpatient data, health insurance claims, and scientific literature studies, drawing from an extensive pool of nearly 69,000 reliable epidemiological sources [36]. Smoke health impacts, particularly in relation to COPD and lung cancer, are evaluated in this study. COPD and TB&L impacts due to ambient and household air pollution are considered in this study. Due to the long-term effects of smoke exposure and the practical challenges in collecting localized data, this study utilizes global health data encompassing records from various regions affected by smoke-related deaths. The study aims to demonstrate the harmfulness of smoke exposure from both ambient and household sources, and thus both were considered in the analysis.

2.2. Data-Driven Models

The general and well-known machine learning models employed in environmental modeling, including Support Vector Regression, Multi-layer Perceptron, and tree-based ensemble algorithms, were selected for this study to compare their performance. These models were chosen based on their established effectiveness in handling complex relationships and predicting environmental phenomena, ensuring robust analysis and interpretation of wildfire-related data. Support Vector Regression (SVR) is a data-driven model that aims to discover a curve that minimizes the difference between the predicted and actual values. SVR models have a maximum acceptable error for prediction, denoted as Epsilon (\(\varepsilon\)), which is determined during the hyper-tuning phase. The predicted value should fall within the support vector range of \([-\varepsilon, +\varepsilon]\), ensuring that the residuals remain smaller than \(\varepsilon\). The SVR method can be expressed as [37]:

\[
    f(x) = (w, \phi(x)) + b
\]

where \(f(x)\) is the output value that should fall within the support vector for all modelled data; \(\phi(x)\) = Kernel function; \(w\) = weight vector; \(b\) = bias applied to the function estimation. During the training phase, the weight vector is fine-tuned to achieve the lowest possible error, specifically the root mean square error, in the prediction. The kernel function plays a crucial role by transforming the input data into meaningful output. Multi-layer Perceptron (MLP) is the other powerful artificial neural network model that can be utilized to model environmental issues [38]. In order to reach the best neurons’ weight, an optimizer from the quasi-Newton methods is utilized. The equation of MLP is as follows:

\[
    y = f\left(\sum_{j=1}^{m} (w_{ij}x_j) + b_i\right)
\]

where \(y\) = the dependent variable; \(i\) = considered data point; \(f\) = activation function; \(m\) = the number of independent values; \(w_{ij}\) = weight linked to the \(j\)th independent values; \(x_j\) represents the \(j\)th of independent values; \(b_i\) = bias or intercept put on. In this study, a non-linear activation function is utilized (ReLU) as follows [39]:

\[
    f(x) = \max(0, x)
\]

Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Natural Gradient Boosting (NGBoost) are ensemble learning techniques that use the decision tree model as the base [40]. The choice of ensemble methods in this study is based on their superior performance in addressing environmental issues. The idea of using several models together as ensemble models can prevent overfitting, and past studies have demonstrated their higher performance in comparison with single models [41–43]. As a simple ML model,
decision trees are able to show how the input variables relate to the response variable [44]. In each iteration, trees strive to identify the ideal division of the parent node into two child nodes. This division, known as the optimum split, is determined based on minimizing the error, such as Mean Squared Error (MSE). The objective is to achieve a reduction in MSE through the process of finding the optimal split as follows:

\[
\Delta i (s, t) = i (t) - \frac{N_{tL}}{N_t} i (t_L) - \frac{N_{tR}}{N_t} i (t_R)
\]

where \( \Delta i (s, t) = \) the reduction in MSE; \( N_t, N_{tL}, N_{tR} \) is the sample size of the parent node and the left (L) and right (R) child nodes, respectively; and \( i (t) \) is the MSE. RF is a popular ensemble ML modelling approach that has been used to model wildfire-related problems [45]. Random Forest (RF) constructs numerous individual decision tree models by employing bootstrap aggregating (bagging) with different subsets of the training data [46]. RF utilizes these bootstrapped samples, which contain randomly selected instances from the training dataset, to make predictions. For instance, if the training dataset is divided into five subsets, five separate decision trees are built solely using the records within each subset. The training process for these five decision tree models occurs using their respective subset of data instead of the entire training dataset. When making predictions for unseen scenarios, the outputs are determined by combining (averaging) the results from the various decision trees that were developed as follows:

\[
y_{pred} = \frac{1}{n_{tree}} \sum_{i=1}^{n_{tree}} y_i
\]

where \( y_{pred} \) is the prediction of the model; \( n_{tree} \) is the number of generated decision trees based on bootstrap samples; \( y_i \) is the prediction of the \( i \)th tree based on \( i \)th bootstrap samples.

XGBoost is one of the most popular ensemble models for environmental modelling [47,48]. The XGBoost model tries to minimize the below loss function in each iteration:

\[
l (t) = \sum_{i=1}^{n} l (y_i, \hat{y}_i^{(t-1)} + f_t (x_i)) + \Omega(f_t)
\]

where \( l \) is the loss function or error quantification; \( t \) is the number of iterations; \( \Omega \) is the penalty for complicated models; \( y_i \) is the output of the model; \( \sum_{i=1}^{n} l (y_i, \hat{y}_i^{(t-1)} + f_t (x_i)) \) is associated with the summation of current and previous loss. By optimizing the loss function, the most accurate XGBoost model can be developed based on the input dataset. NGBoost is another ML model that is popular for modelling smoke impact on the community [49]. NGBoost models have three important sections: Base learner, Probability Distribution, and Scoring rule. First, the base learner model can be different models including decision tree, SVR, RF, and MLP. Second, the probability distribution is related to the output types. For continuous output, a probability distribution can be a normal distribution. Third, the scoring rule can be expressed as the following equation:

\[
L(\theta, y) = -\log P_\theta (y)
\]

where \( L \) is the scoring rule, \( P \) is the probability distribution, \( y \) is the output of the model, and \( \theta \) is the parameter of the scoring rule. The scoring rule can assign a score to each output using a probability distribution. Hyperparameter tuning of AI models is conducted by using Grid search as a popular approach for setting hyperparameters [50]. In addition, 5-fold cross-validation is employed for tuning parameters as its usefulness for AI models [51]. Grid search is employed to evaluate the data-driven model’s performance using each combination of predefined hyperparameters and identify the best hyperparameters. As an example for hyperparameter tuning for predicting PM\(_{2.5}\), we tested various kernel types, including radial basis function (RBF) kernel, linear kernel, and polynomial kernel with a degree of 3. The polynomial kernel function was selected. Additionally, we optimized
the regularization parameter, choosing 10 from the range [0.1, 1, 10]. The strength of the regularization is inversely proportional to C. We also selected an epsilon value of 0.3 from the range [0.01, 0.1, 0.3, 0.5]. The epsilon parameter specifies the epsilon-tube within which no penalty is associated in the training loss function for points predicted within a distance epsilon from the actual value.

2.3. Removing Outliers

Data-driven models predict new situations based on previous experience. The quality of their prediction is densely dependent on previous experience (available data records). If there are a few outliers in the input dataset, the prediction performance will reduce. Consequently, in this study, outliers are detected by using Cook distance which is popular for wildfire research studies [52]. Cook distance is a method that can identify the influence of each data record on the regression model. Outlier data records that have a strong impact on the regression model can be detected easily by Cook distance as follows [53]:

\[ D_i = \frac{\sum_{j=1}^{n} \left( \hat{y}_j - \hat{y}_{j(i)} \right)^2}{p \ \text{MSE}} \]  

(9)

where \( \hat{y}_j \) is the jth fitted response value; \( \hat{y}_{j(i)} \) is the jth fitted response value, where the fit does not include observation i; MSE is the Mean Squared Error; p is the number of coefficients in the regression model. In this study, Cook’s distance was calculated using the NumPy library in Python. The missing data have been interpolated using linear interpolation of other data records. In this study, the actual values have been used without normalization.

2.4. Model Evaluation and Interpretation

The assessment of model performance involves calculating various goodness-of-fit statistics when applying the models to both the training and evaluation datasets. The selected goodness-of-fit metrics included Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Mean Bias Error (MBE). The evaluation dataset was previously unseen by any of the models and served as an independent measure to evaluate and compare the effectiveness of different methods employed in terms of their fit. RMSE, MAE, MAPE, and MBE formulations are mentioned as follows (\( y_p \) is the predicted value and \( y_o \) is the observed or true value):

\[ \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left( y_p - y_o \right)^2} \]  

(10)

\[ \text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| y_p - y_o \right| \]  

(11)

\[ \text{MBE} = \frac{1}{N} \sum_{i=1}^{N} (y_p - y_o) \]  

(12)

\[ \text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left( \frac{y_o - y_p}{y_o} \right) \times 100 \]  

(13)

Feature importance of tree-based models can be expressed as Gini Importance [54]; assuming only two child nodes (binary tree), the Gini Importance can be calculated as follows:

\[ ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \]  

(14)

where \( ni_j \) is the importance of node j; w is the weighted number of samples reaching node j; C is the impurity value of node j; left (j) is the child node from left split on node j; and
right (j) is the child node from right split on node j. Dividing the summation of all Gini Importance related to nodes that split one feature by the summation of all Gini Importance can be used to reach the importance of each feature.

Shapely Additive Explanations (SHAP) provide a means to comprehend the sensitivity and influence of input variables in machine learning (ML) models [31]. SHAP values provide a unified measure of feature importance by attributing each feature’s contribution to the prediction. This research utilizes the mathematical formulation of SHAP proposed by Lundberg, et al. [55]. SHAP values offer the capability to showcase the contribution of each input variable and identify their importance and whether they have a positive or negative impact on the output. The estimation of SHAP values is commonly carried out through various approximation methods, such as kernel SHAP, deep SHAP, and tree SHAP. For example, when a feature with a red value has a SHAP value of 10, it means that high values of this input can increase the outputs by 10. The choice of approximation method depends on the specific machine-learning technique employed. The SHAP formulation is as follows:

$$SHAP(X_j) = \sum_{S \subseteq N(j)} \frac{k!(p-k-1)!}{p!} (f(S \cup \{j\}) - f(S))$$

(15)

where p = the total number of features; N(j) = a set of all possible combinations of features excluding Xj; S = a feature set in N\(\setminus\{j\}\); f(S) = the model prediction with features in S; f(S \cup \{j\}) = the model estimation with features in S plus feature Xj; k = 1 for the features used in the prediction of the data; 0 for features not used.

3. Results

Figure 1 visualizes the relationship among various input and output variables in the dataset to predict the daily mean of PM$_{2.5}$. The data records are related to one year from 1 June 2019 to 31 May 2020, which is the major fire season in Australia [8]. The data records of each season are demonstrated with one color to show season impacts on weather and fuel variables. Brighter data records are related to summer and spring. In Figure 1, darker points are related to fall and winter. If the variables for the vertical and horizontal axis are the same, the probability density functions (PDFs) of the variable are depicted. By increasing the temperature and solar radiation, the fuel moisture content decreases. According to the variables’ PDF for various seasons, it is evident that the majority of PM$_{2.5}$ levels are associated with typical conditions, which fall below the daily mean threshold of 12.5 µg/m$^3$. These levels are considered indicative of good air quality, as defined by the Environmental Protection Authority Victoria, Victoria State Government of Victoria. However, there are a few exceptions observed during days with significant wildfires, where PM$_{2.5}$ levels exceed the normal range. The exceptions where PM$_{2.5}$ levels exceeded the normal range due to significant wildfires were on 20 December 2019, 22 December 2019, 23 December 2019, 2 January 2020, and 5 January 2020, with PM$_{2.5}$ levels exceeding 100 µg/m$^3$ on these dates. In addition, winter points have a narrow distribution with high fuel moisture (M$_{dl}$), low temperature (T), and solar radiation (S). The M$_{dl}$/solar radiation plot clusters data records based on the season, where summer and winter data records are on the two opposite sides. Summer data records demonstrate lower fuel moisture and higher solar radiation than winter, while fall and spring records exhibit intermediate values. For average fuel moisture, the values are as follows: winter has an average fuel moisture of 14.3, while summer has an average fuel moisture of 7.7. According to rain (R) plots, the data reveal that the majority of days had rainfall below 10 mm, with only a few exceptional cases exceeding 20 mm in spring and fall. The average temperature for winter is 14.13 °C, whereas the average temperature for summer is 33.20 °C.
Figure 1. Correlation among weather variables, fuel variables, and wildfire smoke from June 2019 to May 2020. PM$_{2.5}$ is the daily mean of fine particulate matter with a diameter ≤2.5 µg/m$^3$; M$_{dl}$ is fuel moisture content in percentage; T is the ambient temperature that is reported by the weather station in Celsius; S is solar radiation that is reported by the weather station in megajoule per square meter; R is rain that is reported by the weather station in millimeters.

Table 1a compares the performance of data-driven models for predicting PM$_{2.5}$ based on weather and fuel variables. The input variables for the models are fuel moisture content, ambient temperature, ambient relative humidity, solar radiation, rain, and wind speed. It is worth noting that outlier data records are identified by using Cook’s distance [56]. Outliers’ removal has a positive impact on the prediction performance of data-driven models. Based on Table 1a, ensemble tree-based models, especially Random Forest and NGBoost models (see Methods), can predict PM$_{2.5}$ better than other models. Since the square of the disparity between predicted and observed values yields a greater impact from anomalies, RMSE emphasizes the significance of outliers more than MAE. During the study period, the maximum PM$_{2.5}$ value recorded was 560, which is significantly higher than the average of 14. If this extreme value was used in the model without removal, the model might attempt to reduce the error for this outlier, potentially ignoring trends in the data around the average. This could lead to a reduction in the model’s accuracy. Extreme values can
also impact statistical power, making it challenging to detect a true effect if one exists. Thus, based on MAE, NGBoost is selected as the most accurate model for predicting PM$_{2.5}$ based on weather and fuel variables. NGBoost RMSE, MAE, MAPE, and MBE are 23.4, 8.0, 66%, and $-1.5$, respectively.

Table 1. The performance of data-driven models for predicting (a) PM$_{2.5}$ based on weather and fuel variables (air temperature, relative humidity, solar radiation, rain, wind speed, season, and fuel moisture); (b) the number of deaths related to COPD; (c) the number of deaths associated with TB&L—the input variables for prediction of the number of deaths are sex, age, ambient and household pollution, and year—the prediction performance is evaluated against the test dataset. The test dataset contains 20% of the total data records that were randomly selected.

<table>
<thead>
<tr>
<th></th>
<th>(a) PM$_{2.5}$</th>
<th>(b) COPD</th>
<th>(c) TB&amp;L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>MAE</td>
<td>MAPE</td>
</tr>
<tr>
<td>Support Vector Regression (SVR)</td>
<td>28.3</td>
<td>8.5</td>
<td>54%</td>
</tr>
<tr>
<td>Multi-layer Perceptron (MLP)</td>
<td>23.7</td>
<td>8.6</td>
<td>84%</td>
</tr>
<tr>
<td>Random Forest (RF)</td>
<td>23.2</td>
<td>8.1</td>
<td>71%</td>
</tr>
<tr>
<td>Extreme Gradient Boosting (XGBoost)</td>
<td>30.1</td>
<td>10.6</td>
<td>81%</td>
</tr>
<tr>
<td>Natural Gradient Boosting (NGBoost)</td>
<td>23.4</td>
<td>8.0</td>
<td>66%</td>
</tr>
</tbody>
</table>

Given the reasonable performance of the NGBoost model in predicting PM$_{2.5}$, it could be employed to address two significant inquiries. First, the model’s ability to predict PM$_{2.5}$ in various seasons is analyzed. The results can demonstrate NGBoost to be a season-invariant predictor for PM$_{2.5}$. Figure 2a–d are associated with the prediction performance of the model for spring, summer, fall, and winter data records. As Figure 2b illustrates, there are 62 data records with PM$_{2.5}$ levels exceeding 12.5 µg/m$^3$ in summer. In addition, Figure 2 shows data records that are used in the training and tuning process in red (80% of total data records), and data records that are used in the test process in blue. Figure 2e interprets the NGBoost model and shows solar radiation (feature importance = 0.35), fuel moisture (feature importance = 0.29), and ambient temperature (feature importance = 0.21) are the three most influential variables on PM$_{2.5}$. The importance of features S, M, and T exceeds twice that of the fourth variable, which is wind speed (with a feature importance of 0.08). The significance of seasons (spring, summer, fall, and winter) in predicting PM$_{2.5}$ is below 0.01, indicating that our predictor is season-invariant.

To evaluate the applications of data-driven models for consequence evaluation of smoke exposure impacts on health, Table 1b,c compare the prediction performance of models. The data-driven models in Table 1b,c use sex, age, pollution type (ambient and household), and year on the impact as input to predict the number of deaths related to COPD and TB&L. The number of deaths associated with COPD and TB&L can be a good indicator of the health impact of smoke exposure. Table 1 illustrates that the NGBoost model has reasonable prediction performance for COPD as well as TB&L, so it is selected for the next step (Figure 3). NGBoost demonstrates the ability to predict the quantity of COPD-related deaths with a Mean Absolute Error (MAE) of 5.6, as well as the quantity of TB&L-related deaths with an MAE of 6.4. NGBoost shows the best performance among
these models based on the results, as anticipated. This model leverages other models as base learners, thereby enhancing prediction performance. The selection of the NGBoost model was based on its performance metrics, particularly Mean Absolute Error (MAE), which is more suitable for models with outliers compared to Mean Squared Error (MSE). From Table 1, it shows that NGBoost has better MAE for two out of three problems, which justifies its selection for the next step as shown in Figure 3.

![Figure 2](image_url)

**Figure 2.** The prediction performance of the NGBoost model. The selected data-driven model for predicting PM$_{2.5}$ uses the weather and fuel variables (air temperature, relative humidity, solar radiation, rain, season, wind speed, and fuel moisture) as input variables. Plots of the observed PM$_{2.5}$ versus the predicted PM$_{2.5}$ for (a) spring (b) summer (c) fall (d) winter. The solid black line is the line of perfect agreement. The dashed lines indicate the ±35% error interval. (e) shows the importance of each input variable in the model development.
TB&L-related deaths with an MAE of 6.4. NGBoost shows the best performance among these models based on the results, as anticipated. This model leverages other models as base learners, thereby enhancing prediction performance. The selection of the NGBoost model was based on its performance metrics, particularly Mean Absolute Error (MAE), which is more suitable for models with outliers compared to Mean Squared Error (MSE). From Table 1, it shows that NGBoost has better MAE for two out of three problems, which justifies its selection for the next step as shown in Figure 3.

Figure 3. Prediction performance of the NGBoost model. The model output is the number of deaths associated with COPD and TB&L by using input variables (sex, age, household and ambient air pollution, and year of impact). Plots of the observed versus the predicted number of deaths associated with COPD (a) and TB&L (b). The dashed lines indicate the ±35% error interval. Red points are the training dataset (80% of data records), and blue points are the test dataset (20% of data records). SHAP interpretation of the selected model to find the influential variables for COPD-related deaths (c) and TB&L-related deaths (d).

Figure 3a,b show the observed versus predicted plots of NGBoost models for the number of COPD-related deaths and TB&L-related deaths, respectively. The test datasets (20% of data records) are shown as red data points, representing training and tuning data records. Figure 3a,b demonstrate that the selected model can predict most data records within an acceptable error range of less than 35%. Figure 3c,d interpret the nominated model to identify the most influential variables affecting COPD and TB&L deaths, respectively.
For COPD-related deaths, as shown in Figure 3c, ambient air pollution has the greatest impact, followed by household air pollution, both of which have a positive impact of fewer than 20 deaths. For TB&L-related deaths, as shown in Figure 3d, household air pollution has an impact of about 20 deaths. The placement of red data records for ambient and household air pollution on the positive side of the SHAP (Shapley Additive Explanations) figure indicates their positive impact on the number of deaths. Furthermore, the impact of the year is found to be minimal, as it exhibits the lowest Shapley means. However, a few red data points are observed on the right side, indicating a marginally positive effect on the number of deaths in recent years. In terms of TB&L deaths (Figure 3d), age ranks second after household air pollution, suggesting that age is a more critical factor than ambient air pollution. However, the red color (older age) predominantly indicates the majority of age-related data points situated on the right side, exhibiting positive SHAP values. These observations indicate that older individuals exhibit a higher vulnerability to TB&L compared to their younger counterparts. Similarly to COPD-related deaths, the year holds the least influence on TB&L-related deaths as well. The importance of PM$_{2.5}$ as a key indicator of air pollution and its significant impact on health outcomes, including mortality related to respiratory issues, is illustrated in Figure 3. Changes in PM$_{2.5}$ during this year have been demonstrated in Figure 4.

![Figure 4. Changes in PM$_{2.5}$ µg/m$^3$.](image)

4. Discussion

This study aims to offer a comprehensive understanding of the exposure to wildfire smoke, taking into account the influence of weather and fuel variables such as temperature, rainfall, wind speed, and fuel moisture. Several data-driven models are evaluated, and it is found that the ensemble tree-based models achieve the highest level of accuracy in predicting the test dataset. This finding corroborates previous research that has assessed the predictive performance of machine learning models in similar contexts [41–43]. The best practice in previous studies involves evaluating the importance of features to elucidate data-driven models [57]. Our analysis of feature importance reveals that solar radiation, temperature, and fuel moisture are the primary predictors of PM$_{2.5}$. Notably, these observations align with previous studies [58], strengthening the validity and applicability of the findings. For example, solar radiation is identified as the most influential variable, so on days with fires, practitioners should anticipate higher PM$_{2.5}$ levels. Consequently, they can implement policies to address this anticipated situation. This paper also explores the season-invariant capability of the chosen data-driven model. The findings demonstrate
that the boosting models exhibit season-invariant predictive performance for PM$_{2.5}$. While the PM$_{2.5}$ concentration during summer averages 44 µg/m$^3$, significantly higher than the winter average of 9 µg/m$^3$, the plots for PM$_{2.5}$ prediction in other seasons show a relatively similar pattern. This observation highlights the generalization of the selected model in capturing and predicting PM$_{2.5}$ variations across different seasons.

In addition, this study evaluates the impact of smoke exposure on public health outcomes. Specifically, ensemble machine learning models are employed to estimate the number of deaths associated with two significant diseases of COPD and TB&L. It is essential to focus on the interpretability of ensemble data-driven models [59]. The utilization of SHAP for interpreting our models shows that ambient air pollution is the most influential variable for COPD deaths, whereas it ranks as the third most important variable for TB&L deaths. The utilization of SHAP for interpreting our models reveals that ambient air pollution is the most influential variable for COPD deaths, whereas it ranks third for TB&L deaths. Specifically, an increase in ambient air pollution can correlate with a potential increase of around 20 COPD deaths. This demonstrates how the input impacts the output. As a sanity check, an increase in ambient and household air pollution, as well as age, is associated with a higher number of deaths. In the context of TB&L, age exhibits a higher mean SHAP value compared to COPD. This finding suggests that age plays a more significant role in influencing the number of deaths related to TB&L than it does for COPD. This observation highlights the importance of considering age as a crucial factor when assessing and addressing the impact of smoke exposure on health outcomes, specifically in relation to TB&L [60]. The impact of the year on the number of deaths is relatively insignificant; however, recent years demonstrate a growing importance (as indicated by red points with positive SHAP values). Further research is needed to validate this observation.

5. Conclusions

This study presents a novel framework that establishes a connection between pre-disaster variables (including fuel and weather variables) and post-disaster variables (the health impacts) in the context of smoke exposure (PM$_{2.5}$). By comprehensively considering all these variables and their respective impacts, decision-makers can gain a deeper understanding of the complex interplay between smoke exposure and its consequences. This holistic approach empowers decision-makers to identify optimal solutions that contribute to the achievement of a sustainable world. The integration of pre- and post-disaster variables in this study provides a comprehensive perspective that can enhance our ability to mitigate the adverse impacts of smoke exposure and promote sustainable practices.

By simultaneously considering relationships between pre- and post-disaster variables, the study enables more effective planning and management to mitigate wildfire risks. For instance, the findings provide actionable insights for authorities to implement targeted fire prevention strategies, such as proactive monitoring of fuel moisture content. Moreover, the health impact projections can inform public health preparedness and interventions, enabling timely protection measures for vulnerable groups during wildfire events. Adaptive strategies, such as distributing masks and installing indoor filtration systems, could significantly mitigate smoke exposure risks. Ultimately, this integrated approach offers a holistic view of wildfire impacts across interconnected spheres, supporting the development of cross-sectoral strategies for resilience. It is suggested that future studies also conduct sensitivity analyses and compare the results with SHAP and feature selection methods. This comparison could further enhance our understanding of variable significance and model robustness. Future research could validate and strengthen the modelling with expanded datasets over larger spatiotemporal scales. Overall, linking pre- and post-wildfire factors is an important paradigm to advance scientific understanding and promote sustainability. A limitation of this study pertains to the difference in the time frame and geographical scope between the two research questions. Subsequent research endeavors may consider replicating the methodology for a consistent time frame and region and reporting the findings accordingly. Future research studies should consider including asthma and respiratory
infections, which are more common and affect all age groups, to better understand the health impacts of smoke exposure.

The period for analyzing pre-wildfire parameters on wildfire smoke pertains to the 2019–2020 grassfires. For health data, we consider all records from global health data related to deaths from smoke exposure to highlight the dangerous nature of such exposure, irrespective of the specific time period. The use of data from routine ambient monitoring does not necessarily indicate that the patients were exposed to the PM$_{2.5}$ levels as stated. Therefore, it is suggested that future studies use personal air quality monitoring data to analyze the impact of air pollution on health more accurately. There could be other causes of exacerbations or deaths for the study population other than air pollution that could not be controlled for, given the retrospective nature of the study. We acknowledge that data limitations and generalizability issues could affect the findings. Specifically, the data used may not fully represent all geographic regions or population subgroups, which could limit the generalizability of the results.

Author Contributions: Conceptualization, S.K. and M.A.; methodology, S.K. and M.A.; software, S.K.; validation S.K. and M.A.; formal analysis, S.K.; investigation, S.K. resources S.K. and M.A.; data curation, S.K.; writing—original draft preparation, S.K.; writing—review and editing, all authors; visualization, S.K.; supervision, M.A.; project administration, M.A.; funding acquisition, M.A. All authors have read and agreed to the published version of the manuscript.

Funding: The authors are grateful for support from the Australian Research Council (ARC) through the Linkage Project funding scheme (LP180101080).

Institutional Review Board Statement: The codes were used for quantifying the effect of the curing level on the rate of fire spread using Python language (version 3.9). Dataset visualization is conducted by using the Seaborn Python library [61] which can be found at https://seaborn.pydata.org/index.html, accessed on 18 June 2024. Regression trees, support vector regression (SVR), and Random Forest are developed using the data analysis scikit-learn library [62] which can be found at https://scikit-learn.org/stable/, accessed on 18 June 2024. Extreme Gradient Boosting (XGBoost) is developed by using the XGBoost Python package which can be found at https://xgboost.readthedocs.io/en/stable/, accessed on 18 June 2024. Natural Gradient Boosting (NGBoost) is developed by using the NGBoost Python library which can be found at https://github.com/stanfordmlgroup/ngboost, accessed on 18 June 2024. SHAP (SHapley Additive exPlanations) is utilized for interpretation by its Python package which can be found at https://shap.readthedocs.io/en/latest/, accessed on 18 June 2024.

Informed Consent Statement: Not applicable.

Data Availability Statement: The analyzed data records in this study are publicly available and can be categorized into three classes. The first class is related to weather variables from 1 June 2019 to 31 May 2020 collected from Climate Data Online (Bureau of Meteorology website. The second class is associated with air pollution variables (PM$_{2.5}$) obtained from the Data download facility of the NSW Department of Planning and Environment which can be found at https://www.dpie.nsw.gov.au/, accessed on 18 June 2024. The third class is devoted to the health impact of air pollution gained from the Global Health Data Exchange query tool that can be found at https://vizhub.healthdata.org/gbd-results/, accessed on 18 June 2024.

Acknowledgments: The authors are grateful for support from the ASCII Lab members at Monash University and their constructive feedback on progressive iterations of this work.

Conflicts of Interest: The authors declare no conflict of interest.

References
2. Rodrigues, M.; Cunill Camprubi, À.; Balaguer-Romano, R.; Coco Megía, C.J.; Castañares, F.; Ruffault, J.; Fernandes, P.M.; Resco de Dios, V. Drivers and implications of the extreme 2022 wildfire season in Southwest Europe. Sci. Total Environ. 2023, 859, 160320. [CrossRef] [PubMed]
3. Cong, J.; Gao, C.; Han, D.; Li, Y.; Wang, G. Stability of the permafrost peatlands carbon pool under climate change and wildfires during the last 150 years in the northern Great Khingan Mountains, China. Sci. Total Environ. 2020, 712, 136476. [CrossRef] [PubMed]


10. He, J.; Huang, C.-H.; Yuan, N.; Austin, E.; Seto, E.; Novosselov, I. Network of low-cost air quality sensors for monitoring indoor, outdoor, and personal PM2.5 exposure in Seattle during the 2020 wildfire season. Atmos. Environ. 2022, 285, 119244. [CrossRef]

11. Schweizer, D.; Preisler, H.; Entwistle, M.; Gharibi, H.; Cisneros, R. Using a Statistical Model to Estimate the Effect of Wildland Fire Smoke on Ground Level PM2.5 and Asthma in California, USA. Fire 2023, 6, 159. [CrossRef]


14. Ghali, R.; Akhloufi, M.A. Deep learning approaches for wildland fires using satellite remote sensing data: Detection, mapping, and prediction. Fire 2023, 6, 192. [CrossRef]


21. Coker, E.S.; Buralli, R.; Manrique, A.F.; Kanai, C.M.; Amegah, A.K.; Gouveia, N.; Association between PM2.5 and respiratory hospitalization in Rio Branco, Brazil: Demonstrating the potential of low-cost air quality sensor for epidemiologic research. Environ. Res. 2022, 214, 113738. [CrossRef]


28. Akdemir, E.A.; Battye, W.H.; Myers, C.B.; Aneja, V.P. Estimating NH 3 and PM 2.5 emissions from the Australia mega wildfires and the impact of plume transport on air quality in New Zealand. Environ. Sci. Atmos. 2022, 2, 634–646. [CrossRef]


44. Khanmohammadi, S.; Cruz, M.G.; Mohammadi Golafshani, E.; Bai, Y.; Arashpour, M. Application of artificial intelligence methods to model the effect of grass curing level on spread rate of fires. *Environ. Model. Softw.* 2024, 173, 105930. [CrossRef]

45. Li, Y.; Li, G.; Wang, K.; Wang, Z.; Chen, Y. Forest Fire Risk Prediction Based on Stacking Ensemble Learning for Yunnan Province of China. *Fire* 2023, 7, 13. [CrossRef]


Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.