Wildfire Air Quality Prediction:
A Data-Driven Approach

Subhankar Dhar, San Jose State University, USA*

 https://orcid.org/0000-0003-1493-451X

Jerry Gao, San Jose State University, USA

ABSTRACT

Wildfires are extremely harmful to the environment. While producing gaseous pollutants and particles that cause smoke, wildfires also release carbon dioxide (CO_2), a greenhouse gas that will continue to warm the planet after the wildfire ends. This article delves into the impact of wildfires and air quality on human living conditions. The authors' machine learning models use wildfire data to forecast air quality with detailed indexes and geographic information during a wildfire. The work evaluates the performance of each machine learning model via statistical metrics like mean absolute error (MAE), R-squared (R2), and root mean squared error (RMSE). The experimental results used neural networks to predict a specific value for carbon monoxide (CO), ozone, and PM2.5. These are both promising and accurate, providing meaningful insight into air quality within a region. This work will be useful for cities, governments, citizens, and public safety.

INTRODUCTION

Wildfires can increase air pollution and cause severe damage to air quality by emitting carbon dioxide (CO2), carbon monoxide (CO), and other greenhouse gases that contribute to global warming and environmental hazards (Castelli et al., 2020). In addition, they can damage forests that otherwise remove CO2 from the air and inject aerosols into the atmosphere. Smoke from wildfires can cause serious health disorders and respiratory diseases (Camia & Amatulli, 2009; de Groot et al., 2007) by reducing the size of green forests and eliminating CO2 in the air (Reid et al., 2019). Pollutants like particulate matter (PM) 2.5, ozone (O3), nitrogen dioxide (NO2), and CO are major parameters in the air quality index (AQI) of a region (Tian et al., 2011).

*Corresponding Author

This work delves into the impact of wildfires on air quality, using data science tools and technologies to predict the impact of air quality after a wildfire event. To achieve this goal, the authors must define key measurements with respect to air. For example, air has several important measurement metrics that affect air quality, including the CO level, ozone level, and PM2.5 (small inhalable particles with diameters no more than 2.5 micrometers) index. These pollutants can have severe negative impacts on human health, including wheezing, coughing, sore eyes, and throats. The authors, therefore, develop machine learning models to understand the health effects caused by wildfires and strengthen public awareness surrounding such events.

First, the authors develop a data-driven machine learning framework to predict air quality during and after wildfires. In addition, the study addresses the limitations of existing methods. The research provides insight into the significant impact of wildfires, enabling people and governments to prioritize environmental protections and prevention efforts. Residents can also monitor air quality changes more closely and prepare for potential wildfires.

Second, the authors use two sets of machine learning models to predict air quality during wildfire seasons. The study provides detailed explanations on how to determine and tune parameters. Analysis helps the public understand and anticipate air quality changes. It also enables individuals to compare and improve upon existing models.

Third, the study includes results driven by statistical analysis and machine learning. It aims to prove the significance of wildfires and predict air quality with improved accuracy. Descriptive and inferential statistical analysis tools, along with visualization diagrams and graphs, depict the difference between air quality before and after wildfires. The study's regression and deep learning models generate precise metrics to increase accuracy.

Fourth, the study presents a data-driven approach for deploying an online wildfire air quality alert system. It uses real-life data to develop a proof-of-concept implementation. This prototype demonstrates the novel functionalities of the authors' framework, benefiting both the public and city officials.

The article is organized as follows. The next section discusses related works. Then, the article describes the geographical location of the study, comprehensive data parameters, data pre-processing, and training and test data preparation. Wildfire air quality prediction models and experimental results are explored before navigating the system design, implementation framework, and visualization tools. Lastly, the study discusses the conclusion and future work.

RELATED WORK

Preisler et al. (2015) highlighted the shortcomings of research when predicting wildfire impacts from PM2.5 concentration at ground-level monitors in California. While most researchers rely on satellite-based observational tools, this work combined models with an autoregressive statistical model, incorporating weather and seasonal factors to identify thresholds for predicting unusual events. The study focused on ground-based monitoring of PM2.5 levels, with data consisting of hourly values of PM2.5 and meteorological data. Data was gathered from the United States Department of Agriculture Forest Services. Unexpectedly, the study found that smoke plumes could identify seasonal wildfire influence with high accuracy.

Reid et al. (2015) and Jaffe et al. (2013) evaluated the contribution of O3 caused by wildfires in the atmosphere in western U.S. The studies developed several statistical models to estimate the maximum daily eight-hour average (MDA8) O3 emissions and created methodology functions for three western areas in the U.S. The residual of the statistical model can provide information on O3 emissions that cannot be explained by normal wildfires.

Similarly, Preisler and Westerling (2007) developed a statistical model for forecasting fire-danger and producing one-month ahead wildfire-danger probability in the western U.S. The model predicted a monthly average temperature and drought severity index to demonstrate significant potential

wildfire areas over historical data. Each of these studies focused on a different target variable and collected input features under different circumstances, locations, and methods. It was, therefore, a challenge to apply the same methods to different datasets (Jaffe et al., 2013; Preisler & Westerling, 2007; Preisler et al., 2015).

Guan et al. (2020) highlighted the impact of PM2.5 caused by a large wildfire in the southeastern U.S. in November 2016. The study investigated the wildfire via the community multiscale air quality model and weather research and forecasting model. It revealed the effect of the aerosol on worsening air quality after a wildfire.

Reid et al. (2019) used machine learning to create daily surface concentration maps for PM2.5 and O3 during an intense wildfire in California in 2008. The study linked daily exposures to the counts of respiratory hospitalizations and emergency department visits at the zip code level. Their work found that PM2.5 plays a significant role when considering the association between wildfires and respiratory hospitalizations. However, the associations with ozone were confounded.

Yao et al. (2018) studied British Columbia's wildfire season from 2010 to 2015 via five data sources: (1) PM2.5 measurements; (2) fire activity; (3) ecozone; (4) meteorology; and (5) elevation. The authors acquired one-hour average PM2.5 measurements from more than 70 air quality monitoring stations. They calculated the average elevation of each grid cell in the study area with data from the GTOPO30 product.

Watson et al. (2019) aimed to use machine learning techniques to predict ozone emission during wildfires in California. They tested several machine learning models for predicting ground-level O3 during wildfires. Their work assessed model accuracy using MDA8, the same index used in the current O3 model. The study also used leave-one-location-out cross-validation (LOGO CV) as the evaluation metric. They found that among the 10 machine learning models, gradient boosting had the highest accuracy and lowest LOGO CV estimated RMSE. Random forest (RF) was the second-best predictor. They also noted that the ranking of predictors may vary depending on the evaluation metric, such as 10-fold CV. They determined that differences in the evaluation results will be more significant when using flexible models like gradient boosting or RF.

The most important takeaway from the current study is that ensemble tree models have highly flexible mean structures when considering the output as the mean value. This is because the mean structure characterizing the relationship between covariates and O3 is likely to include interactions, nonlinearities, and possible discontinuities.

REGION OF STUDY AND DATASETS

To fulfill the objectives and expectations of the current research, the authors included data from three components: (1) wildfire; (2) weather; and (3) air quality. This resulted in three large datasets. The study focused on assessing the impact of wildfires on counties in northern California. Therefore, the authors carefully selected datasets with high-quality data for the relevant regions. They prioritized datasets provided by local governments (considered reliable sources). They also disregarded datasets with excessive null values, difficulty in accessing data at the desired scale, outdated information, or delayed updates (Huang et al., 2021). Table 1 provides the data types and their sources.

Data Preprocessing

This section addresses the data processing involved in preparing the dataset for training a model that predicts air quality. The dataset comprises historical wildfire, weather, and air quality data. It is collected from various sources. Data processing, a critical aspect of this research, can be challenging. To accomplish the authors' research goals, machine learning tasks were used to identify input and output datasets. The authors aimed to predict the impact of northern California wildfires on air quality. Thus, their output dataset consisted of air quality, fire, and weather data. Both weather conditions

Table 1. Data types and sources

Data	Data Type	Source	Time Range
Wildfire	SQL Lite	Kaggle Wildfire dataset* This publication provides a spatial database of wildfires that took place in the U.S. from 1992 to 2015. The database was originally developed to support the national Fire Program Analysis (FPA) system. It is the third update to the publication. The data on wildfires were obtained from the reporting systems of federal, state, and local fire organizations.	1992-2015
Weather	CSV/txt file	Statewide Integrated Pest Management Program: Current and past weather data for approximately 400 weather stations throughout California**.	1951-present
Air Quality	CSV	U.S. Air Protection Agency (EPA) []	1990-present

*https://www.kaggle.com/datasets/rtatman/188-million-us-wildfires
**https://ipm.ucanr.edu/WEATHER/wxactstnames.html

and wildfire factors significantly impact air quality; therefore, fire and weather data served as input features and air quality data served as output features.

In the feature engineering process, the aim was to reduce the dataset dimensionality and eliminate collinearity through variable selection. According to a report by Dominick et al. (2012), air pollutant concentration (e.g., ground-level O3, PM2.5, and NO2) is affected by meteorological factors and local topography. Meteorological conditions like air quality are dependent on processes

Table 2. Datasets and important features

Dataset	Important Features
Fire	• Fire_code • Fire_name • Fire_year • Discovery_doy • Discovery_time • Stat_cause_descr • Cont_doy • Cont_time • Fire_size • Fire_size_class • Latitude • Longitude
Weather	• Station • Datetime • Precip • Temp_max: • Temp_min:
Air Quality	• Date • Ozone Concentration • CO Concentration • $PM_{2.5}$ Concentration • Daily_AQI_INDEX • COUNTY/COUNTY CODE • Location

like air pollutant emission, transportation, chemical transformations, disposition (wet and dry), and dispersion (Demuzere et al., 2009).

Wildfire Data Processing

After reviewing multiple publicly available wildfire datasets, the authors chose a dataset provided by the Department of Agriculture. It contained a spatial database of wildfires that occurred in the U.S. from 1992 to 2015, totaling 1.88 million geo-referenced wildfires records. The study used the records that occurred in Santa Clara County and Sonoma County in northern California. Records in the wildfire dataset are labeled by county code, simplifying the online search with corresponding county codes.

Weather Data Processing

The weather dataset, with one section for Santa Clara County and one for Sonoma County, included precipitation, temperature, humidity, and other weather-related information. The main tasks analyzed and predicted impacts on air quality in terms of air metrics based on the California wildfire dataset. Given the datasets, the authors chose four machine learning-based data algorithms for each aspect. These were traditional multi-linear regression (MLR), generalized boosting model (GBM), RF, and artificial neural network (ANN) model. The last two models were suggested by previous research papers (Ramona-Gottschling & Gottschling, 2016), in which the researchers compared 11 models and illustrated the favorable performance of the GBM and RF in detail. In particular, the work explored the performance of the two models on predicting PM2.5 concentrations after wildfire events in California.

Air Quality Data Processing

Regarding air data, the authors' dataset used O3, CO, and PM2.5. All five years of daily pollution indexes were collected with labeled dates and geographic location information. The authors needed individual pollutant's air concentration from Santa Clara County and Sonoma County from 2000 through 2015. Data were downloaded from the U.S. Environmental Protection Agency (EPA).

In terms of air quality, the authors analyzed the relationship between wildfires and substances in the air (PM2.5, O3, and CO). This was achieved by comparing the content of each substance before and during the wildfire.

Air quality data contains the following four air pollutant indexes: (1) MDA8 for O3; (2) MDA8 for CO; (3) daily mean for PM2.5 concentration; and (4) AQI. For other columns, some descriptive features were removed, including site ID or site name. Only date and location information were maintained for visualization purposes.

This study used a five-step data analysis approach. Step one found relevant datasets for the analysis. These included outdoor air quality data provided by the EPA. The authors also found wildfire datasets for their study on the National Oceanic and Atmospheric Administration (NOAA) website. It included wildfire location, burning area, wildfire time, and lighting conditions. Step two included the necessary data cleaning process and exploratory data analysis. For instance, the authors could combine those datasets to create a single comprehensive dataset on desired air quality perspectives. Each combined dataset contained all features and aspects within their study. Step three performed feature engineering to reduce the correlation between features and converting categorical features into numerical features using encoding methods. Step four fed data into evaluated performance of different models in terms of multiple statistical metrics.

Data sources and standards are not uniform. Therefore, the authors understood that they needed to apply some form of data scaling or centering. Hence, the study performed the required transformation for later modeling because the data sources were inconsistent. The min-max method was used to set a reasonable range for the normalized data and prevent inconsistency in number size. To achieve min-max normalization, the authors imported the MinMaxScaler from sklearn.preprocessing to construct the scalar model in Python.

PROPOSED MODELS

After merging and cleaning the datasets, the authors had a dataset consisting of six independent variables and three dependent variables. This number was suitable for machine learning tasks. The authors had three dependent variables to predict. They, therefore, needed to run each model three times. Next, the researchers explored major approaches for solving these problems by applying statistical and machine learning models.

Linear regression is widely used in all kinds of machine learning projects. It is also simple enough to be used as a performance baseline (Noorian, 2015; Preisler et al., 2015; Yao et al., 2018). Therefore, the authors chose linear regression as the first model for analysis.

For predicting air quality, the authors used ANN. This is a widely used machine learning model that can solve non-linear problems and predict output values based on input parameters from training data. In this study, the authors implemented an ANN model with three hidden layers and activation functions to calculate PM2.5, CO, and O3 values using input values (Elia et al., n.d.; Jain et al., n.d.; Tonini, n.d.; Zhana et al., 2017). Despite its apparent simplicity, the ANN algorithm is highly effective at replicating relationships within datasets, which makes it an ideal solution for the current air quality prediction study.

Recurrent neural network (RNN) is a type of deep learning algorithm that is useful for time-series data like air quality measurements. Overall, RNN is a powerful tool for air quality prediction. It can capture the complex temporal dynamics of pollution levels and provide accurate forecasts based on past measurements.

This study used an RNN model to predict air quality levels based on past measurements. Unlike other neural networks, RNNs have a memory of previous inputs, which allows them to better model temporal dependencies (Wang & Wang, n.d.; Xhu et al., n.d.; Zhana, 2017). This is important for air quality prediction because pollution levels are affected by factors that change over time, such as weather patterns, traffic, and industrial activity. The authors' RNN model had multiple layers, with each layer using a combination of recurrent and activation functions to generate an output based on the input data from the previous time steps. The authors used the output to predict next step air quality levels based on patterns and trends observed in the historical data.

The last model was the gated recurrent unit (GRU), a type of RNN that addresses the problem of vanishing gradients. It is an improvement over traditional RNNs. GRU is like long short-term memory (LSTM) in design; however, it uses an updated gate and reset gate to selectively transfer information to the output (Watson et al., 2019; Yao et al., 2018). These gates are trained to remember (without deleting) information, even if it is not immediately relevant to the current prediction. Compared to LSTM, GRU is computationally more efficient and simpler to modify. Thus, it was chosen for this study.

EXPERIMENTAL STUDY

This section analyzes the proposed models for predicting air quality during wildfires. First, it describes the steps the authors took to prepare the dataset, including six independent variables and three dependent variables. Then, it explores statistical and machine learning approaches to analyze the problem. A linear regression was used as the baseline model, as it is a commonly used and simple machine learning technique. The authors also deployed ANN with three hidden layers to predict PM2.5, CO, and O3 values. While simple, the ANN algorithm proved to be effective in emulating relationships in data sets. However, after examining the data visualization graphs, the authors realized that the input feature might not be a simple linear relationship. As such, the authors used deep learning models like RNN and GRU to capture the complex relationships between weather and wildfire data.

As noted, RNN internal memory can consider previous calculations and results when making decisions. This proved to be useful for the current project and related data. GRU, an improved version of RNN, solved the vanishing gradient problem and was also computationally efficient. The authors

chose GRU over LSTM models because they are computationally more efficient and simpler to modify (Huang et al., 2021).

Experimental Results

The authors experimented with linear regression, ANN, RNN, and GRU. All the models demonstrated distinct strengths and weaknesses in predicting outcomes (Huang et al., 2021). For each of the four models, the authors used six features as input and selected one AQI as the output each time. To predict three AQIs, the authors ran each model three times, resulting in three sets of test results. Figure 1 displays the six input features and three AQIs.Figure 2 shows the performance of the linear regression model, which performed poorly. All three AQIs received very low scores. This indicated that the linear regression model was not suitable for the authors' problem. It also suggested there may be nonlinear associations among the independent variables and dependent variable.

Figure 1. Input features and three AQIs

```
air_features = ['FIPS_CODE', 'FIRE_SIZE', 'duration', 'mean_precipitation', 'mean_min_temp', 'mean_max_temp']

X = fire_df[air_features]
y = fire_df[['ozone_value', 'CO_value', 'PM2.5_value']]
y
```

	ozone_value	CO_value	PM2.5_value
0	0.040511	0.932919	11.961372
1	0.056826	0.406627	14.274514
2	0.058498	0.452012	11.096789
3	0.057687	0.464976	10.256487
4	0.054478	0.440554	10.371258
...	...	...	...
3584	0.044820	0.379127	10.009098
3585	0.039641	0.287561	7.895305
3586	0.039641	0.287561	7.895305
3587	0.047409	0.263785	9.454478
3588	0.052835	0.287186	11.164340

Figure 2. Linear regression
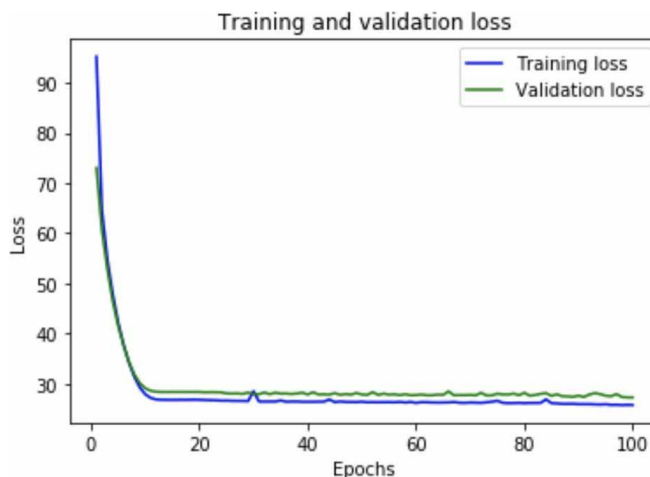
The study constructed an ANN model with two hidden layers. It used the ReLU activation function and calculated the loss through the mean squared error method. Figures 3 through 5 display the training loss and validation loss on O3, PM2.5, and CO values with the ANN model. The ANN model had an overfitting issue on CO value prediction; however, it performed well on O3 and PM2.5 value prediction. Both the training loss and validation loss decreased as the number of training epochs increased. However, this may be because the model was trained for too long. The authors should, therefore, reduce the number of training epochs.
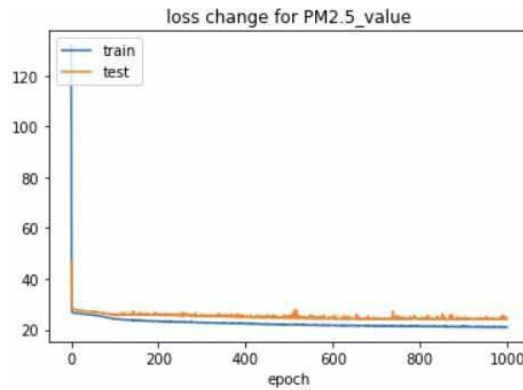
Figure 3. Training Loss vs. validation loss on ozone value prediction with the ANN model
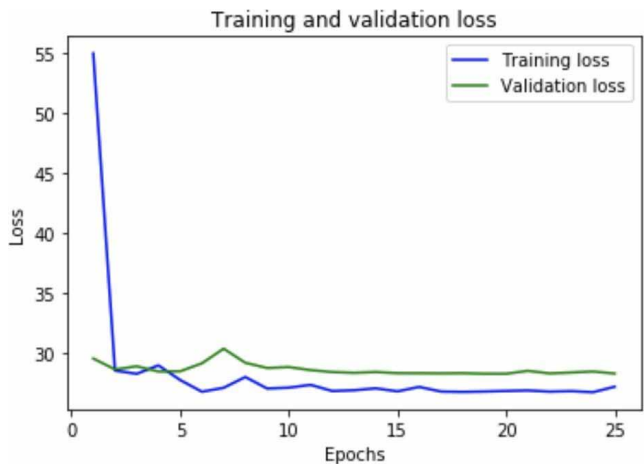
Figure 4. Training loss vs. validation loss on PM2.5 value prediction with the ANN model

```
linear_model(fire_data, 'ozone_value')
linear_model(fire_data, 'PM2.5_value')
linear_model(fire_data, 'CO_value')

regression model score for ozone_value is 0.3361198894913202
regression model score for PM2.5_value is 0.07918711866345374
regression model score for CO_value is 0.2503233295299714
```

Figure 5. Training loss vs. validation loss on CO value prediction with the ANN model
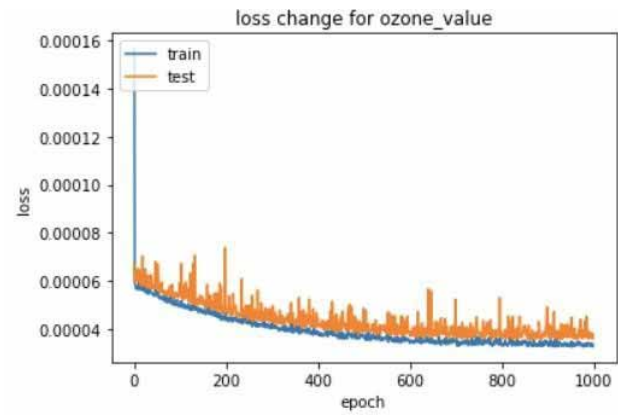
The authors built an RNN model with three hidden layers and set the return sequence as true. Next, they calculated the loss through the mean squared error method, using rmsprop as the optimizer when compiling the model. Figures 6 through 8 display the training loss and validation loss on CO, O3, and PM2.5 values with the RNN model. The RNN model performed best on PM2.5 values, with training losses for all three AQIs decreasing. The validation loss of CO fluctuated more than the other two AQIs. Still, there was a downward trend.

Figure 6. Training loss vs. validation loss on CO value prediction with the RNN model

Figure 7. Training loss vs. validation loss on ozone value prediction with the RNN model
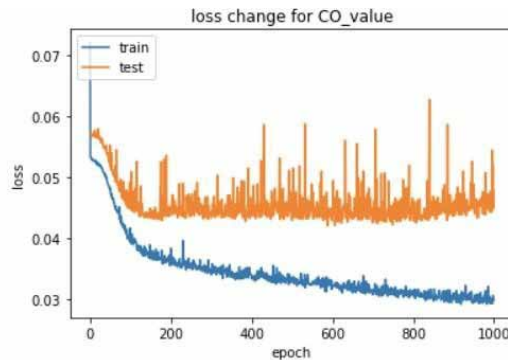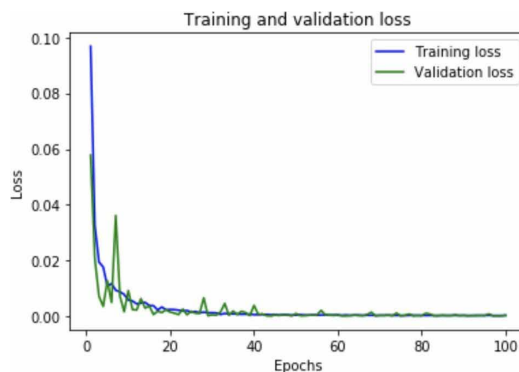

Figure 8. Training loss vs. validation loss on PM2.5 value prediction with the RNN model

The authors built a GRU model with three hidden layers, set the return sequence as true, and used *tanh* as the activation function. They calculated the loss through the mean squared error method and used SGD as the optimizer when compiling the model. Figures 9 through 11 display the training loss and validation loss on CO, O3, and PM2.5 values with the GRU model. The GRU model outperformed the ANN and RNN models, with training losses for all three AQIs decreasing. The validation loss for all three AQIs decreased significantly as the number of training epochs increased.

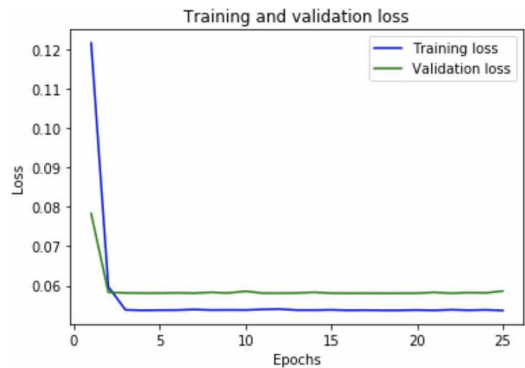**Figure 9. Training loss vs. validation loss on CO value prediction with the GRU model**

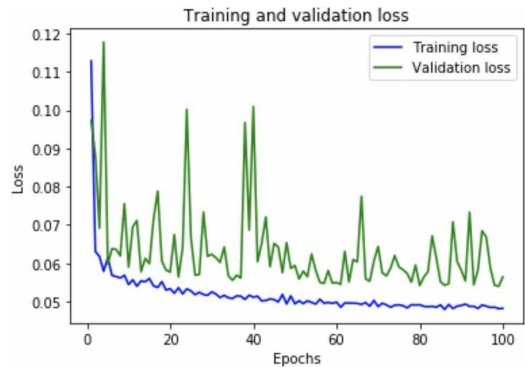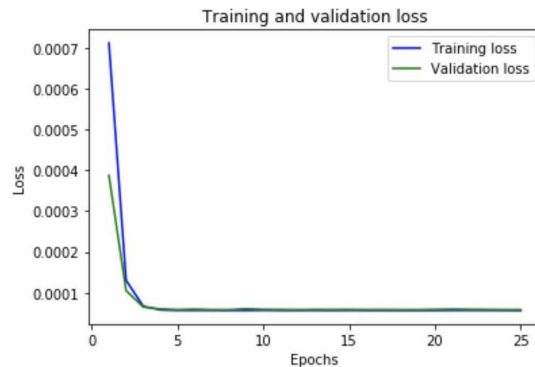**Figure 10. Training loss vs. validation loss on ozone value prediction with the GRU model**

Figure 11. Training loss vs. validation loss on PM2.5 value prediction with the GRU model

SYSTEM IMPLEMENTATION

System Requirements

The proposed system focused on individuals whose lives were affected by the wildfires to predict the air quality impact of California wildfires (Huang et al., 2021). Users can use this application to predict the impact of wildfires on air quality by looking at the PM2.5, CO, and O3 information (data collected over 15 years. The air quality impact due to wildfire application system will benefit people with health conditions related to air quality sensitivity and those working to contain wildfires (e.g., firefighters, first responders, and people living close to wildfire areas). In addition, the system will actively learn about fires as they occur throughout the year.

The deep learning models can help users learn about and understand the relationships between air quality indicators and fire parameters. However, the system has limitations. For instance, multiple factors could negatively affect the machine learning and deep learning models' prediction accuracy. Although fire data has been collected from the past 15 years, there are only tens or hundreds of fires per year in California. Most wildfires occur in the summer and autumn (June to November), making the prediction less accurate in the winter and spring. Furthermore, the dynamic nature of air quality means that other potential features could affect the air, such as ocean currents, geological activities, and human activities. The current model will focus, therefore, on analyzing air quality due to wildfire activities.

The high-level data analytics in this project displayed air quality before, during, and after wildfires. It also captured information for major fires that left a large impact.

### System Design and Architecture

The authors' prototype is a Web-based application with a Python Flask framework. Most of the artificial intelligence-powered function components are written in Python. The authors used an AWS cloud environment in the backend to deploy the application. It is seamlessly integrated with the front-end. The study deployed the authors' application on an AWS cloud platform that supports scalable computing capabilities (Huang et al., 2021).

For data management, the authors generated two datasets through an API provided by the EPA weather data. These include file_data (filtered by duration) and AQI data (containing all air quality data collected by sensors from 2000 to 2015 for each county in California). As explained, these datasets served as input variables. When the Web application was deployed in the production mode, the machine learning and deep learning models would be operational and available online.

User Interface Design

The study used Tableau to design the visualizations and interactive dashboard on the online application for user interface design and data visualization. The visualization included two maps that served as filters for each other. It also included a legend to help users select a date range and move forward along the timeline automatically to represent the data as tooltips on map points. The indicator legend displayed a continuous color range to represent fire size, with orange representing the smallest and red representing larger fires (> 100000). See Figure 12.

Figure 12. Legends in the dashboard
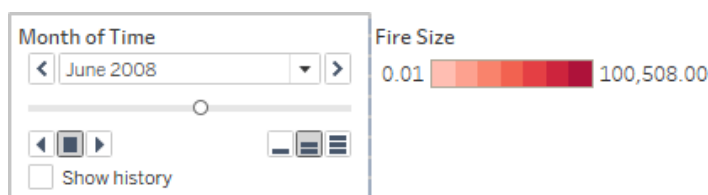
For example, Figure 13 depicts all wildfires in June 2008 across California. The left map shows the exact location of each wildfire (latitude and longitude). The right map displays wildfires that occurred during a specified time by county. Clicking on a county displays a tooltip with the exact numbers for the summation of fire size, CO, O3, PM2.5, and temperature within a given time. The interactive visual created in Tableau has significantly improved the use of the Website.

In addition to Tableau visualizations, the authors developed a report generation feature that can be accessed by clicking the "generating report" button. This report demonstrates the statistical analysis and changing plots for air quality throughout the duration of the wildfire. Figures 14 and 15 are examples of how the air changes with wildfires that lasted more than 100 days. The CO value shows a significant increase after the fire has burned or across several weeks. The O3 value decreases as the fire burns.
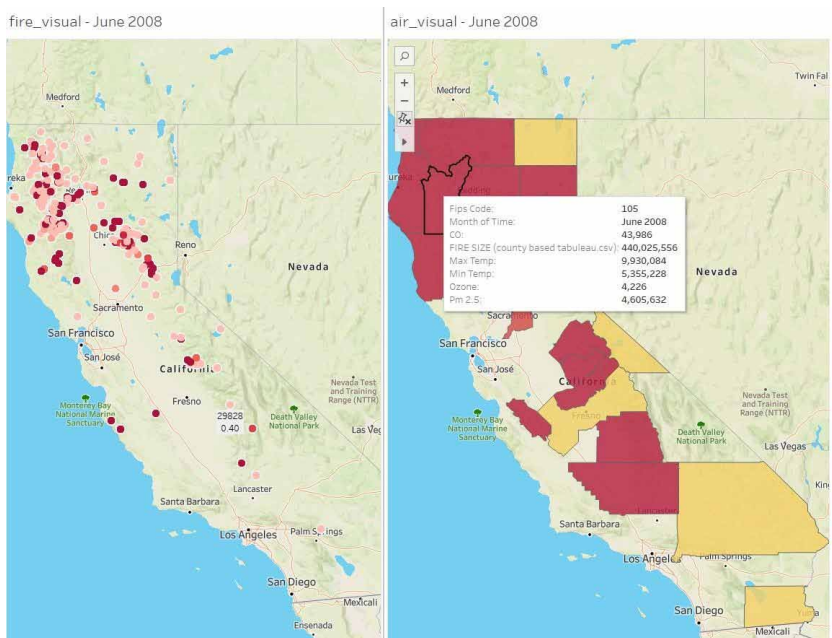
Figure 13. California wildfire (June 2008)

Figure 14. Change of CO value caused by wildfires
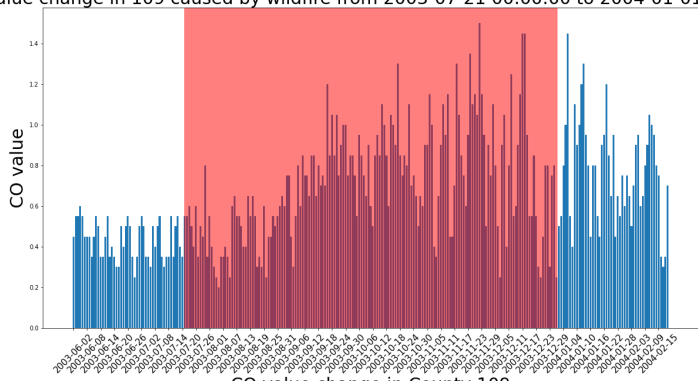
Figure 15. Ozone value change caused by wildfires

Figure 16. Change in PM2.5 in 2006 at county 19
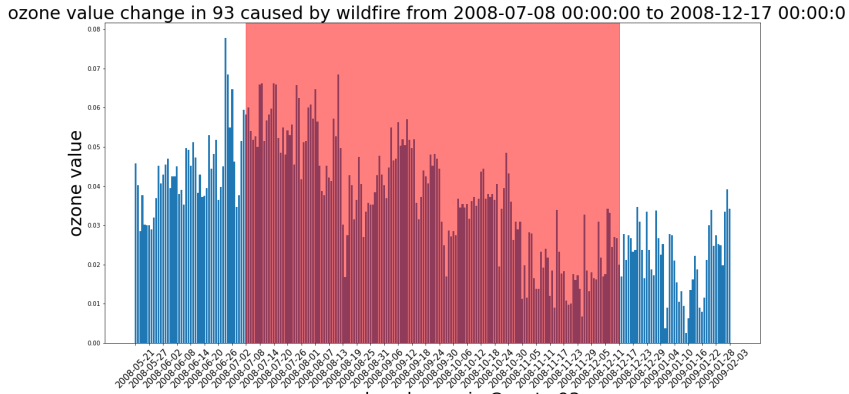
The authors generated 21 bar charts for the three air quality indices (O3, PM2.5, and CO). Users can select the pollutant they need to view. Furthermore, the authors developed three-dimensional (3D) plots to visualize the trend for air quality indices and the relationship of fire size and air quality. Each 3D plot represents a single long duration fire for one county in one year. The x-axis represents days for one year from January 1 through December 31. The y-axis represents the fire size. The z-axis represents the air quality value for one AQI.

Figures 16 and 17 display the change in PM2.5 (2006 and 2008) at county 19. Through these visualizations and reports, users can gain a better understanding of the relationship between wildfires and air quality.
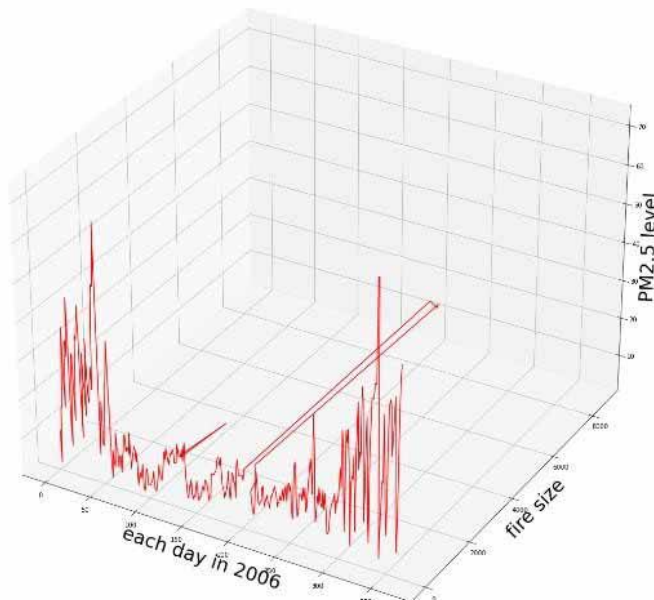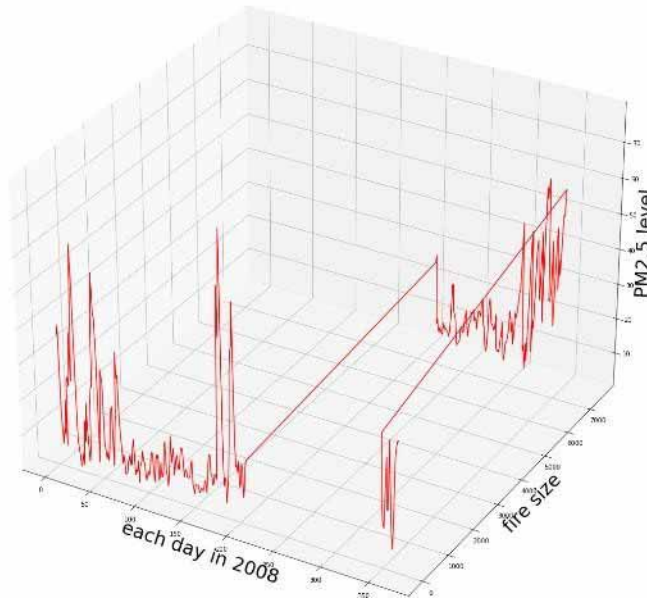
Figure 17. Change in PM2.5 in 2008 at county 19

One long duration fire happened in 2008 at County 019, Please see the PM2.5 change

Prototype Implementation

The authors conducted experiments with various machine learning models, including linear regression, ANN, RNN, and GRU. Each has its own strengths and weaknesses in predicting outcomes. The study aimed to develop a prototype that gives users a seamless experience and delivers accurate predictions. To achieve this, the authors implemented the GRU model, which has demonstrated the highest accuracy compared to other models. Although the study did not include other deep learning models in the prototype, the authors provided brief introductions and training results for models like RF, ANN, and logistic regression.

During a wildfire season, predicting air quality is crucial when making life-saving decisions. The authors aimed to deliver superior results and optimized outcomes. They fine-tuned parameters for each layer and continuously trained their machine learning models to minimize loss. For this specific case, they chose a GRU model with five layers, a return sequence feature, and *tanh* activation functions. The experiments show that the GRU model yields excellent results. Thus, the authors were ready to integrate it into the system.

To implement the prototype, the authors needed to ensure that all required features and functional components were included:

- Provide accurate predictions of air quality based on relevant fire parameters.
- Offer insightful and meaningful data visualizations based on statistical and exploratory data analysis.
- Provide an explanation of the generated prediction result, such as the meaning of the predicted variable value based on the machine learning models.
- Offer the latest summary statistics on the main page of the application along with instant news updates regarding the fire to users of interest (e.g., firefighters and county residents).

The application has three main tabs. The first tab, Overview, displays statistics, insights, and current information about a wildfire or the fire season. This page includes financial losses, millions of acres burned, number of studies performed in the wildfire field, and number of people injured or dead due to the wildfire. Additionally, three of the authors' best-performing machine learning and deep learning models are listed on this page.

The second tab, Data Preparation, illustrates the key parts of the data cleaning and preparation process in the authors' project. First, it presents the filtering process, corner/nullable/invalid data processing, and merging dataset. Second, it prepares the information into the machine learning ready state dataset.

The third tab, Result, displays both our artificial intelligence-powered model prediction results and the authors' insightful data visualization graphs and animations. The first portion of this tab displays best-performing prediction scores in terms of error loss. The authors also enable a form to allow users to perform current or future wildfire impact on air quality. The information will be processed on the back end, producing a prediction value displayed to the user. The output includes O3, CO, and PM2.5 value. Additionally, based on these values, the authors determine whether the predicted air quality is healthy, okay, or hazardous for people going outside without a mask.

In the second portion of the third tab, the authors provide two major categories of data visualization. The first category shows the specific air quality metric before, after, and during a fire. It is displayed in a two-dimensional fashion, with CO, O3, or PM2.5 as the Y-axis and time of interest as the X-axis. The second category is an interactive, animated heat map, which demonstrates the past 15 years of California fires with respect to fire size per county and specific location. These

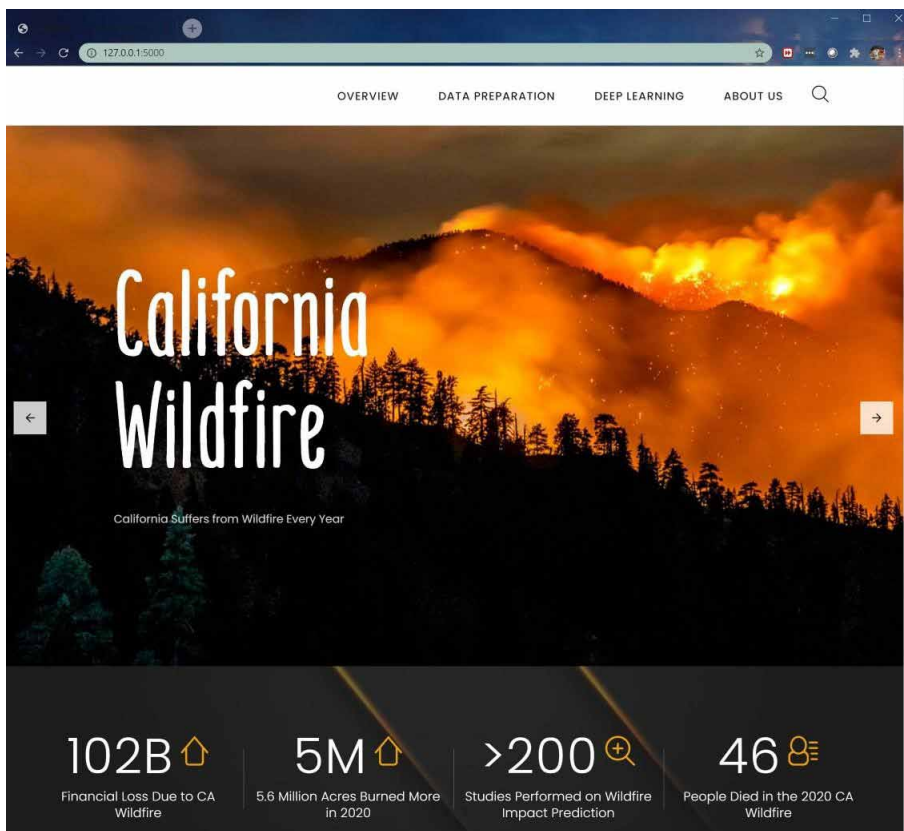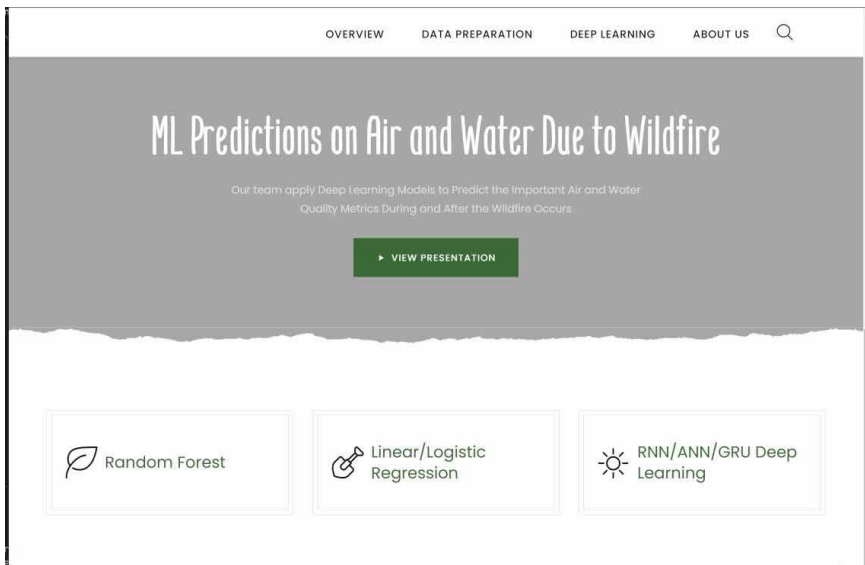**Figure 18. California wildfire portal**

Figure 19. Predictions on air and water due to wildfire

are displayed side by side. Users can toggle the 15-year time interval to see months with fires, fire size, and impact on air quality.

The authors also included a user interface for raw data source (see Figure 20). This interface enables users to access the raw data used in the research and understand the data cleaning and preparation process.

Visualization

After the authors implemented various models, they developed a prototype with a visualization tool (see Figure 22). This enables users to receive notifications of future air quality in terms of CO, PM2.5,
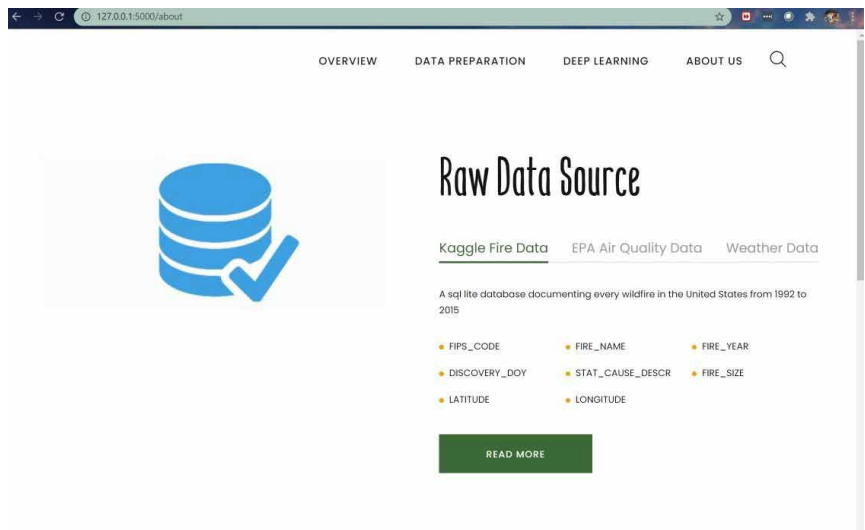
Figure 20. Raw data sources

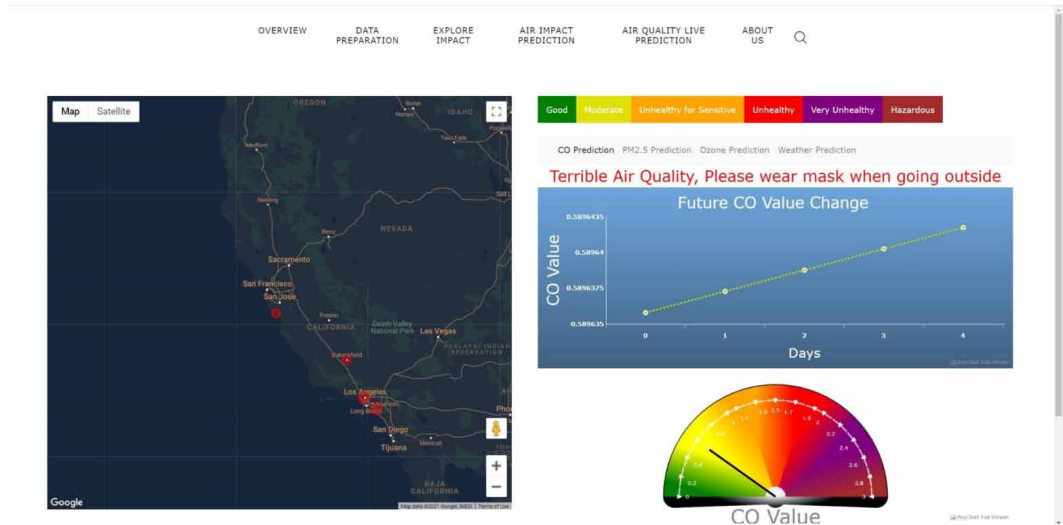**Figure 21. AQI categories defined by EPA**

Figure 22. Dashboard showing CO value prediction
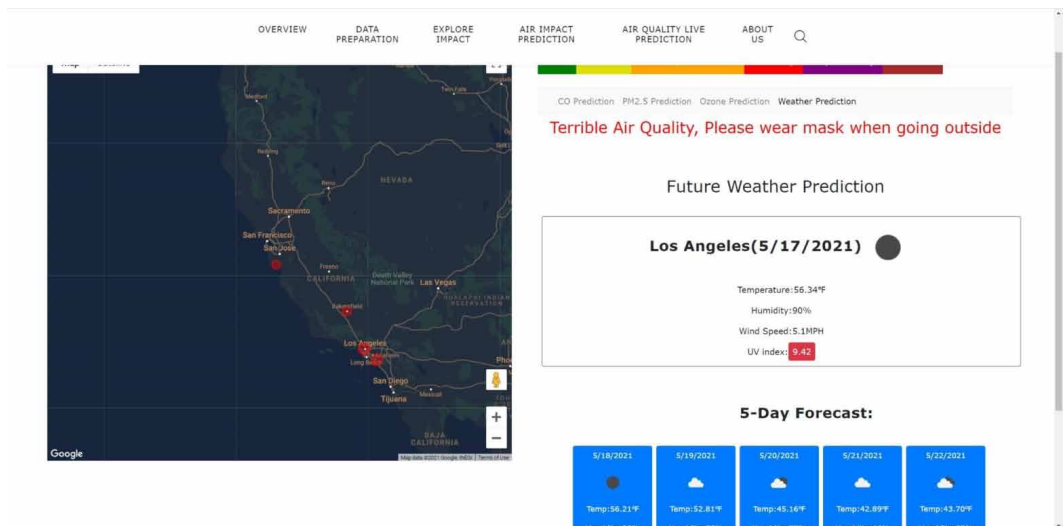
and O3 concentration indexes. Thus, they can act accordingly, save time, and save lives during and after a wildfire.

As mentioned, AQI is an index utilized by government agencies to measure the level of air pollution. The EPA reports that AQI values range from 0 to 500, with higher values indicating higher levels of pollution. To interpret AQI values, it is important to reference the classification in Figure 21. The authors hope this study will become a trustworthy source of weather, air, and wildfire reporting systems, with clear visualizations including line charts, interactive maps, and 3D plots. They aim to serve the purpose of providing appropriate predictions and offer recommendations (Huang et al., 2021).

This prototype has two main parts for visualization. These can display the experimental results along with data analytics. The first part is the authors' data visualization result from the data analytics

section. They developed the Tableau interactive visualization, which allows users to observe changes over 15 years related to wildfires in California. Furthermore, the authors display their data analytics results, such as 2D and 3D wildfire-related charts and graphs.

The authors also developed a map-based UI (User Interface) that displays statistics of active wildfires in California and uses the authors' deep learning model to predict air quality for the next five days. When the user clicks on the county where the fire is located, the visualization tool generates predictions for CO, O3, and PM2.5. The map-based UI displays relevant information like the fire's position and impact radius. The four tabs on the right display air quality levels for the next five days, allowing users to plan their outdoor activities accordingly. These predictions are powered by artificial intelligence and derived from the study's deep learning model. Additionally, the authors included a fourth tab. It includes valuable information like weather forecasts for the next five days, temperature, humidity, wind speed, and ultraviolet index. Overall, the map-based UI provides a user-friendly way to visualize data and predictions related to active wildfires in California, empowering users to stay informed and make informed decisions.

The next feature of the visualization tool allows users to input their own wildfire metadata and predict wildfire air quality values using the methodology as described. On the left-hand side, the authors created a form where users can configure their own fire metadata, including county name, fire size, fire duration, mean precipitation, and min-max temperature. Three linear gauge labels are displayed on the right-hand side, with each representing a single air quality metric prediction value. These values are shown on the top of the tab and the pointer inside the linear gauge label. Overall, this feature provides a user-friendly way to input individual data and receive predictions about future wildfire air quality levels.

CONCLUSION AND FUTURE WORK

This study investigated the impact of wildfires on air quality and developed a Web-based application to provide accessible information on air quality patterns after a wildfire event. The authors collected data from publicly available trustworthy sources and used machine learning models to extract data for their application. Then, the authors trained multiple models, choosing a GRU model based on its

Figure 23. Dashboard showing weather prediction

Figure 24. Impact of wildfires on air quality prediction

Air Quality Index		
AQI Category and Color	**Index Value**	**Description of Air Quality**
Good Green	0 to 50	Air quality is satisfactory, and air pollution poses little or no risk.
Moderate Yellow	51 to 100	Air quality is acceptable. However, there may be a risk for some people, particularly those who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups Orange	101 to 150	Members of sensitive groups may experience health effects. The general public is less likely to be affected.
Unhealthy Red	151 to 200	Some members of the general public may experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy Purple	201 to 300	Health alert: The risk of health effects is increased for everyone.
Hazardous Maroon	301 and higher	Health warning of emergency conditions: everyone is more likely to be affected.

high accuracy and low loss rate. The prototype Website showcases the authors' research findings, datasets, visualizations, and live air quality predictions through a map with ongoing wildfires.

The proposed framework serves two functions. First, individuals in northern California can monitor current wildfire situations and receive air quality predictions for the next five days. The model also advises on whether the predicted air quality is dangerous for human health and recommends whether to stay indoors. Second, firefighters and researchers can use the artificial intelligence-powered model to predict air quality based on imaginary fire scenarios and estimate their impact on air quality in northern California. Lastly, the authors present a data-driven approach for deploying a Web-based wildfire air quality alert system. The proof-of-concept implementation demonstrates the novel functionalities of the proposed framework, benefiting both the public and city officials. Through this study, individuals can take precautions to keep themselves safe and follow the authors' predictions or recommendations to prepare for potential impacts caused by wildfires.

However, the current solution only considers wildfire and weather data. Thus, it cannot cover all factors that impact air quality, such as industrial exhaust emissions and natural disasters like sandstorms. The authors acknowledge the limitations of the sample input data, noting room for improvement in

modeling, tuning, and data preprocessing. In future work, the authors plan to incorporate factors that contribute to air quality during a wildfire to gain better insights based on other datasets.

The current approach used a trained neural network to predict CO, O3, and PM2.5 values. It is more accurate and provides better insights into regional air quality. The Web-based application offers historical wildfire impact analysis, appropriate charts, and explanations for easy understanding of the authors' approach and prior wildfires in northern California.

The current study limited its scope to California. It also used precipitation and maximum and minimum temperatures as proxies for weather. However, future work could incorporate additional features like wind direction, wind speed, and atmospheric pressure. This would allow the authors' machine learning model to produce more accurate predictions and improve its overall capability and robustness. Furthermore, the authors can expand their scope to include the western portion of the U.S. or the entire country.

The prototype, including the online application, provides California residents with real-time and future air quality information, as well as a visual representation of a wildfire's extent. By regularly checking the application, individuals can make informed decisions about outdoor activities and adjust their schedules if the air quality is poor.

The effects of wildfires on people and the economy are significant in California. Therefore, the authors aim to assist residents and firefighters impacted by wildfires through the use of their air quality prediction study. The work has provided people who may be impacted by wildfires with predictions and recommendations to prepare for potential impacts caused by wildfires. These recommendations include circulating ambient air with air filters, limiting outdoor time, and wearing masks. The authors hope their research can also be used by city governments and citizens concerned with public safety.

REFERENCES

Camia, A., & Amatulli, G. (2009). Weather factors and fire danger in the Mediterranean. In E. Chuvieco (Ed.), *Earth observation of wildland fires in Mediterranean ecosystems* (pp. 71–82). Springer-Verlag, doi:10.1007/978-3-642-01754-4_6

Castelli, M., Martins Clemente, F., Popovič, A., Silva, S., & Vanneschi, L. (2020). A machine learning approach to predict air quality in California [New Models, New Technologies, New Data and Applications of Urban Complexity from Spatio-temporal Perspectives]. *Complexity*, *2020*, 2020. doi:10.1155/2020/8049504

de Groot, W. J., Field, R. D., Brady, M. A., Roswintiarti, O., & Mohamad, M. (2007). Development of the Indonesian and Malaysian fire danger rating systems. *Mitigation and Adaptation Strategies for Global Change*, *12*(1), 165–180. doi:10.1007/s11027-006-9043-8

Demuzere, M., Trigo, R. M., Vila-Guerau de Arellano, J., & Van Lipzig, N. P. M. (2009). The impact of weather and atmospheric circulation on O3 and PM10 levels at a rural mid-latitude site. *Atmospheric Chemistry and Physics*, *9*(8), 2695–2714. doi:10.5194/acp-9-2695-2009

Dominick, D., Juahir, H., Latif, M. T., & Aris, A. Z. (2012). An assessment of influence of meteorological factors on PM sub (10) and NO sub (2) at selected stations in Malaysia. *Sustainable Environment Research*, *22*(5), 305–315.

Elia, M., D'Este, M., Ascoli, D., Giannico, V., Spano, G., Ganga, A., Colanelo, G., Lafortezza, R., & Sanesi, G. (n.d.). *Estimating the probability of wildfire occurrence in Mediterranean landscapes using artificial neural networks*. Academic Press.

Guan, S., Wang, D. C., Gao, Y., Zhang, T., & Pouliot, G. (2020). Impact of wildfire on particulate matter in the southeastern United States in November 2016. *The Science of the Total Environment*, *724*, 138354. doi:10.1016/j.scitotenv.2020.138354 PMID:32272416

Huang, C., Li, S., Qiao, Y., & Yu, Y. (2021, December). *Wildfire air quality prediction: Project report.* San Jose State University.

Jaffe, D. A., Wigder, N., Downey, N., Pfister, G., Boynard, A., & Reid, S. B. (2013). Impact of wildfires on ozone exceptional events in the western U.S. *Environmental Science & Technology*, *47*(19), 11065–11072. doi:10.1021/es402164f PMID:23980897

Jain, P., Coogan, S. C. P., Subramanian, S. G., Crowley, M., Taylor, S., & Flannigan, M. D. (2020). A review of machine learning applications in wildfire science and management. *Environmental Reviews*, *28*(4), 478–505. doi:10.1139/er-2020-0019

Noorian, F. (2015). *Risk management using model predictive control* [Thesis].

Preisler, H. K., Schweizer, D., Cisneros, R., Procter, T., Ruminski, M., & Tarnay, L. (2015). A statistical model for determining impact of wildland fires on Particulate Matter ($PM_{2.5}$) in Central California aided by satellite imagery of smoke. *Environmental Pollution*, *205*, 340–349. doi:10.1016/j.envpol.2015.06.018 PMID:26123723

Preisler, H. K., & Westerling, A. L. (2007). Statistical model for forecasting monthly large wildfire events in western United States. *Journal of Applied Meteorology and Climatology*, *46*(7), 1020–1030. doi:10.1175/JAM2513.1

Ramona-Gottschling, I., & Gottschling, M. (2016). Taxonomic revision of Rochefortia Sw. (Ehretiaceae, Boraginales). *Biodiversity Data Journal, 4*, E7720. , 10.3897/BDJ.4.e7720

Reid, C. E., Considine, E. M., Watson, G. L., Telesca, D., Pfister, G. G., & Jerrett, M. (2019). Associations between respiratory health and ozone and fine particulate matter during a wildfire event. *Environment International*, *129*, 291–298. doi:10.1016/j.envint.2019.04.033 PMID:31146163

Reid, C. E., Jerrett, M., Petersen, M. L., Pfister, G. G., Morefield, P. E., Tager, I. B., Raffuse, S. M., & Balmes, J. R. (2015). Spatiotemporal prediction of fine particulate matter during the 2008 Northern California wildfires using machine learning. *Environmental Science & Technology*, *49*(6), 3887–3896. doi:10.1021/es505846r PMID:25648639

Tian, X. R., McRae, D. J., Jin, J. Z., Shu, L. F., Zhao, F. J., & Wang, M. Y. (2011). Wildfires and the Canadian forest fire weather index system for the Daxing'anling region of China. *International Journal of Wildland Fire*, *20*(8), 963–973. doi:10.1071/WF09120

Tonini, M., D'Andrea, M., Biondi, G., Esposti, S. D., Trucchia, A., & Fiorucci, P. (n.d.). *A machine learning-based approach for wildfire susceptibility mapping.* The Case Study of the Liguria Region in Italy.

Vila, L., Gómez, I., Martínez-Vega, J., Echavarría, P., Riaño, D., Martín, M. P. (2016). *Multitemporal modelling of socio-economic wildfire drivers in central Spain between the 1980s and the 2000s: Comparing generalized linear models to machine learning algorithms.* 10.1371/journal.pone.0161344

Wang, S. C., & Wang, Y. (n.d.). *Quantifying the effects of environmental factors on wildfire burning areas in the south central US using integrated machine learning techniques.* Academic Press.

Watson, G. L., Telesca, D., Reid, C. E., Pfister, G. G., & Jerrett, M. (2019). Machine learning models accurately predict ozone exposure during wildfire events. *Environmental Pollution*, *254*, 112792. doi:10.1016/j.envpol.2019.06.088 PMID:31421571

Xu, Y., ChakHo, H., Wong, M., Deng, C., Ta-ChienChan, Y., & Knudby, A. (n.d.). *Evaluation of machine learning techniques with multiple remote sensing datasets in estimating monthly concentrations of ground-level $PM_{2.5}$.* Academic Press.

Yao, J., Brauer, M., Raffuse, S., & Henderson, S. B. (2018). Machine learning approach to estimate hourly exposure to fine particulate matter for urban, rural, and remote populations during wildfire seasons. *Environmental Science & Technology*, *52*(22), 13239–13249. doi:10.1021/acs.est.8b01921 PMID:30354090

Zhana, Y., Luo, Y., Deng, X., Chen, H., Grieneisenb, M. L., You, X., & Shen, L. (2017). *Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm.* Academic Press.

Subhankar Dhar is a Professor with the School of Information Systems and Technology at San José State University. He is also an affiliate faculty member of the Silicon Valley Center for Entrepreneurship. He advises early-stage high tech startups and mentors entrepreneurs. Subhankar's research interests are data science, big data analytics, mobile, and cloud computing as well as wireless networks. In addition, he has also worked on projects related to smart cities and his work on Data Science has been funded by NSF. He teaches a variety of courses including computer networks, distributed systems, database systems, and web-based computing. His publications have appeared in reputed international journals and gave presentations to various international conferences. He serves as a member of the editorial board of International Journal of Business Data Communications and Networking. He is a reviewer of papers for various international journals, conferences and scholarly publications. He also served as a member of the organizing committee of various international conferences including IEEE ANTS, IEEE Smart World Congress, Workshop on Large Scale Complex Network Analysis. Subhankar is a senior member of IEEE and has several years of industrial experience in software development including product planning, design, and information systems management.

Jerry Gao is a professor of Computer Engineering at San Jose State University.