Framework to Create Dataset for Disaster Behavior Analysis using Google Earth Engine: A Case Study in Peninsular Malaysia for Historical Forest Fire Behavior Analysis

Yee Jian Chew *, Shih Yin Ooi *, Ying Han Pang, Zheng You Lim

Posted Date: 1 April 2024

doi: 10.20944/preprints202404.0027.v1

Keywords: Disaster behavior; forest fire behavior; forest fire dataset; data extraction framework; Google Earth Engine; remote sensing; Malaysia; ChatGPT; Noteable; Large Language Model

Article

Framework to Create Dataset for Disaster Behavior Analysis using Google Earth Engine: A Case Study in Peninsular Malaysia for Historical Forest Fire Behavior Analysis

Yee Jian Chew, Shih Yin Ooi *, Ying Han Pang and Zheng You Lim

Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, Melaka, Malaysia; chewyeejian@gmail.com, syooi@mmu.edu.my, yhpang@mmu.edu.my
* Correspondence: syooi@mmu.edu.my

Abstract: This research presents a comprehensive framework for efficiently generating forest fire datasets from Google Earth Engine data sources. The primary contribution of this work lies in providing a methodology to swiftly extract forest fire factors without the need for permissions or access to private datasets, rendering the dataset openly accessible and shared without barriers. Furthermore, given that the remote sensing data used is a global dataset, it can be applied in any region without restrictions. In this study, Peninsular Malaysia is chosen as a case study to demonstrate the framework's effectiveness. The generated dataset includes essential variables including the climate and environment, landcover, topography, and anthropogenic factors facilitating the analysis of fire occurrences. The methodology empowers data scientists, enabling them to leverage their analytical skills on the extracted dataset without requiring specialized remote sensing knowledge. Additionally, this study also showcases the adoption of large language models, specifically GPT-4 with the Noteable plugin, as a tool for conducting preliminary analyses on the generated dataset. Sample analyses reveal that several key features, including the KBDI, LST, PDSI, climate water deficit, and precipitation, significantly impact forest fire occurrences in Peninsular Malaysia. Despite the successful application of the GPT-4 with Noteable plugin, certain limitations and challenges are identified, highlighting the necessity for further validation of the tool's applicability and limitations. This study encourages future research to (1) adopt the proposed framework in other regions, (2) explore more detailed analyses encompassing all variables, and (3) leverage machine learning for advanced forecasting.

Keywords: disaster behavior; forest fire behavior; forest fire dataset; data extraction framework; Google Earth Engine; remote sensing; Malaysia; ChatGPT; noteable

1. Introduction

Forest fires pose a significant ecological and societal challenge due to their destructive impact and potential harm to human communities. Recent advances in machine learning and data analysis serve as promising technology for understanding and mitigating these disasters [1,2]. However, a substantial obstacle has persisted—the tedious and laborious process of collecting and preparing the essential data for analysis [3–6]. Many existing projects have relied on government intervention or costly data acquisition methods, limiting their accessibility and scalability. This study addresses this critical issue by presenting a simplified approach that harnesses the power of Google Earth Engine (GEE) [7]. Our primary objective is to streamline the behavioral analysis of forest fires, making it more accessible to researchers and practitioners.

 doi:10.20944/preprints202404.0027.v1

A recurring question pertains to the emphasis on real-time detection over behavioral analysis of forest fires. While real-time detection is unquestionably essential [8–12], it often necessitates collaboration from multiple stakeholders, including government agencies, firefighters, and satellite providers, making it resource-intensive and costly. Our contribution focuses on enhancing the understanding of forest fire dynamics and historical patterns, contributing to more effective long-term forest fire management. To provide context for our research, previous studies related to historical forest fire data extraction leveraging remote sensing data will be reviewed. This review will help us ascertain whether any previous efforts within the forest fire domain have attempted to simplify the data collection and analysis process, serving as a foundational background study. It will also highlight the existing knowledge gap and set the stage for our innovative approach within the broader field of fire ecology and management.

Our research objective aims to develop a simplified approach using GEE to locate historical fire locations and extract relevant factors. This simplification of data acquisition is crucial because many advanced machine learning algorithms have primarily focused on algorithmic aspects, often overlooking the complexity of data collection and preprocessing. In many cases, such projects have required significant governmental involvement and substantial resources. Our approach seeks to shift the focus toward making the data readily available for analysis and machine learning purposes. The research question underlying our work is fundamental. It revolves around the importance of extracting historical fire data for analysis and machine learning purposes. Without access to this historical data, the scope of analysis remains limited, and the potential for comprehensive insights remains untapped. Moreover, it should be noted that despite the recent advancements in the fields of remote sensing and machine learning, a wealth of satellite data is readily available [13], yet it remains significantly underutilized, primarily due to the challenges associated with the lack of experts, accessible tools, and methodologies.

In this paper, our method that leverages GEE to efficiently locate historical fire locations, extract relevant factors, and conduct data analysis on pertinent factors is presented. This streamlined approach simplifies the entire process, empowering researchers to focus on refining algorithms or integrating new datasets. Additionally, it eliminates the need for resource-intensive data downloads and local processing, democratizing access to this type of analysis. Our study aims to facilitate the behavioral analysis of forest fires, offering a foundational framework for future research in this domain. Often, the initial hurdle researchers face is finding a starting point. Our approach addresses this challenge by simplifying the data acquisition and analysis stages. Our methodology effectively employs GEE's Python API and Geemap [14] package to achieve these objectives. While this paper emphasizes forest fire disasters, it is hypothesized that this methodology can be readily extended to address other types of disasters by simply providing the coordinates of the incidents.

On the other hand, to showcase the usefulness of the generated forest fire dataset, it is necessary to perform a set of analyses to validate its applicability. While traditional methods for analyzing the dataset are still considered the most preferred approach, this paper takes a distinct approach considering the proliferation of advancements in Large Language Models (LLMs). Specifically, ChatGPT [15], particularly the GPT-4 variant with the Noteable plugin [16] is exploited to demonstrate the utility of the forest fire dataset. At the end of the analysis, the potential uses, advantages, and limitations of this plugin are also discussed.

Section 2 discusses previous studies that have focused on extracting remote sensing data for forest fire analysis. In Section 3, the proposed framework for streamlining remote sensing data extraction for historical forest fire datasets is deliberated. This section also includes a comprehensive list of forest fire attributing factors, along with their sources and detailed information. Section 4 outlines the case study in Peninsular Malaysia, where the framework is applied to generate a forest fire dataset tailored to this specific location. Section 5 conducts sample analysis using LLM, specifically GPT-4 with the Noteable plugin to validate the generated forest fire dataset's applicability. Additionally, the section also delves into the potential applications and limitations of this analysis approach. Section 6 serves as the conclusion, summarizing the paper's contributions,

 doi:10.20944/preprints202404.0027.v1

while Section 7 outlines future work to enhance the proposed framework and validate the adoption of LLM in data analysis.

2. Literature Reviews and Background Study

In a recent systematic review conducted by [17], an extensive exploration was undertaken to unveil the most influential factors affecting forest fires. Their exhaustive analysis combed through a total of 144 factors from 94 publications spanning the years from 2001 to 2021. Among the factors that emerged as highly significant were slope, elevation, aspect, land cover, NDVI (Normalized Difference Vegetation Index), temperature, precipitation, windspeed, and more. Notably, the prevalence of these factors was attributed to their wide global availability in existing databases.

To perform data analysis [18,19] and develop machine learning model [3–6,20,21] in the domain of forest fires, it is essential to possess a dataset containing both fire locations and associated factors. For a comprehensive exploration of the application of data analysis and machine learning in the domain of forest fires, one can refer to the review papers [2,22].

Creating these datasets involves various researchers using their own pipelines, which may combine multiple datasets from global sources and government agencies, along with various Geographic Information System (GIS) [23,24] tools. These tasks are tedious and time-consuming, particularly for data scientists and machine learning engineers without a background in GIS. Moreover, it's essential to note that obtaining government data typically entails multiple layers of permissions and requests, making it a challenging and often non-shareable resource. Hence, the objective of this paper is to streamline the data collection process for building a forest fire dataset using publicly available datasets and resources, with the aim of increasing accessibility for data scientists.

A study closely aligned with our research was conducted by [25]. Their study aimed to assemble a comprehensive forest fire dataset, encompassing historical fire incidents and associated factors in Australia. This dataset served as the foundation for their investigation into the primary factors contributing to forest fires and the construction of machine learning models for predicting forest fire incidents. Their research relied upon key tools such as the GEE code editor JavaScript API [7] and made use of multiple global public satellite datasets as well as government datasets. However, replicating their methodology and adapting it to different geographical locations presented considerable challenges when using the provided scripts [26].

In contrast, our work emphasizes the use of globally publicly available datasets from the GEE catalog to ensure its flexibility in adapting to various locations. While government data or any private data are not considered in this paper, researchers can seamlessly integrate their datasets into the framework to enrich their analyses. Importantly, we prioritize the development of a swift data extraction method. Our primary purpose is to provide a means to quickly gather data for understanding forest fire occurrences in specific locations, establishing a robust foundation for other researchers to utilize this framework. Notably, none of the previous efforts in the forest fire domain have attempted to simplify the data collection process. Given these challenges, simplifying the process of data collection for forest fires can empower data scientists to leverage their expertise in better analyzing and understanding fire behavior in different locations.

3. Methodology

3.1. Proposed Framework

In this work, GEE is leveraged as the primary big data platform for accessing remote sensing data, providing a robust foundation for meaningful analysis. It has been widely adopted for various tasks, showcasing its versatility [27–31]. The significant increase in the availability of publicly accessible remote sensing data [13] presents a promising opportunity for research and analysis. However, obtaining historical forest fire data remains a formidable challenge. This challenge arises because the ownership of such data is predominantly vested in governmental authorities, which often necessitates a complex approval process for data access and utilization. In Figure 1, a

 doi:10.20944/preprints202404.0027.v1

comprehensive visual representation of the proposed framework employed is presented. It illustrates both the process of extracting historical forest fire locations and highlights the multifaceted factors that contribute to the occurrence of these fire locations.

In this study, the MCD64A1 Burnt Area (BA) dataset [32] is utilized to extract historical forest fire locations. As a default, this study leverages all the data with complete year availability (i.e., from 2001-01-01 to 2022-12-31) in MCD64A1 to harvest the historical fire points. This dataset offers a comprehensive global record of burned areas with a spatial resolution of 500 meters, facilitating the monitoring and analysis of wildfire and land cover change dynamics. MCD64A1 is chosen for this study instead of FireCCI51 v5.1 [33] and Globfire Fire Event [34] from the GEE catalog. This decision is based on two primary factors: its extended temporal availability within GEE and its efficacy in identifying small fires in our study area, Peninsular Malaysia [35].

To derive historical fire points, the MCD64A1 dataset is filtered based on the year of interest and the specific region. All detected fire locations are then extracted as fire points at a 1km² resolution, as shown in Figure 1, Step 1. To extract the year and month of the fire incidents, the '*system:index*' from MCD64A1 is utilized, with the first four digits denoting the year and the subsequent two digits indicating the month. Since MCD64A1 provides tentative '*BurnDate*' values ranging from 0 to 366, the date of month of each location are calculated using these burnt day values. A spatial resolution of 1km² is employed in this study, considering the size of burnt areas in the study location, Peninsular Malaysia, which predominantly comprises small fires, typically less than 100 hectares [36–43]. This spatial resolution is also consistent with the resolutions of other datasets used in the study, which are generally larger than 1km². It's important to note that in locations with larger-scale fires, increasing the resolution size may be necessary to prevent complications in GEE due to computational resource constraints.

Conversely, for the collection of non-fire points, all available burnt area data within the region of interest from the MCD64A1 dataset is utilized. The burnt areas are blended into a single image, depicting the historical burnt regions across all years. In consideration of potential omissions of nearby fires, an additional dilation morphological operation [44] is applied to expand the boundaries of the burnt regions, with the default radius and iteration value set to 2. To obtain the non-fire regions, we invert the selection of burnt area region with the region of interest. Subsequently, the GEE function *ee.FeatureCollectionRandomPoints* is employed to randomly extract non-fire points at a 1km² resolution, with the total number of non-fire points matching the total number of fire points. For the month and day of the fire for non-fire points, we can leverage the most recent available year in the MCD64A1 dataset. This approach is justified by the necessity of incorporating the most up-to-date data concerning non-fire locations to ensure a comprehensive understanding of non-fire occurrences in the current context. The process for extracting non-fire points is elucidated in Step 2 of Figure 1. It's important to emphasize that all the historical fire and non-fire points extracted include their respective coordinates (latitude and longitude), as they are essential for the next step involving importing the fire incidents back into GEE. After the extraction of historical fire and non-fire points, the points will be saved as *.csv* files for subsequent analysis. This decision to extract and store the data as *.csv* files is primarily driven by the consideration that GEE may encounter issues such as crashes or connection losses during processing. Therefore, using *.csv* files allows for the continuity of the process without the need to start the extraction of historical points from the beginning in case of disruptions in GEE operations.

 doi:10.20944/preprints202404.0027.v1

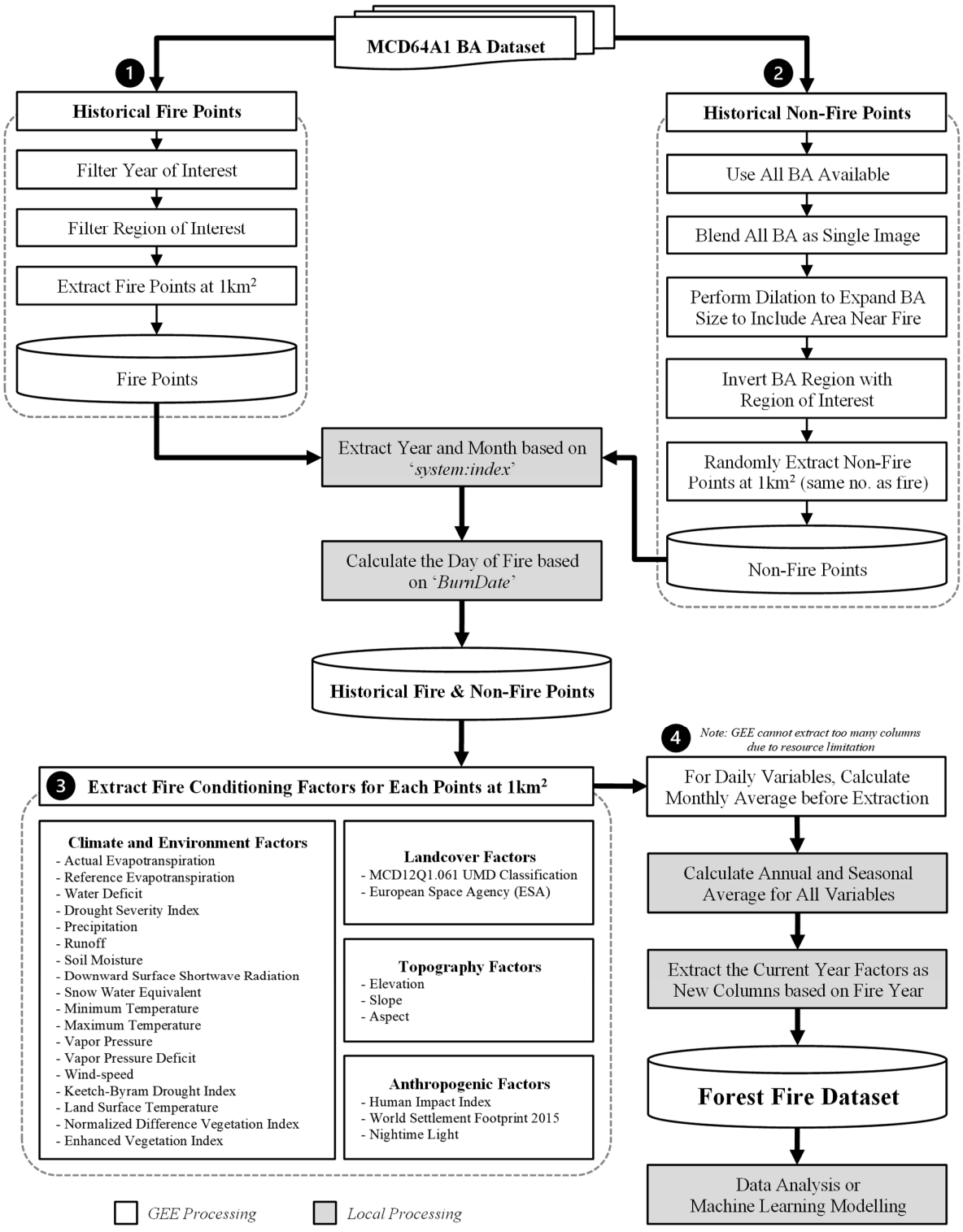

Figure 1. Overall Methodology to Build a Forest Fire Dataset.

Factors for each point are extracted based on the coordinates of both fire and non-fire points. Step 3 in Figure 1 lists all the conditioning factors exploited in this work, while Table 1 provides comprehensive information about each factor, including its source, temporal availability, temporal cycle, spatial resolution, etc. As this study aims to alleviate the challenges faced by data analysts and

 doi:10.20944/preprints202404.0027.v1

machine learning engineers in the extraction of remote sensing data, all fire factors based on their temporal availability within the range of BA temporal availability are extracted to maximize data accessibility for future analysts. However, it's worth noting that due to resource constraints within GEE, factors that are available on a daily basis (e.g., KBDI) will undergo processing inside GEE to compute the monthly averages before exporting.

In Step 4, additional processing is conducted locally to derive supplementary factors valuable for fire behavior analysis. For all factors available on a monthly basis, computation of annual and seasonal averages is performed for each year. The seasonal averages are determined by aggregating data over three-month periods: December-February, March-May, June-August, and September-November. This approach proves valuable as it acknowledges the substantial variations in fire behavior across seasons, which can be attributed to factors such as weather conditions, vegetation growth, and human activities. This aligns with established research findings [45,46] highlighting the importance of effectively capturing and analyzing these seasonal patterns. Subsequently, to gain a more comprehensive understanding of fire behavior, the factors associated with fire incidents for the specific year are extracted. This information is incorporated as a new column, referencing the year of each fire event obtained from MCD64A1. Historically, this step has been perceived as a complex and labor-intensive process, as generating datasets with high temporal resolution can pose practical challenges, as noted in prior studies [47]. Consequently, the forest fire dataset is assembled, encompassing all factors, including the monthly, annual, seasonal, and current-year fire-influencing variables, rendering it prepared for in-depth analysis. It is important to acknowledge that this dataset may contain missing data, mandating additional processing before effective utilization in analyzing data or training machine learning models.

One of the key advantages of the proposed framework is its universal applicability, as it relies solely on globally available, publicly accessible datasets. Given their public accessibility, the datasets generated can be readily shared and distributed without concerns related to copyright or privacy issues. For instance, the forest fire dataset produced in our study location, Peninsular Malaysia, is accessible at https://doi.org/10.5281/zenodo.10050852 [48] In this study, the GEE Python API is employed instead of GEE JavaScript. This choice allows for additional analysis and processing, tapping into the widespread utility of Python in data science and machine learning. Furthermore, Python offers the advantage of code reuse for various geospatial and data analysis tasks beyond the GEE environment. In conjunction with the GEE Python API, we also employ GeeMap [14,49], a Python package tailored for interactive geospatial analysis and visualization within the GEE framework. To facilitate the replicability of our proposed methodology and encourage its adoption in other geographical locations, the entire source code manipulated in this paper is readily accessible on GitHub https://github.com/chewyeejian/GEE_FrameworkForestFireDataset. It's important to note that the provided source code primarily focuses on our study location, Peninsular Malaysia. Modifying the default Country Feature Collection is necessary to adapt the code for use in different locations.

3.2. Forest Fire Attributing Factors Data Source and Details

Table 1 provides a comprehensive overview of the factors harnessed in this study, offering insights into their respective categories, sources, temporal availability, temporal cycle, spatial resolution, and more. As emphasized earlier, all the datasets used in this research are globally sourced, ensuring their adaptability across various locations without any hindrance. The selection of factors in this paper is founded on the prevalence of their usage and their potential high correlation with forest fire incidents, as indicated in existing literature [17]. It's important to highlight that the Human Impact Index (HII) from the Wildlife Conservation Society [50] is the sole dataset not directly available in the official GEE dataset catalog. However, it can be conveniently accessed through GEE.

 doi:10.20944/preprints202404.0027.v1

Table 1. Fire Conditioning Factors Data Source and Details.

Category	Source of Data	Temporal Availability	Temporal Cycle	Spatial Resolution	Annual Average	Monthly Average	Seasonal Average	Data Layers	Unit
Climate & Environment	TerraClimate [51]	1958-01-01 to 2022-12-01	Monthly	4km²	✓	✓	✓	Actual Evapotranspiration (AET)	mm
								Water Deficit (DEF)	mm
								Palmer Drought Severity Index (PDSI)	-
								Reference Evapotranspiration (PET)	mm
								Precipitation (PR)	mm
								Runoff (RO)	mm
								Soil Moisture (SOIL)	mm
								Downward Surface Shortwave Radiation (SRAD)	w/m²
								Snow Water Equivalent (SWE)	mm
								Minimum Temperature (TMMN)	°c
								Maximum Temperature (TMMX)	°c
								Vapor Pressure (VAP)	kPa
								Vapor Pressure Deficit (VPD)	kPa
								Wind-speed (VS)	m/s
	Rainfall [52]	2007-01-01 to 2023-09-12	Daily	4km²	✓	✓	✓	Keetch-Byram Drought Index (KBDI)	-
	MOD11A2.061 Terra [53]	2000-02-18 to 2023-08-29	8-day	1km²	✓	✓	✓	Land Surface Temperature (LST)	K
	MOD13Q1.061 Terra [54]	2002-02-18 to 2023-08-13	16-day	250m	✓	✓	✓	Normalised Difference Vegetation Index (NDVI)	-

 doi:10.20944/preprints202404.0027.v1

								Enhanced Vegetation Index (EVI)	-
Landcover	MCD12Q1.061 MODIS [55]	2001-01-01 to 2022-01-01	Annual	500m	✓			Annual University of Maryland (UMD) Classification (LC_Type2)	16 classes
	European Space Agency (ESA) [56] (static)	2021-01-01 to 2022-01-01	Annual	10m	✓			Landcover (Map)	11 classes
Topography	NASADEM [57] (static)	2000-02-11 to 2000-02-22		30m	✓			elevation	m
								Slope (derived from DEM)	degrees
								Aspect (derived from DEM)	degrees
Social Economic / Anthropogenic factors	Wildlife Conservation Society [50]	2001-01-01 to 2020-01-01	Annual	300m	✓			Human Footprint / Human Impact Index (HII)	-
	Deutsches Zentrum für Luft- und Raumfahrt [58]	2015-01-01 to 2016-01-01	Annual	10m	✓			World Settlement Footprint 2015 (settlement)	-
	VIIRS [59]	2012-04-01 to 2021-01-01	Annual	500m	✓			Nighttime light (average)	nanoWatts/sr/cm^2
Burn Area	MCD64A1.061 MODIS [32]	2000-11-01 to 2023-07-01	Monthly	500m	-	-	-	BurnDate	-

4. Application of the Proposed Framework in The Study Area – Peninsular Malaysia

4.1. Study Area – Peninsular Malaysia

In this study, Peninsular Malaysia serves as the chosen study location for assessing the proposed framework's effectiveness. This selection stems from the limited previous efforts to create a publicly available forest fire dataset for analytical purposes in this region [60]. Most prior works relied on private datasets sourced either from the Malaysian Government or private agencies [61]. In our approach, we utilize the level 2 administrative boundaries of Malaysia [62], refined to encompass only the states within Peninsular Malaysia. Employing these administrative boundaries enables the extraction of both fire and non-fire points to include their respective states and districts, which, in turn, facilitates subsequent analysis based on states or districts.

Following the extraction of historical fire and non-fire points in Steps 1 and 2 of Figure 1, a total of 5,557 fire points and 5,526 non-fire points are collected. However, visualizing such a substantial volume of points within GEE is not feasible due to the constraint, which would lead to a "*Request payload size exceeds the limit: 10,485,760 bytes*" error. To address this, QGIS is employed to visualize all

 doi:10.20944/preprints202404.0027.v1

the points, ensuring their inclusion in the figure, as depicted in Figure 2. It's important to note that for sample visualization, displaying a subset of the points is achievable in GEE.

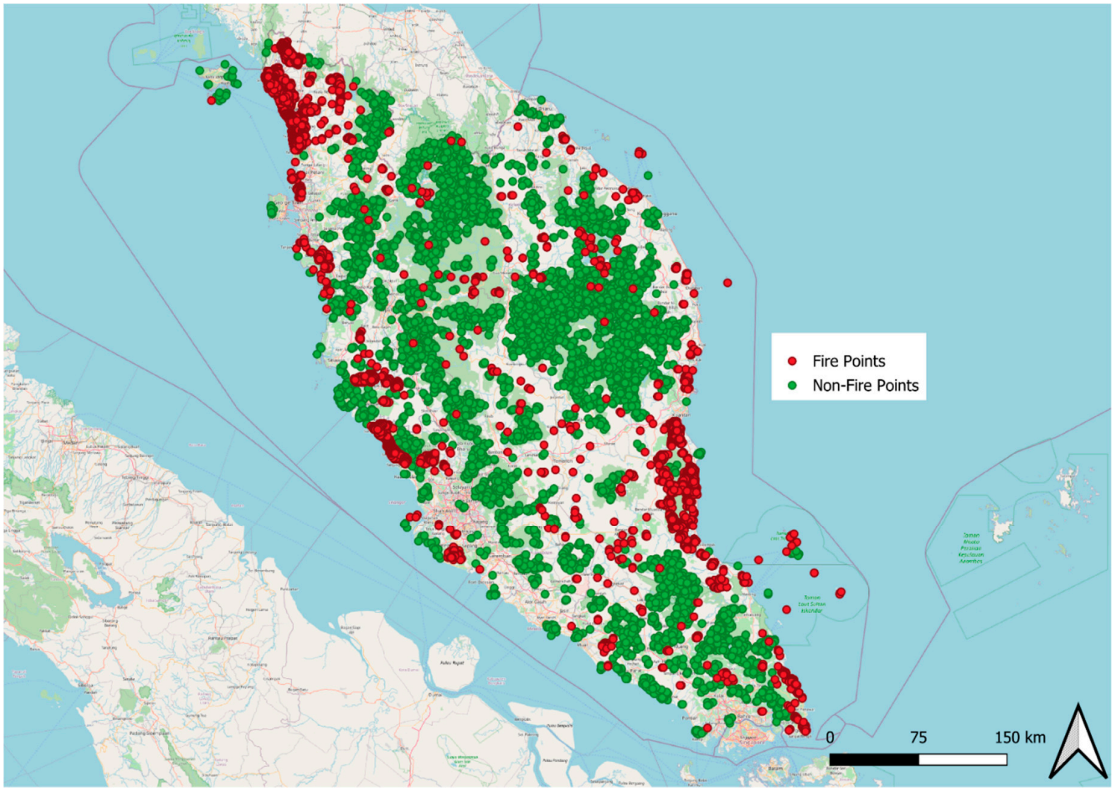

Figure 2. Fire and Non-Fire Points Distribution Illustrated in QGIS.

4.2. Peninsular Malaysia Forest Fire Dataset Description

In this Forest Fire Dataset, a total of 11,083 rows and 7,040 columns are present. This dataset comprises 11,083 instances, including 5,557 fire points and 5,526 non-fire points, each characterized by 7,040 features. After removing all columns that exclusively contain null values (i.e., *ADM2_REF*, *ADM2ALT2EN*, *ADM2ALT1EN*), a total of 7,037 columns remain, making them available for comprehensive analysis. The 5,557 fire points represent burned areas detected from January 1, 2001, to December 31, 2022, at a spatial resolution of 1km². In contrast, the 5,526 non-fire points default to the latest date in the analysis (December 31, 2022), reflecting the present context of non-fire scenes. The 7,073 columns encompass all monthly, annual, and seasonal factors detailed in Table 1, administrative boundary features, and burnt date information sourced from MCD64A1. The full, unprocessed dataset is freely accessible at https://doi.org/10.5281/zenodo.10050852 [48].

Figure 3 depicts the annual distribution of fire points from 2001 to 2022, revealing significant peaks in 2005 and 2014, which corresponds with the previous analysis of FIRMS hotspots [63]. An analysis of the datasets to identify missing data is presented in Figure 4, highlighting the top 20 features with the highest percentage of missing data. Among these variables, Land Surface Temperature (LST), nighttime light, and the HII stand out as those with the most substantial missing data. For LST, the missing data likely originates from its source. Regarding nighttime light and the HII, the high percentage of missing data can be attributed to the use of the latest available date (December 2022) as the reference date for non-fire points. The substantial missing data is caused by the limited temporal coverage of nighttime light data, which extends only until January 2021, while the HII covers data only until January 2020. While addressing missing data is not the primary focus of this paper, future studies may consider strategies such as substituting missing values with data from the previous year or month or using overall averages to fill the gaps. This paper will conduct a preliminary analysis to determine whether the generated datasets can offer insights into the behavior

 doi:10.20944/preprints202404.0027.v1

of fires in Peninsular Malaysia. In the next section, we will conduct a preliminary analysis to assess the potential of the generated datasets in providing insights into fire behavior in Peninsular Malaysia.

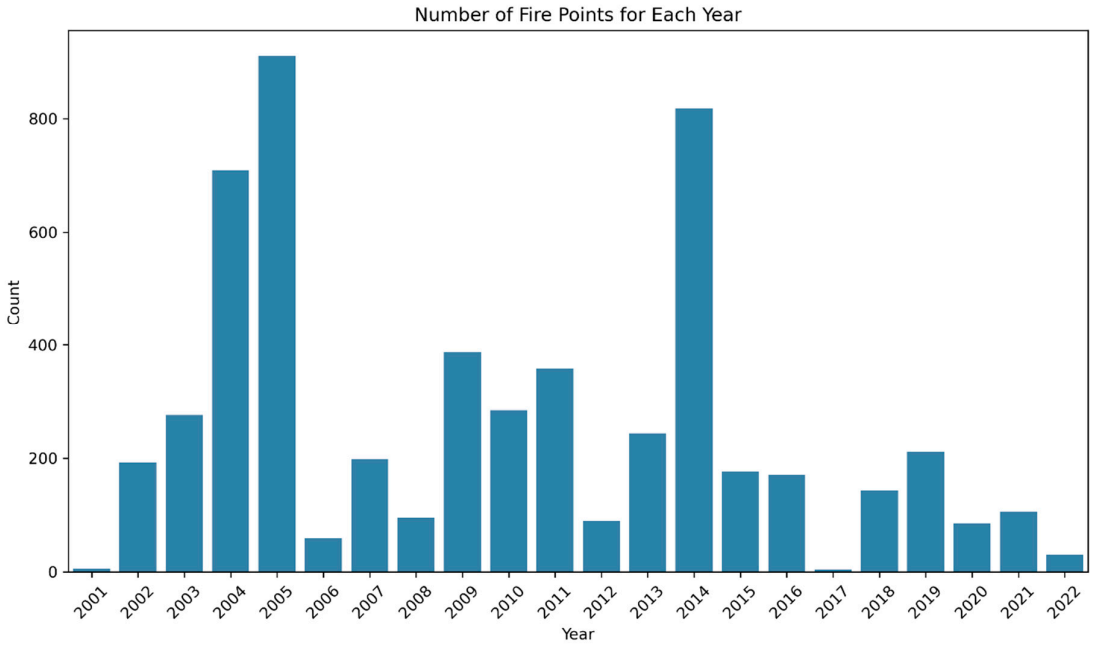

Figure 3. Number of Fire Points for Each Year from 2001 to 2022.

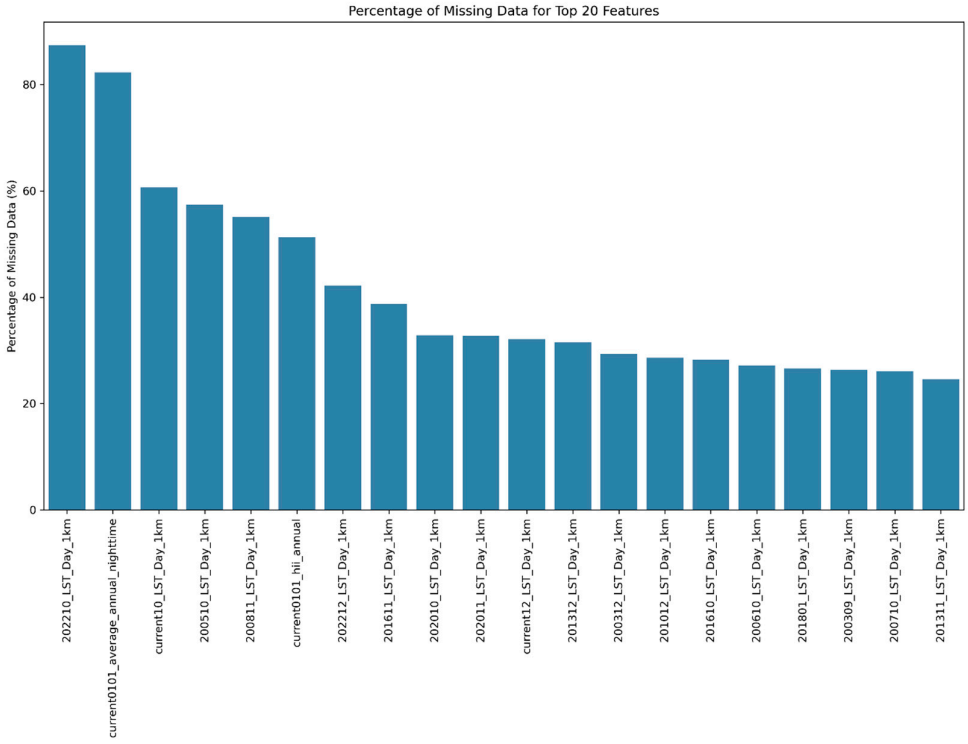

Figure 4. Percentage of Missing Data for Top 20 Features from Full Forest Fire Dataset.

 doi:10.20944/preprints202404.0027.v1

5. Assessing Forest Fire Dataset Leveraging Large Language Model

5.1. ChatGPT (GPT-4) and Noteable Plugin

The primary objective of this subsection is to evaluate the suitability of the forest fire dataset for fire behavioral analysis. It's important to emphasize that this assessment does not encompass a comprehensive analysis of the dataset. Instead, our focus is on ensuring that the dataset contains the necessary information and variables needed for in-depth analysis, paving the way for future research. To streamline the evaluation process, the dataset has been filtered to include only all annual key features. It's worth mentioning that the dataset filtering process specifically targeted variables containing the keyword 'annual,' resulting in the inclusion of only dynamic variables. Static variables such as the ESA landcover class, elevation, slope, and aspect are excluded from this analysis. This preprocessing is conducted locally to filter the dataset. This section intends to determine whether the generated forest fire dataset can provide valuable insights into the behavior of fires in Peninsular Malaysia.

For the analysis phase, the significant rise in the popularity of LLMs, such as ChatGPT [15], has created new opportunities for exploring and leveraging their capabilities. In this study, we employ GPT-4 in conjunction with the Noteable Plugin to conduct our sample analysis in the forthcoming discussions. Although Microsoft and Meta have recently released an open-source LLM called Llama 2 [64], it is not utilized in this work due to the unavailability of corresponding plugins. Despite GPT-4 being a closed-source model, it is employed in this study as the available plugin is exclusive to this platform. It should be emphasized that our primary focus is on the analysis, not the specific LLM model adopted.

Noteable, originally designed as a collaborative notebook platform, facilitates team data utilization and visualization through its secure cloud-based deployment, no-code visualizations, and expertly designed collaborative features, offering a unified data workspace for businesses. With the advent of ChatGPT and its increasing adoption, a new plugin has emerged, Noteable Plugin [16] with GPT-4, extending the platform's functionalities, enabling the creation of notebooks that encompass exploratory analysis, data visualization, machine learning, and data manipulation through natural language prompts. One key feature of this plugin is its ability to generate all the Python scripts utilized for the analysis, encompassing figures, charts, tables, and more within the Noteable platform. This approach promotes open data science by providing not only the results but also the Python scripts used to generate them [13]. Consequently, researchers and analysts can access these scripts to reproduce the same results, enhancing source code reusability.

In the forthcoming discussions, we delve into the analysis conducted through GPT-4 with the Noteable Plugin. It's essential to acknowledge that while this tool offers valuable insights, it may not be without minor errors. These imperfections are intentionally retained to emphasize that the plugin, while powerful, is not flawless and may require further refinement in the generated Python scripts for improved results. Nonetheless, it remains a valuable addition to a researcher's or data analyst's toolkit, offering a swift method for conducting preliminary analyses. To enhance transparency, the prompt history through ChatGPT via https://chat.openai.com that is used to trigger the plugin is made accessible and can be found at https://github.com/chewyeejian/GEE_FrameworkForestFireDataset. Additionally, the scripts generated by the plugin through the prompts are also provided at the same GitHub repository. Hence, the second contribution of this study is to evaluate the suitability of ChatGPT with the Noteable plugin as a tool for analysis, highlighting its potential and limitations.

5.2. Termination of the Noteable

The Noteable plugin was announced to be discontinued in December 2023, though no specific reasons were provided for this decision [65]. Despite its discontinuation, we postulate that the findings detailed within this paper might significantly encourage the future development and adoption of similar tools within the academic research domain. It is important to note that, while direct access to active notebooks on the Noteable platform has ceased, an archived copy of the

 doi:10.20944/preprints202404.0027.v1

notebook has been preserved and is accessible through our GitHub repository, https://github.com/chewyeejian/GEE_FrameworkForestFireDataset. In light of this, for those seeking alternatives, there are several Chatgpt-4 plugins available that offer coding assistance, including Code Interpreter [66] and Code Copilot [67].

5.3. Sample Analysis of Forest Fire Dataset in Peninsular Malaysia through GPT-4

In this subsection, the forest fire dataset generated earlier for Peninsular Malaysia is explored and analyzed to assess its usability. As previously mentioned, the dataset has been locally filtered to encompass only the annual average features to facilitate a simplified analysis. This analysis is carried out solely through ChatGPT prompts, which trigger the Noteable environment to generate the results and analysis. As mentioned in Section 5.1, the corresponding prompt history and the associated Noteable notebook are available at GitHub repository shared earlier. This analysis aims to offer an initial glimpse into the potential utility of the forest fire dataset for understanding fire behavior in Peninsular Malaysia.

To begin the analysis, GPT-4 with the Noteable plugin is employed to establish a connection between ChatGPT prompts and the Noteable platform. After the successful establishment of the linkage, ChatGPT is directed through prompts to execute various operations on the dataset. For instance, it is instructed to first analyze the dataset, determining the number of rows, the total number of columns, and listing all the column names. This reveals that the dataset comprises a total of 11,083 rows, aligning with the total number of rows in the full dataset. However, after the local filtering process to include only the annual average features, only 40 columns remain available, as detailed in Table 2. To refine the dataset for fire-focused analysis, a selection of columns that are irrelevant to the analytical objectives is removed. This includes the elimination of columns such as date/shape categories inherited from MCD64A1 (i.e., *system:index*, *Shape_Leng Shape_Area, date, year*, *month*, *day*, *validOn*), various administrative boundary categories (i.e., ADM0_EN, ADM1_EN, ADM2_EN, *ADM0_PCODE*, *ADM1_PCODE*, *ADM2_PCODE*), and others (i.e., *longitude*, *latitude*).

Following the data filtering process, an assessment to discover the missing features is carried out, akin to the analysis conducted on the full dataset. As depicted in Figure 5, variables such as nighttime light, HII, and KBDI exhibit a high percentage of missing data. This trend aligns with the observations made in the analysis of the full dataset and underscores the temporal availability of the data, which does not extend to 2022. This limitation arises from the chosen reference date for non-fire points. For the LST, the LST annual average does not appear to present an issue. However, the high percentage of missing data in the full dataset (Figure 4) may be attributed to the missing of monthly values. It is essential to emphasize that the primary objective in this analysis is to showcase the tool's effectiveness in evaluating the dataset. Therefore, this preliminary examination does not include any methods for replacing missing values.

Table 2. Features Information in Filtered Forest Fire Dataset.

Feature Name	Description	Feature Name	Description
system:index	System-generated from MCD64A1	current_aet_annual	Actual Evapotranspiration
longitude	Longitude Coordinate of Fire Points	current_def_annual	Climate water deficit
latitude	Latitude Coordinate of Fire Points	current_pdsi_annual	Palmer Drought Severity Index
fire	Fire Occurrence (binary class)	current_pet_annual	Reference Evapotranspiration

 doi:10.20944/preprints202404.0027.v1

date	Date from Administrative Boundaries refer to the Shape	current_pr_annual	Precipitation Accumulation
ADM1_PCODE	Administrative level 1 code	current_ro_annual	Runoff
ADM2_PCODE	Administrative level 2 code	current_soil_annual	Soil Moisture
Shape_Leng	Shape Length (from MCD64A1)	current_srad_annual	Downward Surface Shortwave Radiation
ADM0_EN	Country Name	current_swe_annual	Snow Water Equivalent
ADM1_EN	Administrative level 1 name	current_tmmn_annual	Minimum Temperature
ADM2_EN	Administrative level 2 name	current_tmmx_annual	Maximum Temperature
validOn	Validation Date from Administrative Boundaries refer to the Shape	current_vap_annual	Vapor Pressure
Shape_Area	Shape area (from MCD64A1)	current_vpd_annual	Vapor Pressure Deficit
ADM0_PCODE	Country code	current_vs_annual	Wind Speed at 10m
BurnDate	Date in 0-365 (from MCD64A1)	current_EVI_annual	Enhanced Vegetation Index
year	Year of Fire Observation	current_NDVI_annual	Normalized Difference Vegetation Index
month	Month of Fire Observation	current_LST_annual	Land Surface Temperature
day	Day of Fire Observation	current_KBDI_annual	Keetch-Byram Drought Index.
current0101_hii_annual	Human Impact Index	current0101_LC_Type2_annual	Land Cover Classification of UMD (Numeric)
current0101_average_annual_nighttime	Nighttime Brightness	current0101_LC_Type2_annual_classname	Land Cover Classification of UMD (Classname)

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2024 doi:10.20944/preprints202404.0027.v1

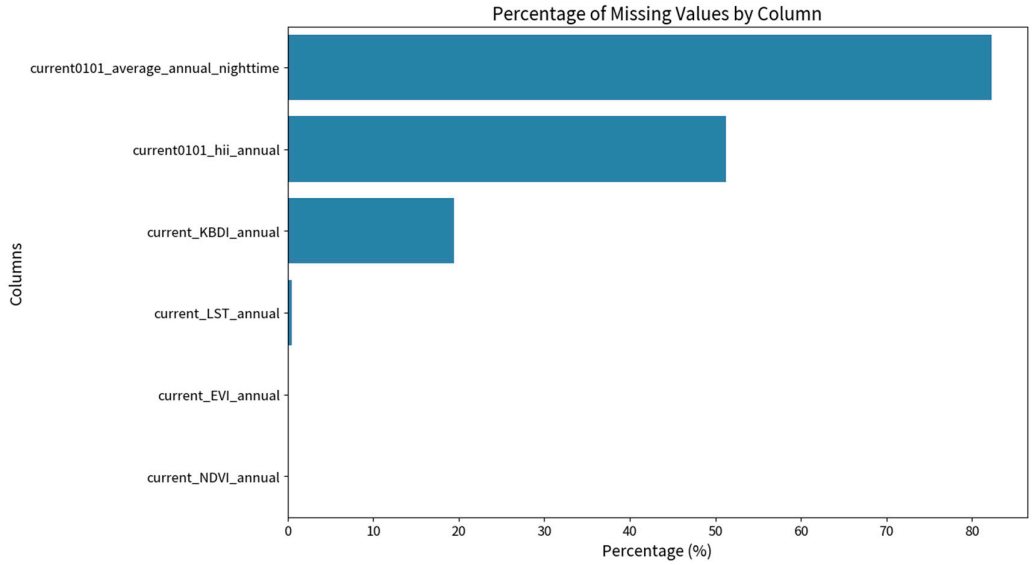

Figure 5. Percentage of Missing Data in the Filtered Forest Fire Dataset.

To provide further insights into the dataset, Table 3 presents the statistical mean and standard deviation of key features in relation to the fire class. Several noteworthy observations emerge from this table. First, the total count of average annual values observed for *current0101_average_annual_nighttime* and *current0101_hii_annual* is 0 for the non-fire category, aligning with our earlier discussions regarding the reference date for non-fire scenarios. Additionally, the '*current_swe_annual*' (snow water equivalent) feature remains at 0 for both fire and non-fire classes, which is reasonable given the absence of snowfall in Malaysia throughout the entire year. From the table, higher mean values are observed for KBDI, LST, AET, DEF, PET, and VPD for fire conditions, indicating drier conditions. Lower values of PDSI, PR, and RO also denote drier conditions. On the other hand, higher values of SRAD, TMMN, TMMX, and VS suggest more favorable conditions for fire incidents.

Table 3. Statistical Mean and Standard Deviation of the Key Features against Fire Class.

Features	Fire = 1			Non-Fire (Fire = 0)		
	Count	Mean	Standard Deviation	Count	Mean	Standard Deviation
current0101_LC_Type2_annual	5557	9.613101	3.652621	5526	3.088853	2.465596
current0101_average_annual_nighttime	1960	1.33929	2.232185	0	-	-
current0101_hii_annual	5403	2094.723	881.2572	0	-	-
current_EVI_annual	5553	0.369616	0.059076	5526	0.460531	0.064722
current_KBDI_annual	3404	95.86203	55.7931	5526	21.63871	9.450237
current_LST_annual	5522	304.4709	1.561426	5510	299.5443	2.200019
current_NDVI_annual	5553	0.567431	0.078325	5526	0.684663	0.075106
current_aet_annual	5557	104.8773	9.327677	5526	96.14751	8.020696
current_def_annual	5557	7.798602	7.765711	5526	0.57292	0.925342
current_pdsi_annual	5557	0.453241	2.904917	5526	3.93935	1.623397
current_pet_annual	5557	112.6757	12.85509	5526	96.72054	8.350228
current_pr_annual	5557	202.7966	36.59999	5526	269.2244	48.8007

 doi:10.20944/preprints202404.0027.v1

current_ro_annual	5557	97.5236	41.03856	5526	173.054	52.46889
current_soil_annual	5557	95.14247	52.39215	5526	94.80768	27.17345
current_srad_annual	5557	185.4051	24.57128	5526	169.1053	7.917929
current_swe_annual	5557	0	0	5526	0	0
current_tmmn_annual	5557	23.47415	0.549855	5526	21.33283	2.059547
current_tmmx_annual	5557	32.31917	0.718615	5526	29.97259	2.362977
current_vap_annual	5557	3.065104	0.080479	5526	2.808721	0.217333
current_vpd_annual	5557	0.825758	0.098175	5526	0.623304	0.215493
current_vs_annual	5557	1.834698	0.287861	5526	1.327443	0.267973

5.3.1. Sample Analysis - Boxplot Analysis with GPT-4 and Noteable Plugin

Box plot analysis, also known as a whisker plot, which visually represents data distribution using a five-number summary: minimum, first quartile, median, third quartile, and maximum is presented in Figure 6. In non-fire conditions, the *BurnDate* appears insignificant due to the reference date being set to -1 with the year of 2022. However, in fire scenarios, the median typically falls around day 60, with the first-third quartile ranging from approximately day 40 to 150. This suggests that fires are generally prevalent from February to around May.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2024 doi:10.20944/preprints202404.0027.v1

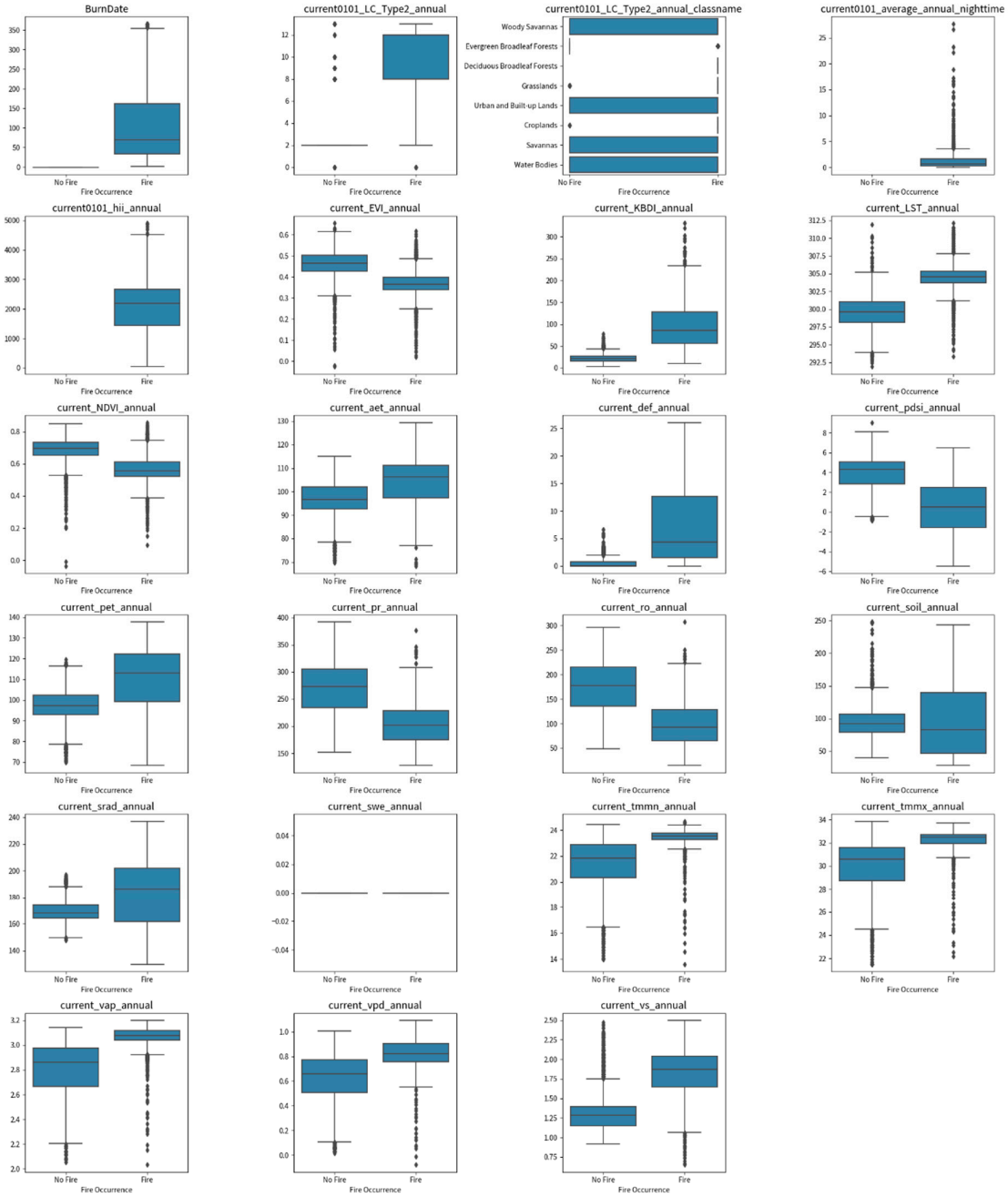

Figure 6. Boxplot Analysis for each Key Features for Fire and Non-Fire Points.

For the features current_KBDI_annual, current_LST_annual, current_aet_annual, current_def_annual, current_pdsi_annual, current_pet_annual, current_pr_annual, current_ro_annual, current_tmmn_annual, current_tmmx_annual, current_vap_annual, current_vpd_annual, and current_vs_annual, distinct medians are apparent when comparing fire scenario (i.e., fire=1) to non-fire scenario (i.e., fire=0). This disparity suggests that these features exhibit varying central tendencies in areas with fire. It is important to note that the black line inside the blue box in the diagram corresponds to the median.

The results obtained from the box plot analysis appear to be quite reasonable because most of the median values observed from the meteorological variables suggest conditions favorable for fire incidents. For instance, higher values of KBDI indicate drier conditions, elevated LST signifies higher temperatures, increased AET and PET values represent more significant water loss due to

 doi:10.20944/preprints202404.0027.v1

evaporation, lower PDSI and higher DEF (climate water deficit) values imply moisture deficits, lower PR (precipitation) levels indicate reduced moisture conditions, and decreased RO (runoff) values denote less water flowing from the land to the surface, signifying drier conditions or reduced water availability. In addition, higher values of SRAD indicate more solar energy reaching the Earth's surface, while the higher median temperature values (TMMN and TMMX) suggest a more favorable environment for fire occurrence. Elevated VPD values, associated with dry air, can promote the rapid drying of vegetation, while higher VS values may accelerate fire spread. Conversely, the lower median values for NDVI and EVI underscore the presence of less green vegetation at fire-affected sites. This analysis offers valuable insights into the interplay between various features and fire incidents.

5.2.2. Sample Analysis – T-tests Statistical Tests with GPT-4 and Noteable Plugin

T-tests statistical analysis are valuable tools for comparing the means of two groups and determining whether the observed differences between them are statistically significant. In this analysis, t-tests were exclusively applied to the numeric columns, and any columns containing missing data were thoughtfully excluded from the analysis. The t-test methodology involves comparing the means of two distinct groups and generating *p*-values, which quantitatively express the statistical significance of the observed differences. These *p*-values are subsequently compared to a predefined significance level, typically set at 0.05, to make informed decisions regarding the null hypothesis. The groups in question here are as follows: group 1 comprises variables associated with fire scenarios (identified by a "fire=1"), while group 2 encompasses variables linked to non-fire scenarios (designated by a "fire=0). The t-tests, as illustrated in Table 4, yielded both t-statistic values and corresponding *p*-values, aiding in the assessment of statistical significance within the dataset.

Table 4. t-test Statistics and *p*-value test.

Features	Group 1	Group 2	T-Statistics	p-value	Statistically Significant
BurnDate	Filter Fire Condition ('fire' = 1)	Filter Non-Fire Condition ('fire' =0)	-93.2815	0	**Significant**
current0101_LC_Type2_annual			-110.2646	0	
current_EVI_annual			77.2107	0	
current_KBDI_annual			-76.9397	0	
current_LST_annual			-135.6047	0	
current_NDVI_annual			80.4098	0	
current_aet_annual			-52.8367	0	
current_def_annual			-68.8715	0	
current_pdsi_annual			78.0405	0	
current_pet_annual			-77.5257	0	
current_pr_annual			81.0322	0	
current_ro_annual			84.3792	0	
current_srad_annual			-47.0552	0	
current_tmmn_annual			-74.6871	0	
current_tmmx_annual			-70.6441	0	
current_vap_annual			-82.2645	0	
current_vpd_annual			-63.5849	0	

 doi:10.20944/preprints202404.0027.v1

current_vs_annual			-96.0227	0	
current_soil_annual			-0.4226	0.6726	**Not Significant**
current0101_average_annual_nighttime			-	-	**NaN** (Missing Data)
current0101_hii_annual			-	-	**NaN** (Missing Data)
current_swe_annual			-	-	**NaN** (Constant)

In general, the magnitude of a t-statistic value serves as a measure of the difference between the sample mean and the hypothesized population mean. A larger t-statistic magnitude indicates a more substantial disparity between the sample mean and the hypothesized population mean, while a magnitude close to 0 suggests that the means of both groups are quite similar. Examining the table, we observe that t-statistic values for most key features exhibit a substantial magnitude, with the exception of *current_soil_annual*, which hovers closer to 0. Furthermore, the *p*-values generated for the majority of key variables, except for *current_soil_annual*, *current_0101_average_annual_nighttime*, *current0101_hii_annual*, and *current_swe_annual*, are extremely low, effectively reaching 0. This indicates that the differences observed between fire and non-fire conditions are indeed statistically significant for most key features. In the case of *current_soil_annual*, the t-statistic and *p*-value reveal their statistical insignificance. However, it should be noted that *current_0101_LC_Type2_annual* is not ideally suited for this analysis as it represents a categorical attribute converted to a numeric form, rendering it less relevant. Similarly, *BurnDate* should not be included as a key feature or predictor in this analysis; its inclusion is intentional, as these results were generated directly through the ChatGPT prompts.

5.2.3. Sample Analysis – Variance Inflation Factor with GPT-4 and Noteable Plugin

The Variance Inflation Factor (VIF) is a crucial metric used to identify multicollinearity in regression analysis. For this analysis, only numeric variables are taken into consideration, and any columns containing null variables are excluded. Multicollinearity refers to a scenario in which two or more predictor variables in a regression model exhibit high correlations. In Figure 7, the VIF analysis reveals that certain variables, such as current_pet_annual and current_aet_annual, exhibit very high VIF values, signifying their strong correlations with other predictor variables in the dataset. To address multicollinearity issues, it is generally recommended to either eliminate or combine features with high VIF values.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 1 April 2024 doi:10.20944/preprints202404.0027.v1

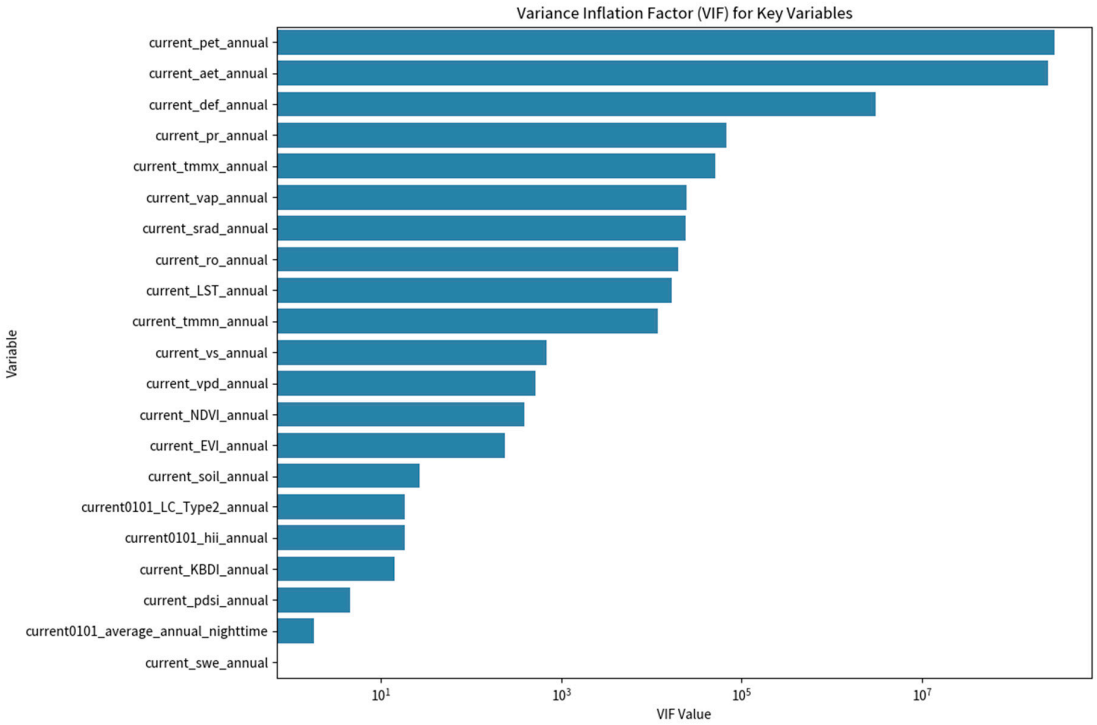

Figure 7. Variance Inflation Factor for Key Features.

5.4. Limitation of the GPT-4 with Noteable Plugin

In exploring the capabilities of the GPT-4 with Noteable Plugin, we observed a mix of successful analyses and encountered challenges with certain statistical tests and feature importance examinations. While some of the analyses, such as basic statistical information, boxplot analysis, t-test statistics, and VIF analysis, yielded valuable results as shown in the previous subsection, several other tests faced errors and inaccuracies. It is important to emphasize that the suggestions for various statistical and feature importance tests are generated by ChatGPT. This subsection delves into the tests and analyses that presented challenges and limitations, offering insights where further refinement may be necessary.

The first test encountering errors is the feature importance ranking utilizing the Random Forest machine learning model. This analysis aims to rank features based on their importance in predicting the fire class, offering insights into which features strongly influence fire occurrences. However, the prompts consistently detect only one class, preventing model training and the calculation of feature importance through Random Forest. A similar issue also arises when examining logistic regression coefficients. This anomaly could be attributed to the unaddressed missing data, which may require further attention. In addition, attempts to conduct multiple statistical tests in a single prompt using numerous suggestions from ChatGPT responses resulted in the conversation crashing. This issue occurred during multiple retries on logistic regression coefficients, recursive feature elimination, chi-squared tests, mutual information, and ANOVA (analysis of variance) tests. The crash may be attributed to exceeding a certain output limit. It is important to note that these errors or limitations may also be caused by the imperfect prompts used to instruct ChatGPT to perform specific tasks, highlighting the need for more precise prompts to ensure accurate execution. While the plugin has demonstrated its capability to generate graphs, charts, and analysis insights in Section 5.3 through multiple prompts, it is equally important for data users to have a solid grasp of the specific analysis or tests being conducted. This understanding is crucial for interpreting the analysis results effectively and deriving deeper insights from the generated outcomes.

Advocating for additional studies to validate the usage of the tools in their analysis is essential. While the limitations of the plugins are acknowledged, the results presented in Section 5.3 suggest

 doi:10.20944/preprints202404.0027.v1

that embracing these tools as a part of one's toolkit for preliminary analysis or assessment is warranted. However, as evidenced by the limitations, it remains essential to conduct additional checks to verify the accuracy and intended outcomes of the analyses. From our personal viewpoint, considering that most tests and procedures are proposed by plugin, it is evident that this tool can greatly benefit data analysts and researchers seeking guidance on which available tests to perform.

6. Conclusions

In conclusion, the primary contribution of this paper lies in the proposed framework, which includes the scripts for swiftly generating forest fire datasets from GEE. This methodology is easily replicable for various locations, and the resulting datasets can be freely shared without the need for permissions from government authorities or other organizations. Peninsular Malaysia served as the case study to showcase the effectiveness of the proposed framework. Since the generated dataset is created without the use of any private government or organizational data, it can be openly accessed and shared without restrictions. This framework greatly lowers the barriers for data scientists, enabling them to apply their analytical skills directly to the GEE-extracted datasets, reducing the necessity for in-depth remote sensing knowledge.

The second contribution of this work involves the successful adoption and demonstration of LLM, specifically GPT-4 with the Noteable plugin, as a tool for conducting preliminary analysis on the generated dataset. The sample analysis reveals valuable insights into the fire scenarios in Peninsular Malaysia. Key factors affecting forest fires in this region, based on the preliminary analysis of the annual averages referencing the year of fire, include KBDI, LST, PDSI, DEF, PR, RO, SRAD, TMMX, TMMN, and VPD. Section 5.4 discusses the limitations of GPT-4 with the Noteable plugin. It is important to note that no manual coding was performed during the analysis; rather, the analysis and Python scripts producing the results were generated through simple prompts within the ChatGPT interface. As technology continues to evolve, researchers at the forefront should consider adopting such technologies to enhance their methodologies and analyses.

7. Future Works

In this study, the proposed framework relies solely on the MCD64A1 BA dataset for extracting fire points in the study area. It's important to acknowledge that this dataset may not be entirely accurate and could miss some fire occurrences [68]. However, for the purposes of this study, it provides a rapid means to discover the historical fire location. For future work, improving the collection of historical fire points could involve incorporating other BA datasets such as FireCCI51 v5.1 [33], Globfire Fire Event [34], or region-specific government data. While the provided scripts demonstrate the feasibility of the proposed framework, creating a GUI version could enhance user convenience.

While the GPT-4 with Noteable plugin has effectively showcased its utility for analysis, it is essential to emphasize the need for subsequent validation to confirm the tool's applicability and limitations. The sample analysis conducted with the GPT-4 and Noteable plugin focused solely on the annual averages, specifically referencing the year of fire. However, it is crucial to advocate for a more comprehensive analysis that considers all variables, providing a deeper understanding of the critical factors influencing fires in Malaysia. Furthermore, the dataset can be employed to train a machine learning model as a predictive tool to forecast future fire occurrences. This predictive model would serve as a valuable resource for proactive fire management strategies.

The versatility of the proposed framework extends beyond forest fires. By supplying the geographical coordinates of various disastrous events, it becomes feasible to extract the pertinent remote sensing features from GEE for analyzing the occurrence of these disasters, using the same methodology applied in this study. It's important to underscore that a substantial amount of remote sensing data employed in this research is accessible through the GEE data catalog. However, it's worth noting that the GEE community catalog [69] offers access to a multitude of additional datasets, further broadening the scope of potential applications. Future work could explore the utilization of

 doi:10.20944/preprints202404.0027.v1

these diverse datasets for different types of environmental assessments and disaster management efforts.

Author Contributions: Conceptualization, Y.J.C.; Methodology, Y.J.C. and S.Y.O.; Formal analysis, Y.J.C.; Investigation, Y.J.C. and S.Y.O.; Visualization, Y.J.C. and Z.Y.L.; Writing – Original Draft, Y.J.C.; Funding Acquisition, S.Y.O.; Resources, S.Y.O.; Project Administration, S.Y.O.; Supervision, S.Y.O. and Y.H.P.; Writing – Review and Editing, S.Y.O. and Y.H.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research work was supported by a Fundamental Research Grant Schemes (FRGS) under the Ministry of Education and Multimedia University, Malaysia (Project ID: FRGS/1/2020/ICT02/MMU/02/2).

Data Availability Statement: The complete Python scripts for extracting fire factors through GEE, the filtered forest fire dataset containing only the annual variables, a copy of the Noteable Python scripts generated by the ChatGPT prompt, and the ChatGPT conversation history used to trigger the plugin for generating the analysis can be easily accessed at the following GitHub repository, https://github.com/chewyeejian/GEE_FrameworkForestFireDataset. The full unprocessed forest fire dataset for Peninsular Malaysia is available at https://doi.org/10.5281/zenodo.10050852 [48].

Acknowledgments: We also extend our heartfelt gratitude to Dr. Morgan Crowley for the valuable suggestions and advice provided on the proposed framework.

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

References

1. Arif, M.; Alghamdi, K.K.; Sahel, S.A.; Alosaimi, S.O.; Alsahaft, M.E.; Alharthi, M.A.; Arif, M. Role of Machine Learning Algorithms in Forest Fire Management: A Literature Review. *J Robot. Autom* **2021**, *5*, 212–226.
2. Bot, K.; Borges, J.G. A Systematic Review of Applications of Machine Learning Techniques for Wildfire Management Decision Support. *Inventions* **2022**, *7*, 15.
3. Moayedi, H.; Mehrabi, M.; Bui, D.T.; Pradhan, B.; Foong, L.K. Fuzzy-Metaheuristic Ensembles for Spatial Assessment of Forest Fire Susceptibility. *J. Environ. Manage.* **2020**, *260*, 109867.
4. Bui, D.T.; Bui, Q.-T.; Nguyen, Q.-P.; Pradhan, B.; Nampak, H.; Trinh, P.T. A Hybrid Artificial Intelligence Approach Using GIS-Based Neural-Fuzzy Inference System and Particle Swarm Optimization for Forest Fire Susceptibility Modeling at a Tropical Area. *Agric. For. Meteorol.* **2017**, *233*, 32–44.
5. Bui, D.T.; Hoang, N.-D.; Samui, P. Spatial Pattern Analysis and Prediction of Forest Fire Using New Machine Learning Approach of Multivariate Adaptive Regression Splines and Differential Flower Pollination Optimization: A Case Study at Lao Cai Province (Viet Nam). *J. Environ. Manage.* **2019**, *237*, 476–487.
6. Sevinc, V.; Kucuk, O.; Goltas, M. A Bayesian Network Model for Prediction and Analysis of Possible Forest Fire Causes. *For. Ecol. Manage.* **2020**, *457*, 117723.
7. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-Scale Geospatial Analysis for Everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27, doi:10.1016/j.rse.2017.06.031.
8. Ban, Y.; Zhang, P.; Nascetti, A.; Bevington, A.R.; Wulder, M.A. Near Real-Time Wildfire Progression Monitoring with Sentinel-1 SAR Time Series and Deep Learning. *Sci. Rep.* **2020**, *10*, 1322, doi:10.1038/s41598-019-56967-x.
9. Hodges, J.L.; Lattimer, B.Y. Wildland Fire Spread Modeling Using Convolutional Neural Networks. *Fire Technol.* **2019**, *55*, 2115–2142, doi:10.1007/s10694-019-00846-4.
10. Zhang, G.; Wang, M.; Liu, K. Forest Fire Susceptibility Modeling Using a Convolutional Neural Network for Yunnan Province of China. *Int. J. Disaster Risk Sci.* **2019**, *10*, 386–403, doi:10.1007/s13753-019-00233-1.
11. Jiao, Z.; Zhang, Y.; Xin, J.; Mu, L.; Yi, Y.; Liu, H.; Liu, D. A Deep Learning Based Forest Fire Detection Approach Using UAV and YOLOv3. In Proceedings of the 2019 1st International Conference on Industrial Artificial Intelligence (IAI); IEEE: Shenyang, China, July 2019; pp. 1–5.
12. Wang, S.; Zhao, J.; Ta, N.; Zhao, X.; Xiao, M.; Wei, H. A Real-Time Deep Learning Forest Fire Monitoring Algorithm Based on an Improved Pruned + KD Model. *J. Real-Time Image Process.* **2021**, *18*, 2319–2329, doi:10.1007/s11554-021-01124-9.
13. Gomes, V.C.F.; Queiroz, G.R.; Ferreira, K.R. An Overview of Platforms for Big Earth Observation Data Management and Analysis. *Remote Sens.* **2020**, *12*, 1253.
14. Wu, Q. Geemap: A Python Package for Interactive Mapping with Google Earth Engine. *J. Open Source Softw.* **2020**, *5*, 2305.
15. OpenAI ChatGPT [Large Language Model] Available online: https://chat.openai.com/chat (accessed on 21 October 2023).

16. Noteable Noteable (ChatGPT Plugin for Notebook) Available online: https://noteable.io/chatgpt-plugin-for-notebook/ (accessed on 21 October 2023).
17. Chicas, S.D.; Østergaard Nielsen, J. Who Are the Actors and What Are the Factors That Are Used in Models to Map Forest Fire Susceptibility? A Systematic Review. *Nat. Hazards* **2022**, *114*, 2417–2434, doi:10.1007/s11069-022-05495-5.
18. Pradhan, B.; Suliman, M.D.H. Bin; Awang, M.A. Bin Forest Fire Susceptibility and Risk Mapping Using Remote Sensing and Geographical Information Systems (GIS). *Disaster Prev. Manag.* **2007**, *16*, 344–352.
19. Pu, R.; Li, Z.; Gong, P.; Csiszar, I.; Fraser, R.; Hao, W.-M.; Kondragunta, S.; Weng, F. Development and Analysis of a 12-Year Daily 1-Km Forest Fire Dataset across North America from NOAA/AVHRR Data. *Remote Sens. Environ.* **2007**, *108*, 198–208.
20. Lestari, A.; Rumantir, G.; Tapper, N. A Spatio-Temporal Analysis on the Forest Fire Occurrence in Central Kalimantan, Indonesia. In Proceedings of the Proceeding of the 20th Pacific Asia Conference on Information Systems; Chiayi, Taiwan, 2016; p. 90.
21. Monjarás-Vega, N.A.; Briones-Herrera, C.I.; Vega-Nieva, D.J.; Calleros-Flores, E.; Corral-Rivas, J.J.; López-Serrano, P.M.; Pompa-García, M.; Rodríguez-Trejo, D.A.; Carrillo-Parra, A.; González-Cabán, A. Predicting Forest Fire Kernel Density at Multiple Scales with Geographically Weighted Regression in Mexico. *Sci. Total Environ.* **2020**, *718*, 137313.
22. Abid, F. A Survey of Machine Learning Algorithms Based Forest Fires Prediction and Detection Systems. *Fire Technol.* **2021**, *57*, 559–590, doi:10.1007/s10694-020-01056-z.
23. Esri Introducing ArcGIS Platform | Esri Available online: https://www.esri.com/en-us/home (accessed on 13 March 2021).
24. QGIS Development Team Welcome to the QGIS Project! Available online: https://www.qgis.org/en/site/ (accessed on 13 March 2021).
25. Sulova, A.; Jokar Arsanjani, J. Exploratory Analysis of Driving Force of Wildfires in Australia: An Application of Machine Learning within Google Earth Engine. *Remote Sens.* **2021**, *13*, 10.
26. Sulova, A.; Jokar Arsanjani, J. Github Code: Exploratory Analysis of Wildfires in Australia and Approach for Wildfire Modeling in Google Earth Engine Available online: https://github.com/sulova/AustraliaFires (accessed on 3 October 2023).
27. Velastegui-Montoya, A.; Montalván-Burbano, N.; Carrión-Mero, P.; Rivera-Torres, H.; Sadeck, L.; Adami, M. Google Earth Engine: A Global Analysis and Future Trends. *Remote Sens.* **2023**, *15*, 3675, doi:10.3390/rs15143675.
28. Pham-Duc, B.; Nguyen, H.; Phan, H.; Tran-Anh, Q. Trends and Applications of Google Earth Engine in Remote Sensing and Earth Science Research: A Bibliometric Analysis Using Scopus Database. *Earth Sci. Informatics* **2023**, *16*, 2355–2371, doi:10.1007/s12145-023-01035-2.
29. Pérez-Cutillas, P.; Pérez-Navarro, A.; Conesa-García, C.; Zema, D.A.; Amado-Álvarez, J.P. What Is Going on within Google Earth Engine? A Systematic Review and Meta-Analysis. *Remote Sens. Appl. Soc. Environ.* **2023**, *29*, 100907.
30. Yang, L.; Driscol, J.; Sarigai, S.; Wu, Q.; Chen, H.; Lippitt, C.D. Google Earth Engine and Artificial Intelligence (AI): A Comprehensive Review. *Remote Sens.* **2022**, *14*, doi:10.3390/rs14143253.
31. Chen, H.; Yang, L.; Wu, Q. Enhancing Land Cover Mapping and Monitoring: An Interactive and Explainable Machine Learning Approach Using Google Earth Engine. *Remote Sens.* **2023**, *15*, 4585, doi:10.3390/rs15184585.
32. Giglio, L.; Justice, C.; Boschetti, L.; Roy, D. MODIS/Terra+Aqua Burned Area Monthly L3 Global 500m SIN Grid V061 [Data Set] Available online: https://doi.org/10.5067/MODIS/MCD64A1.061 (accessed on 1 March 2023).
33. Lizundia-Loiola, J.; Otón, G.; Ramo, R.; Chuvieco, E. A Spatio-Temporal Active-Fire Clustering Approach for Global Burned Area Mapping at 250 m from MODIS Data. *Remote Sens. Environ.* **2020**, *236*, 111493.
34. Artés, T.; Oom, D.; De Rigo, D.; Durrant, T.H.; Maianti, P.; Libertà, G.; San-Miguel-Ayanz, J. A Global Wildfire Dataset for the Analysis of Fire Regimes and Fire Behaviour. *Sci. data* **2019**, *6*, 296.
35. Chew, Y.J.; Ooi, S.Y.; Pang, Y.H. MCD64A1 Burnt Area Dataset Assessment Using Sentinel-2 and Landsat-8 on Google Earth Engine: A Case Study in Rompin, Pahang in Malaysia. In Proceedings of the 2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE); 2023; pp. 38–43.
36. Malaysia Kini Hutan Seluas 34 Hektar Terbakar Di Kuantan (A 34-Hectare Forest Burned in Kuantan) Available online: https://www.malaysiakini.com/news/339616 (accessed on 2 August 2021).
37. Astro Awani Kebakaran Hutan Simpan Pekan Tak Membimbangkan (Fire in Pekan Forest Reserve Is Not a Concern) Available online: https://www.astroawani.com/berita-malaysia/kebakaran-hutan-simpan-pekan-tak-membimbangkan-186979 (accessed on 2 August 2021).
38. Bernama Kebakaran Hutan Simpan Pekan: Anggota Bomba, Jabatan Perhutanan Terkandas (Fire in Pekan Forest Reserve: Fire Fighters, Forestry Department Is Stranded) Available online: https://www.utusanborneo.com.my/2018/10/01/kebakaran-hutan-simpan-pekan-anggota-bomba-jabatan-perhutanan-terkandas (accessed on 2 August 2021).

39. Alagesh, T.N. 40ha of Pahang Forest, Peat Land on Fire. *New Straits Times* 2019.
40. Bernama 80 Hektar Hutan Simpan Kuala Langat Terbakar Available online: https://www.bharian.com.my/berita/kes/2020/04/679541/80-hektar-hutan-simpan-kuala-langat-terbakar (accessed on 2 August 2021).
41. Idris, M.N. Kebakaran Hutan Di Selangor Meningkat - Utusan Digital Available online: https://www.utusan.com.my/berita/2020/07/kebakaran-hutan-di-selangor-meningkat/ (accessed on 2 August 2021).
42. Malaymail Kuala Langat Selatan Forest Fire Spreads to over 40 Hectares, Says Selangor Fire Dept 2021.
43. Bernama Large Forest Fire Raging in Perak; Orang Asli Settlement Threatened. *News Straits Times* 2019.
44. Haralick, R.M.; Sternberg, S.R.; Zhuang, X. Image Analysis Using Mathematical Morphology. *IEEE Trans. Pattern Anal. Mach. Intell.* **1987**, *PAMI-9*, 532–550, doi:10.1109/TPAMI.1987.4767941.
45. Zeng, A.-C.; Cai, Q.-J.; Su, Z.; Guo, X.-B.; Jin, Q.-F.; Guo, F.-T. Seasonal Variation and Driving Factors of Forest Fire in Zhejiang Province, China, Based on MODIS Satellite Hot Spots. *Chinese J. Appl. Ecol.* **2020**, *31*, 399–406.
46. Zhu, Z.; Deng, X.; Zhao, F.; Li, S.; Wang, L. How Environmental Factors Affect Forest Fire Occurrence in Yunnan Forest Region. *Forests* 2022, *13*.
47. Parisien, M.A.; Parks, S.A.; Krawchuk, M.A.; Little, J.M.; Flannigan, M.D.; Gowman, L.M.; Moritz, M.A. An Analysis of Controls on Fire Activity in Boreal Canada: Comparing Models Built with Different Temporal Resolutions. *Ecol. Appl.* **2014**, *24*, 1341–1356, doi:10.1890/13-1477.1.
48. Chew, Y.J. Forest Fire Dataset for Peninsular Malaysia (2001-2022) Extracted from Multiple-Source Remote Sensing Data Using Google Earth Engine [Data Set]. *Zenodo* 2023.
49. Wu, Q.; Lane, C.R.; Li, X.; Zhao, K.; Zhou, Y.; Clinton, N.; DeVries, B.; Golden, H.E.; Lang, M.W. Integrating LiDAR Data and Multi-Temporal Aerial Imagery to Map Wetland Inundation Dynamics Using Google Earth Engine. *Remote Sens. Environ.* **2019**, *228*, 1–13, doi:10.1016/j.rse.2019.04.015.
50. Sanderson, E.W.; Fisher, K.; Robinson, N.; Sampson, D.; Duncan, A.; Royte, L. The March of the Human Footprint. **2022**.
51. Abatzoglou, J.T.; Dobrowski, S.Z.; Parks, S.A.; Hegewisch, K.C. TerraClimate, a High-Resolution Global Dataset of Monthly Climate and Climatic Water Balance from 1958-2015. *Sci. Data* **2018**, *5*, 1–12, doi:10.1038/sdata.2017.191.
52. Takeuchi, W.; Darmawan, S.; Shofiyati, R.; Khiem, M. Van; Oo, K.S.; Pimple, U.; Heng, S. Near-Real Time Meteorological Drought Monitoring and Early Warning System for Croplands in Asia. In Proceedings of the Asian Conference on Remote Sensing 2015: Fostering Resilient Growth in Asia; 2015; Vol. 1, pp. 171–178.
53. Wan, Z.; Hook, S.; Hulley, G. MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V061 [Dataset]. NASA EOSDIS Land Processes DAAC; Accessed 2023-02-07 from https://doi. org/10.5067/MODIS/MOD11A2. 061, 2021;
54. Didan, K. MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V061 [Data Set]. NASA EOSDIS Land Processes DAAC; Accessed 2022-12-23 from https://doi. org/10.5067/MODIS/MOD13Q1. 061, 2021;
55. Friedl, M.; Sulla-Menashe, D. MODIS/Terra+ Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V061. *NASA EOSDIS L. Process. DAAC Sioux Falls, SD, USA* **2022**.
56. Zanaga, D.; Van De Kerchove, R.; Daems, D.; De Keersmaecker, W.; Brockmann, C.; Kirches, G.; Wevers, J.; Cartus, O.; Santoro, M.; Fritz, S.; et al. ESA WorldCover 10 m 2021 V200 2022.
57. NASA JPL NASADEM Merged DEM Global 1 Arc Second V001 [Data Set].
58. Marconcini, M.; Metz-Marconcini, A.; Üreyen, S.; Palacios-Lopez, D.; Hanke, W.; Bachofer, F.; Zeidler, J.; Esch, T.; Gorelick, N.; Kakarla, A. Outlining Where Humans Live, the World Settlement Footprint 2015. *Sci. Data* **2020**, *7*, 242.
59. Elvidge, C.D.; Zhizhin, M.; Ghosh, T.; Hsu, F.-C.; Taneja, J. Annual Time Series of Global VIIRS Nighttime Lights Derived from Monthly Averages: 2012 to 2019. *Remote Sens.* **2021**, *13*, 922.
60. Chew, Y.J.; Ooi, S.Y.; Pang, Y.H.; Wong, K.-S. A Review of Forest Fire Combating Efforts, Challenges and Future Directions in Peninsular Malaysia, Sabah, and Sarawak. *Forests* **2022**, *13*, 1405, doi:10.3390/f13091405.
61. Chew, Y.J.; Ooi, S.Y.; Pang, Y.H. Data Acquisition Guide for Forest Fire Risk Modelling in Malaysia. *2021 9th Int. Conf. Inf. Commun. Technol. ICoICT 2021* **2021**, 633–638, doi:10.1109/ICoICT52021.2021.9527495.
62. OCHA Malaysia - Subnational Administrative Boundaries Available online: https://data.humdata.org/dataset/cod-ab-mys (accessed on 20 October 2023).
63. Chew, Y.J.; Ooi, S.Y.; Pang, Y.H.; Wong, K.S. Trend Analysis of Forest Fire in Pahang, Malaysia from 2001-2021 with Google Earth Engine Platform. *J. Logist. Informatics Serv. Sci.* **2022**, *9*, 15–26, doi:10.33168/LISS.2022.0402.
64. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv Prepr. arXiv2307.09288* **2023**.

 doi:10.20944/preprints202404.0027.v1

65. Zhipa, J. Noteable's Unexpected Farewell Available online: https://medium.com/@julia.zhipa/noteables-unexpected-farewell-b18a312346b4 (accessed on 29 March 2024).
66. Paliwal, A. Code Interpreter Available online: https://gptstore.ai/gpts/R3QCSpFIgF-code-interpreter (accessed on 29 March 2024).
67. Promptspellsmith Code Copilot Available online: https://gptstore.ai/gpts/_0GbIdtVSh-codecopilot (accessed on 29 March 2024).
68. Zhu, C.; Kobayashi, H.; Kanaya, Y.; Saito, M. Size-Dependent Validation of MODIS MCD64A1 Burned Area over Six Vegetation Types in Boreal Eurasia: Large Underestimation in Croplands. *Sci. Rep.* **2017**, *7*, 1–9, doi:10.1038/s41598-017-03739-0.
69. Roy, S.; Schwehr, K.; Pasquarella, V.; Trochim, E.; Swetnam, T. Samapriya/Awesome-Gee-Community-Datasets: Community Catalog 2023.