Exploratory analysis of wildfires in Australia and a machine learning approach for wildfire modeling in Google Earth Engine

Aalborg University Copenhagen

MSc Geoinformatics, Institute of Planning

AALBORG UNIVERSITY
STUDENT REPORT

Master Thesis

4th June 2020

Andrea **S**ulova

Abstract

Recent studies suggest that due to climate change the number of wildfires across the globe will be increasing. Recently, massive wildfires hit Australia during the 2019-2020 summer season where 46 million acres of land burnt. This fire disaster is raising questions to what extent the risk of wildfires can be linked to various climate, environmental, topographical, and social factors and how to predict fire occurrence to take preventive measures. This study investigates the Australian wildfires-based on free remotely sensed data from Earth observation to uncover the general insights. In the last few years, machine learning (ML) has demonstrated to be successful in many domains due to its capability of learning from obvious but also hidden relationships. One of the aims of this study is to create an automatized process of creating a fire training dataset at a continental level with an efficient computational expense for the ML algorithms. These results of fire occurrence and no-fire occurrence locations are mapped alongside with fire causal factors. The training dataset is applied to different ML algorithms, such as Random Forest (RF), Naïve Bayes (NB), and Classification and Regression Tree (CART). The ML algorithm with the best performance, the RF model, is used to identify the driving factors using variable importance analysis. Typically, a model can learn certain properties from a training dataset to make predictions. Thus, the overall objective of this study is to disclose the fire occurrence probability across Australia as well as identify the driving factors of wildfires applying the fire occurrence dataset from the 2019-2020 summer season. Improved preventive measures can be implemented in the fire-prone areas to reduce the risk of wildfires in Australia by considering the identified factors.

Keywords: remote sensing, wildfires, Australia, fire severity, random forest, machine learning

Author: Andrea Sulova
Supervisor: Prof. Dr. Jamal Jokar Arsanjani
Education: MSc in Geoinformatics, Master's programme
University: Aalborg University Copenhagen, Denmark

Preface

This master's thesis summarizes my last semester of the master's programme in Geoinformatics at Aalborg University. The main goals of the master's thesis are to uncover the features of Australian wildfires by applying remote sensing data, to identify the driving factors associated with wildfire events during the 2019-2020 season and predict the wildfire locations.

The inspiration for the master's thesis originates from both passion to gain more knowledge related to remote sensing and eagerness to take action against climate change, as one of its direct consequences are wildfires.

I would like to express my gratitude to my supervisor, Prof. Dr. Jamal Jokar Arsanjani, for the exceptional cooperation and his valuable support throughout the entire thesis. Additionally, my deepest gratitude belongs to my boyfriend and my sisters for their constant support and continuous encouragement throughout my years of studies.

I hope this study and its results bring new insights into the wildfires and will be valuable for future research. The JavaScript codes can be obtained from the GitHub repository *https://github.com/sulova/AustraliaFires.*

List of Figures

List of Tables

Definitions and acronyms

Name	Acronym
Application Programming Interface	API
Classification and Regression Tree	CART
Copernicus Global Land Service	CGLS
Convolutional Neural Network	CNN
Cascading Style Sheets	CSS
Digital Elevation Model	DEM
Difference Normal Burn Ratio	dNBR
European Centre for Medium-Range Weather Forecasts	ECMWF
European Centre for Medium-Range Weather Forecast Reanalysis	ERA5
European Space Agency	ESA
The Fire Information for Resource Management System	FIRMS
Google Erath Engine	GEE
Geographical Information System	GIS
Hypertext Mark-up Language	HTML
JavaScript	JS
Middle Infrared	MIR
Machine Learning	ML
Naïve Bayes	NB
Normal Burn Ratio	NBR
Normalized Difference Vegetation Index	NDVI
Normalized Difference Water Index	NDWI
Near-infrared	NIR
National Oceanic and Atmospheric Administration-Advanced Very High-Resolution Radiometer	NOAA-AVHRR
Open Street Map	OSM
Random Forest	RF
Synthetic Aperture Radar	SAR
Shuttle Radar Topography Mission	SRTM
Shortwave Infrared	SWIR
Thermal Infrared Range	TIR

Table of contents

1. Introduction

Australia has been seriously affected by the fire events known as "Black Summer" during the 2019-2020 summer season [1]. At least 46 million acres of land have burnt [2] and "fires near me" has become Google's most searched words in Australia during that fire season [3]. This fire disaster is raising a question to what extent the risk of wildfires can be linked to various climate, environmental and social factors.

Nowadays, wildfire disaster risks are being heightened globally due to climate changes. High temperatures and prolonged dry seasons might result in unprecedented bushfire activity across Australia. The state temperature dataset, originating in 1910, reveals that Australia's warmest year on record was in 2019, with the annual national mean temperature 1.52 °C above average. The dataset also shows the rainfall level was below average in all the capital cities across the 2019-2020 season. Australia's climate in 2019 was the driest year on record driven by record excursions and significant heatwaves in January and December [4]. However, humans might also play a critical role in some wildfire events as the recent study shows in Spain [5]. In this study, most wildfires were most likely triggered by human activities as spatial patterns of wildfire ignition are strongly linked with human access to the natural landscape, with the proximity to urban areas and roads found to be the most important contributory factors.

The satellite remote sensing has become a common tool for large-scale area monitoring of ecosystems as well as spotting threats, e.g. wildfires, across the globe [6]. Multiple studies have been already conducted using remote sensing and applied various approaches, such as Kernel Logistic Regression or Spatial Logistic Regression.

However, recently, ML approaches have rapidly progressed and achieved promising results in the environmental sciences [7]. This led to the analyzes of the recent Australian fires with implementing different ML algorithms, namely, Naive Bayes (NB), Random Forest (RF), and Classification and Regression Trees (CART). This study directly compares ML methods for wildfire mapping, and subsequently, a method with the best achieved performance in both model training and validation is used for mapping the continental wildfire probability in Australia.

Moreover, this thesis aims to evaluate a set of causal variables, i.e., predictor variables, and to identify the dominant factors behind the recent wildfires in Australia.

Modelling many complex environmental and socio-economic independent variables is often a difficult task due to large resource requirements, i.e., complexity as well as heterogeneous data formats. In that respect, most predictor variables, e.g., temperature, precipitation, population, etc., are gathered from the Google Earth Engine (GEE) data catalogue.

A training dataset in ML algorithms is an essential input supporting the model's ability to learn [8]. The process of generating a training dataset for supervised learning is frequently manual. Due to the extensive area and wide time frame of the fire season, it is crucial to create an automated process for generating the most representative set of data for the model training. Therefore, this thesis proposes an extensive automated framework for generating the large training dataset across entire Australia.

1.1 Problem statement and research questions

This study aims to use ML algorithms for predictions of wildfire susceptibility based on wildfires in Australia in the 2019-2020 season and determine potential causal factors from the variable importance analysis. Additionally, the aim of this study is to create an automated process for generating the training dataset of fire occurrence locations over a large area using freely accessible GEE tools and its satellite imagery collections. Hence, the following research questions were framed.

1) ***Research question:*** *What are the main characteristics of the last decade's Australian wildfires from freely available satellite datasets?*

2) ***Research question:*** *Which ML algorithm outperforms other existing models available in GEE for prediction of future fire occurrences?*

3) ***Research question:*** *To what extent are the various causal factors associated with the fire locations?*

1.2 Thesis structure

The structure of the thesis is split into eight chapters as follows:

The *Introduction* chapter is dedicated to providing the reader with the motivation of the thesis and the research questions.

The *Background and theory* chapter gives an overview of the application of satellite remote sensing in wildfires, Sentinel missions and Normalised Burn Ratio (NBR). The ML algorithms sub-chapter presents a summary of 3 supervised ML techniques. The used technology and study area are described in the respective sub-chapters. The last sub-chapter describes the determination of the 2019-2020 fire season period.

The *Exploratory data analysis* chapter investigates the wildfires occurrences in Australian using the remote sensing data.

The *Methodology* chapter contains the three sub-chapters. The first sub-chapter named Data mining and preprocessing and presents how the training data are generated for ML algorithms. The classification sub-chapter presents the application of the tree ML supervised classifications. The last sub-chapter called validation evaluates the performance of ML models.

In the first sub-chapter in the *Results* chapter, the results of an automated workflow of fire occurrence detection are presented. The accuracy of ML algorithms applied in this study is presented in the second section. The third sub-chapter reveals the most important variables presented as "wildfire drivers" in fire season 2019-2020 while the last sub-chapter provides the fire occurrence probability map.

The *Discussion* chapter follows from the results chapter and includes some acknowledgement of the potential strengths and weaknesses of the implementation methods.

The *Conclusion* chapter answers the research questions and presents the impact of this study on the achievement of sustainable development goals.

The *Future* chapter summarizes the potential areas of improvement and further research in the area of the presented work.

2. Background and theory

The following chapter introduces the background and theory gathered through the literature review. This chapter is divided into a few sub-chapters, each focusing on different knowledge domains applied in this work. The first sub-chapter emphases the application of satellite remote sensing used in fire detection. The second and third sub-chapters are focused on the Sentinel satellite missions and the NBR definition respectively. The ML algorithms sub-chapter presents the algorithms for wildfire modelling in GEE while the fifth sub-chapter focuses on the technology used in this work. The study area sub-chapter presents the area of interest and the last sub-chapter defines the time frame of fire season.

2.1 Application of remote sensing in wildfires

Satellites use different sensors which measure the intensity of radiation in a range of the electromagnetic spectrum. Some of these sensors capture visible light or near-infrared radiation (passive sensors), whereas other sensors measure the microwave radiation providing its illumination. The Synthetic Aperture Radar (SAR) uses the microwaves which are capable of penetrating through the smoke at high resolution and imaging regardless of day and night. Therefore, remotely sensed data has played a significant role in the fight against wildfires [9].

This unique way of data collection to respond to fires depends also on rapid revisit rates. Some satellites provide a 24-7 bird's eye perspective by observing the same area as geostationary satellites. The launched Japanese geostationary Himawari-8 satellite provides this perspective on Australia as well as other parts of the Asian-Pacific region.

Guang Hu [10] has demonstrated the potential of using weather satellite data for real-time wildfire monitoring. The real-time information from a satellite on the spatial extent of wildfires can help to mitigate the impact of the fire events, especially in early detection of wildfires due to very high temporal resolution. Even though Himawari-8 provides infrared images with a period of 10 minutes, the provided spatial resolution is 2 km, which is not accurate enough to determine the exact spatial location of arisen fires [10].

The detection of active fires using satellite data is based on temperature, where the fire locations have significantly higher temperatures compared to other backgrounds. The fire spots release electromagnetic radiation based on their temperature and this is captured by thermal sensors of satellites [11]. To distinguish fires from the background, it is important to use multichannel detection over the wavelengths in the infrared range.

Figure 1 presents the comparison of radiance against the respective wavelength detected on different objects. The vegetative background is important in fire identification [11] due to it's the detective emissions contrast. For example, the difference between vegetative and fire radiance in the middle infrared (MIR) is important in determining the active fire.

The generated smoke does not normally interrupt the data acquisition linked to the fires due to the large wavelengths of the MIR range compared to the smoke particles which are commonly < 1µm. Therefore, there is no impact of even thick smoke on the detection of active fires [11].

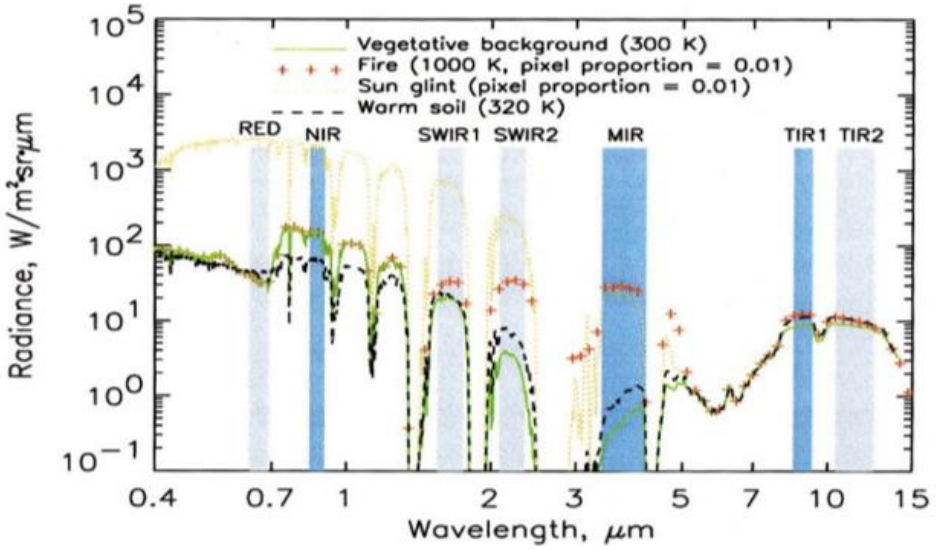


Figure 1 - Spectral radiance of fire against the various typical background as a function of wavelength [11]

2.2 Sentinel missions

The active radar Sentinel-1 satellite and optical Sentinel-2 satellite, provided by European Space Agency (ESA), capture high spatial resolution and 5-day temporal resolution images. The data from both satellite missions can be used to detect and monitor outbreaks of fire, as each sensor has advantages, e.g. cloud penetration of Sentinel-1 and sensitivity to ground moisture of Sentinel-2.

The previous research shows that the Sentinel-1 time-series, in combination with the deep learning framework based on Convolutional Neural Network (CNN) can play a significant role for both detection and tracking temporal progressions of wildfires [12].

In this study, the Sentinel-2 mission is used to detect active fires and burnt areas. This mission is a constellation of twin satellites Sentinel-2A launched by the European Copernicus program on 23 June 2015 and Sentinel-2B followed on 7 March 2017 [13]. Each Sentinel-2 satellite carries a Multi-Spectral Instrument (MSI) which has 13 spectral bands spanning from the visible and the near-infrared (NIR) to the short-wave infrared wavelengths (SWIR) (Figure 2).

The spatial resolution varies from 10 m to 60 m depending on the spectral band and the temporal resolution is 5 days [14]. The Sentinel-2 mission is intended to mostly deliver information for agricultural and forestry practices and applications. The orbital swath width is 290 km. All Sentinel-2 products are projected to the Universal Transverse Mercator (UTM) coordinate system with the World Geodetic System 84 (WGS84) datum [13].

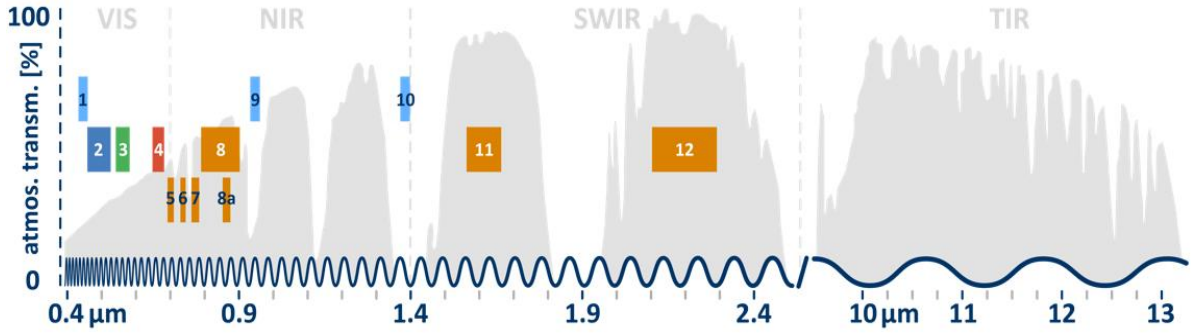


Figure 2 - The overview of Sentinel-2 bands
Source: Reseda, Freie Universität, Berlin

2.3 Normalized burn ratio

Normalized Burn Ratio(NBR) helps to identify burnt areas using the Sentinel-2 dataset. Combining multiple bands in mathematical algorithms can enhance aimed features as each spectral band responds in unique ways to surficial objects, e.g. water content, vegetation, etc. The NBR is frequently used as an index presenting the burnt areas in large fire zones [15]. The NBR formula combines the near-infrared (NIR) and shortwave infrared (SWIR) wavelength [16]. Figure 3 presents the exploiting spectral response curves for burnt areas against healthy vegetation in terms of reflectance as a function of the electromagnetic spectrum. As can be seen, the very high reflectance is for healthy vegetation in the NIR while the low reflectance is in the SWIR portion of the spectrum. This pattern is the opposite of what can be seen in areas devastated by fire. Thus, recently burnt areas demonstrate low reflectance in the NIR and high reflectance in the SWIR.

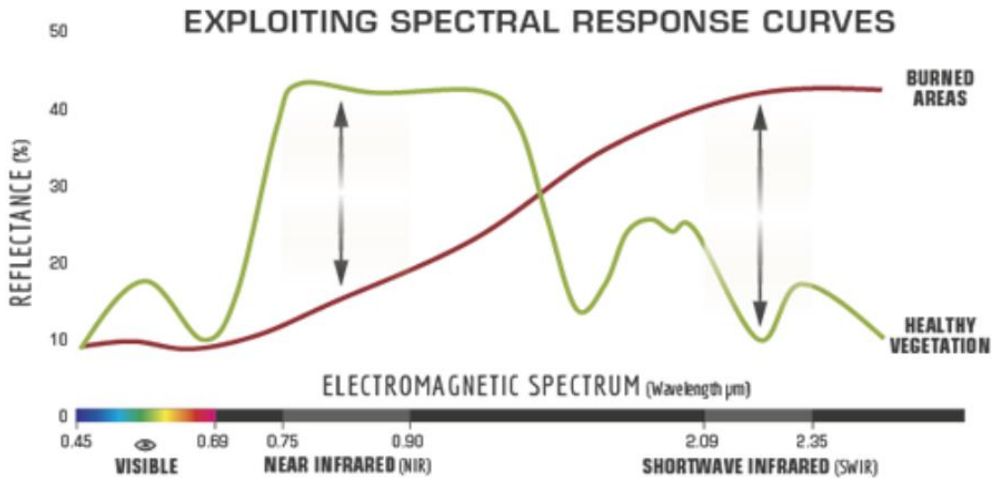


Figure 3 - Contrast of the spectral response curve for heathy vegetation and burnt areas
Source: U.S. Forest Service

Overall, the difference between the spectral response of healthy vegetation and burnt areas reach their peak in the NIR and the SWIR regions of the spectrum.

$$NBR = \frac{NIR_{B8} - SWIR_{12}}{NIR_{B8} + SWIR_{B12}} \quad (1)$$

Where B8 and B12 are the respective satellite bands of Sentinel-2.

$$\Delta NBR = Prefire\ NBR - Postfire\ NBR \quad (2)$$

Higher ΔNBR indicates more severely damaged areas while areas with negative values may indicate increased vegetation following a fire event [17]. ΔNBR is proposed for mapping the burnt severity relied on multispectral images where ΔNBR values can be interpreted based on the United States Geological Survey (USGS) as presented in Table 1 [18] [19].

∆NBR	Burnt Severity
-0,500 – -0,251	High post-fire regrowth
-0,250 – -0,101	Low post-fire regrowth
-0,100 – 0,099	Unburnt
0,100 – 0,269	Low severity
0,270 – 0,439	Moderate–low severity
0,440 – 0,659	Moderate–high severity
0,660 – 1,300	High severity

Table 1 - The burnt severity categories based on ΔNBR according to the USGS

2.4 Machine learning algorithms

One of the main objectives of earth observation is to interpret the observed data, map land use, monitor changes and classify features. ML algorithms can be useful for classifying features, as they can label each pixel to a particular spectral class. The classification, process of assigning the classes to pixels, can be divided into supervised and unsupervised learning [20]. These two techniques depend on user guidance.

The unsupervised classification groups pixels with the common spectral characteristics inherent in the image with no explicit instructions. Thus, unsupervised learning tries to automatically find the structure in data. This method can be used without having previous knowledge of the ground cover in the study site [21]. The popular example of unsupervised learning algorithms is K-means for clustering problems. On the other

hand, supervised learning requires a previously classified sample, i.e., the training dataset. The spectral information from the classified pixels is utilized for training the classification algorithms [21]. This learning is mainly useful in two areas, classification and regression problems.

The algorithm can gradually improve based on the given training dataset. Once a model is trained, the algorithm can be applied to the entire image, and a final classification image is obtained [22].

It is important to fully understand the theory of ML algorithm in order to select and use the model properly. Even though GEE provides 4 available supervised ML algorithms [23], in this study only three supervised ML algorithms are selected based on the literature review.

The following sections describe the CART, NB and RF supervised algorithms used in this study. The second section provides information regarding the accuracy assessment theory applied in ML models. The last section embraces a variable importance analysis.

2.4.1 Classification and regression tree

Classification and Regression Tree (CART) is a model that can be widely used for regression and classification predictive modelling problems. The CART predictive model helps to find a variable based on other labelled variables and the essential benefit of the algorithm are the capability of handling an extensive amount of processed data, ability to capture the non-linearity in the dataset and handle the categorical and numerical features [24]. Moreover, this model can be visualized graphically that enhances the classification model interoperability.

The CART method builds regression or classification models in the form of a tree structure, which consists of nodes and leaf nodes. Each root node represents a single input variable and the leaf nodes of the tree contain an output variable used to make a prediction, e.g., fire (1) and no-fire (0). Thus, the binary tree representation of the CART model makes predictions relatively straightforward.

2.4.2 Random forest

The previously mentioned CART algorithm provides a foundation for Random Forest (RF). The RF model consists of multiple single trees each based on a random sample of the training data which may lead to outperforming the CART model. The drawback of RF is that it is not interpretable as a single CART tree [25].

The RF model is employed to analyze the link between forest fire conditioning factors and the fire occurrence and subsequently used to predict the susceptibility of fires. This algorithm is commonly used for data prediction and suitable for non-linear modelling of forest fire susceptibility [26]. The RF model also allows investigation of the variable importance, which can be used for determining the most important variable from the training dataset [27]. The main benefits of the RF model are that the algorithm avoids the overfitting problem if there are enough trees and can also handle missing values.

The RF algorithm builds many classification trees during the training period and the final output of the model generation process is an average value of the classification results. This structure tree is shown in Figure 4. The purpose of building a decision tree is to generate a model that predicts the value of the objective variable depending on numerous independent input variables.

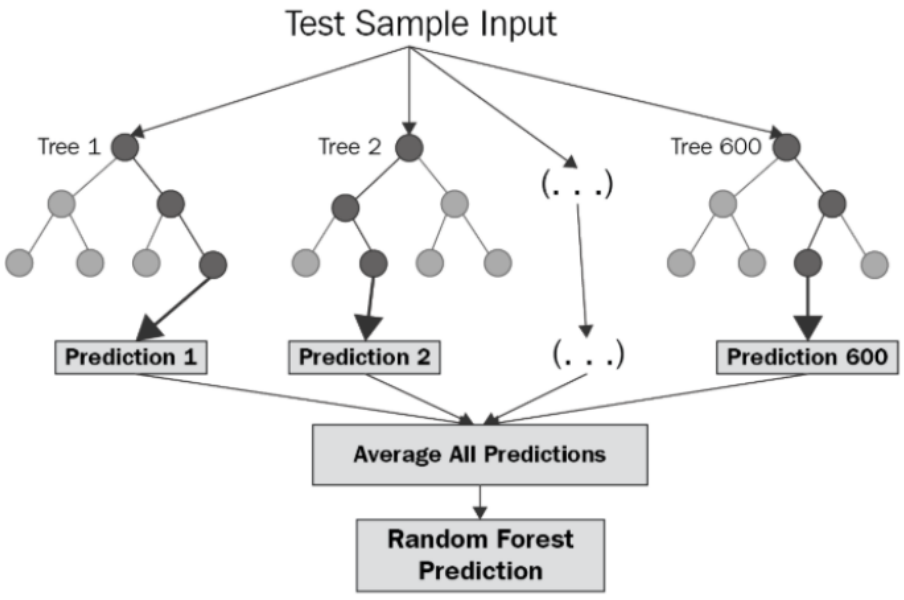


Figure 4 - An example of the RF classification trees structure

2.4.3 Naïve Bayes

Naive Bayes (NB) classifier is a popular algorithm in many powerful ML models. It is known as Naïve because it makes a naive assumption that the presence or absence of a particular element of a class is unrelated to the presence or absence of any other element [28]. This algorithm is founded on the Bayes Theorem presented in Figure 5 created by Thomas Bayes [29].

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

P (A | B) = The probability of A being true given that B is true
P (B | A) = The probability of B being true given that A is true
P (A) = The probability of A being true
P (B) = The probability of B being true

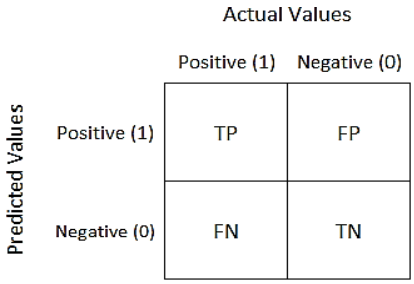
Figure 5 – Bayes theorem

In probability theory and statistics, the Bayes theorem is conditional probability where conditional probability is the probability that something will happen based on that something has already occurred. Thus, by using conditional probability, the probability that a fire event will occur given the knowledge of the prior fire event can be evaluated.

2.5 Accuracy assessment theory

Accuracy assessment gives a general understanding of how the model performs. ML models are prone to overfitting, so it is important to evaluate each ML model using appropriate cross-validation strategies. Thus, the results of selected ML models are validated based on the common accuracy assessment characteristics, such as confusion matrix, overall accuracy and kappa statistics. These characteristics are presented in detail in the next paragraphs.

The confusion matrix is a summary of prediction results on a classification where the number of correct and incorrect predictions are summarized with count values and broken down by each class. There are four basic combinations of predictive and actual values which are explained in Figure 6. This is part of accuracy assessment which provides insight not only into the errors being made by a classifier but more importantly into the types of errors that are being made [30].



True Positive (TP)	*when a model predicts positive and it is true.*
True Negative (TN)	*when a model predicts negative and it is true.*
False Positive (FP)	*when a model predicts positive and it is false.*
False Negative (FN)	*when a model predicts predicted negative and it is false*

Figure 6 - The confusion matrix

Overall accuracy is defined as the percentage of correctly classified results in the confusion matrix. This can be simply computed as shown in equation (3) in percentage [31].

$$Overall\ Accuracy = \frac{(\mathrm{TP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP})}\ x\ 100 \qquad (3)$$

Kappa statistic is one of the most commonly used statistics to test interrater reliability for categorical items. This means that it measures the agreement between more observers, where observers sometimes agree or disagree simply by a chance. The value of kappa statistic is between -1 and 1 and it can be interpreted according to Cohen's kappa in Table 2. The value of 1 means perfect agreement and 0 value is a chance agreement, the most often the value is between 0 and 1. If the value is less than 0, there is worse than a chance agreement (disagreement), which highlights a brutally broken classifier [32].

Kappa	Agreement
< 0	No agreement
0 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81- 1.0	Perfect

Table 2 - Kappa value interpretation according to Cohen (1977)

2.6 Variable importance analysis

The model performance is the key role of the model, but it is just as important to understand how the features of the model contribute to the resulting predictions. ML as the "black box model" can be interpreted and provide insight such as variable importance analysis. This generally refers to how much a given model "uses" that variable to make its predictions. The variable importance is measured by the mean reduction in prediction accuracy [33].

2.7 Technology

In this work, state-of-the-art technology is employed and introduced in the following paragraph.

Google Earth Engine

The Google Earth Engine (GEE) is freely (non-commercial) accessible and available Google product launched in 2010 [34]. This dominant cloud computing platform is designed to store and process massive datasets for scientific analysis and ultimate decision making. Google aims to establish the world's information and make it worldwide accessible and beneficial. GEE has also a commercial license program so it can be purchased for commercial purposes [35].

The predominance of the GEE platform is particularly in handling huge datasets at various scales and building automated programs that can be used at an operational level for many scientists. They use GEE's datasets for forward-thinking in many areas, e.g., flood risk mapping, agriculture, wildfires disaster, Arctic mapping, forest monitoring, land-use change, etc. [36].

This Cloud Storage provides several petabytes of the world's public satellite imagery mostly gathered by NASA's Earth Observing Satellites, e.g. MODIS and Landsat, ESA's Sentinel satellites and many other sources [37]. This cloud storage is available on this website https://earthengine.google.com/datasets/. The vector datasets showing demographic, weather, climate, and digital elevation models and other vector data are also included in these datasets [34]. Datasets can be imported to a scripting environment and users can upload own data for private use. Additionally, any GEE's analysis can be downloaded for use by third-party tools. These datasets should help users to spend more of their time building products and services [35].

Running custom algorithms can be accessed via both the Earth Engine Python and JavaScript application programming interface (API). JavaScript, often abbreviated as JS, is a lightweight and object-oriented programming language. This language is well known for being widely used for web development, i.e., alongside HyperText Mark-up Language (HTML) and Cascading Style Sheets (CSS). The difference between Earth Engine Python and JavaScript application programming interface is mostly in defining functions, defining variables or capitalization of logical operators. The Python API provides a flexible programmatic interface via the Google Colaboratory platform using the Jupyter Notebook interface. This delivers a highly interactive experience without the burden of the local system setup due to a hosted service [35].

2.8 Study area

The study area is the Australian mainland where the wildfires occurred over the 2019-2020 fire season. The Australian mainland includes five states such as New South Wales, Queensland, South Australia, Victoria, Western Australia and major mainland territories, the Australian Capital Territory and the Northern Territory. The map showing the area of interest is presented in Figure 7.

Australia is located between the Indian and Pacific oceans. This world's smallest continent with a heavily concentrated population along the eastern and south-eastern coasts has a wide variety of landscapes, ranging from snow-capped mountains to large deserts. The eastern part of Australia is one of the most fire-prone areas in the world [38].

Several previously undertaken wildfire studies have not analyzed on a state level due to the lack of computing power or absence of datasets over the study areas. Due to the GEE cloud-based spatial processing platform and its multi-petabyte catalogue of satellite imagery it is possible to perform this comprehensive analysis.

Figure 7 - The area of interest defined by the Australian mainland bounds

2.9 Identify the period of the fire season 2019-2020

Many sources are providing different time frames of fire season 2019-2020 and none official source can declare the start and the end date of the fire season. Thus, this section presents identifying the time frame of the recently occurred fire season, as the start and the end date of the fire season not specified officially. This time frame is used as an input for generating the training dataset needed for the ML algorithms.

Input data for representing fire events are gathered from the FIRMS dataset (see more about the FIRMS dataset in chapter 3). The total daily number of fire locations across Australia during the 2019 year and partially 2020 year is shown in Figure 8. This graph reveals the significant growth from September 2019 which decreased almost to 0 in February 2019, precisely between 21st - 22nd February.

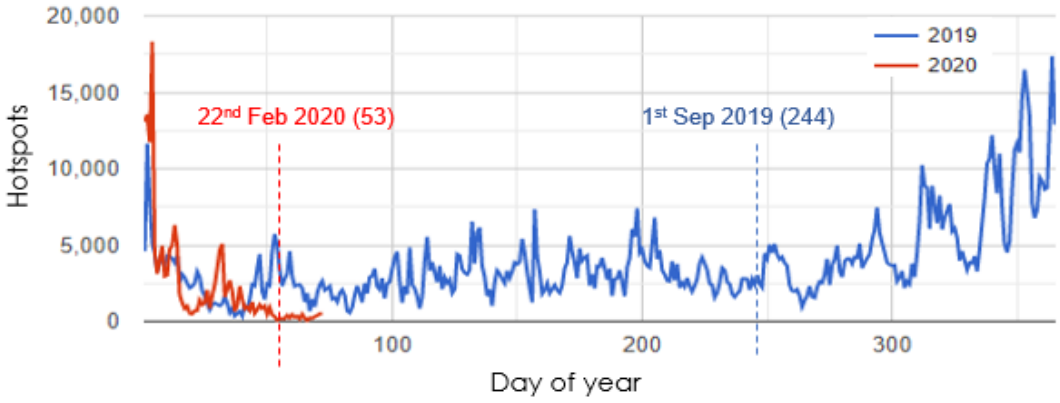


Figure 8 - Total number of fire locations over 2019 and partially for the 2020 year over the Australian mainland

Official information is that Australian's long-running wildfire seasons have been down-graded after heavy rains but without a specific date. Therefore, the overview of the spatial distribution of accumulated monthly precipitation during January and February along with February fire spots across the entire area in Australia is presented in Figure 9. This figure shows that spatial distribution of February precipitation occurred mainly in the north and east part of Australia. The great amount of rain has fallen at fire zones located in the south-east areas that led to stopping them in February. The February fire zones located in the south-west has received less precipitation, but more compared to January precipitation that could lead to stopping active wildfires.

Thus, the time frame is established from 1st September 2019 to 22nd Feb 2020. Input data are gathered from the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) dataset which tracks precipitation back to 1981 [39].

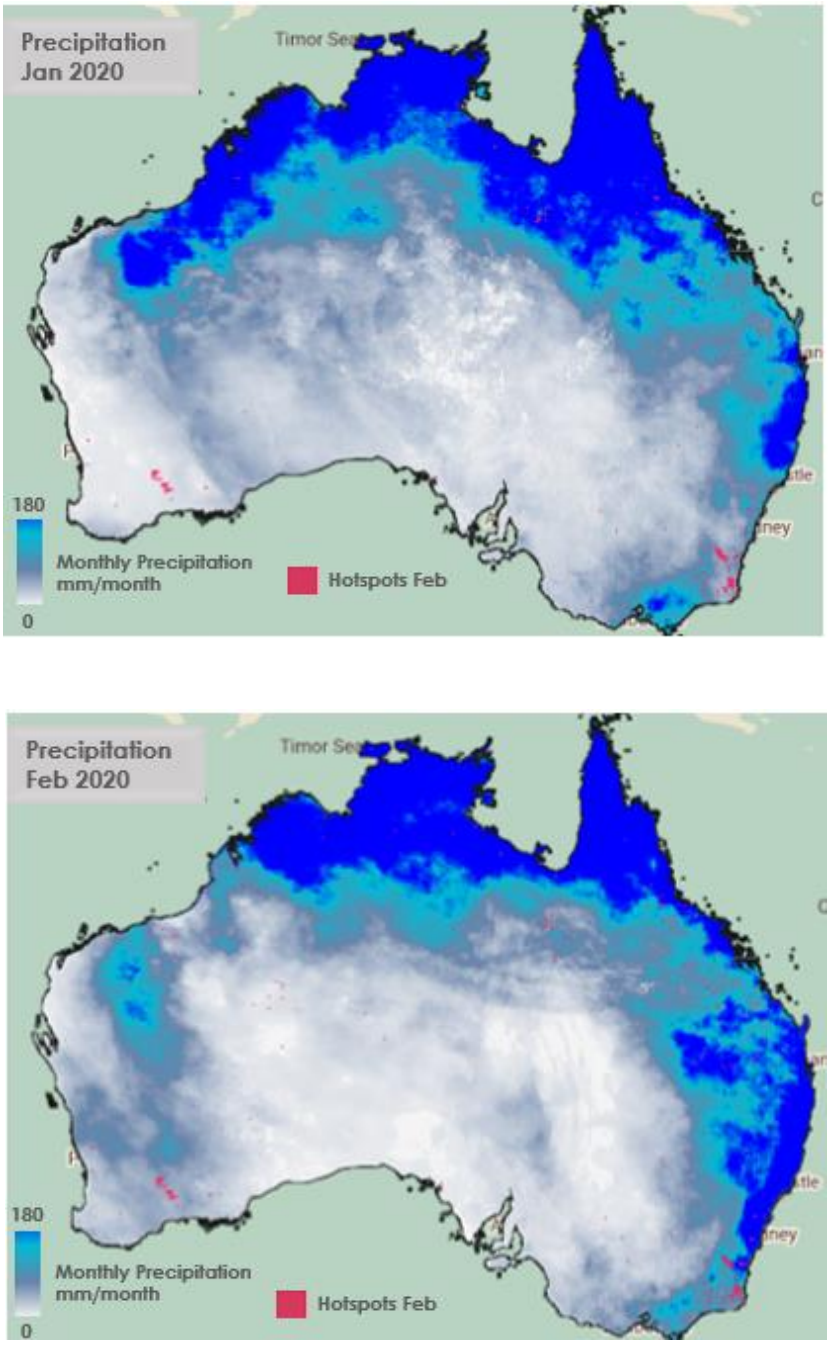


Figure 9 - Spatial distribution of monthly precipitation (mm/month) in Australia during January 2020 - February 2020 using the daily CHIRPS dataset

3. Exploratory data analysis

This chapter presents the exploratory analysis on Australian fires in the 2019-2020 season and compares them with the wildfires from the previous years to outline the main characteristics. The employed datasets for exploratory data analysis include different satellite missions, e.g., VIIRS, MODIS, Sentinel-2. They collected data regularly across the globe. The source codes for generating the figures presented in this sub-chapter are included on the GitHub repository.

To perform this analysis, the European Center for Medium-Range Weather Forecast Reanalysis (ERA5) dataset is used. This dataset is freely available and offers a detailed overview of the atmosphere. The dataset covers the Earth on a 30 km grid and the atmosphere is divided into 137 levels from the surface up to a height of 80 km. This advanced product was released by The European Center for Medium-Range Weather Forecasts (ECMWF) [40]. The ERA5 is part of GEE's datasets consisting of air temperature band as a monthly average at 2 m height with availability from 1979 to present.

Figure 10 presents the mean annual temperature across Australia from 1979 to 2019. As can be seen, the mean annual temperature during these 40 years was the highest in 2019. The difference between the lowest mean annual temperature measured in 2000 and the highest measured in 2019 is approximately 1,8 °C. It is also important to note the highest mean temperature record was broken three times during the last two decades, in 2005, 2013 and 2019. This suggests that Australia is becoming an increasingly warmer place which is most likely due to the global climate change.

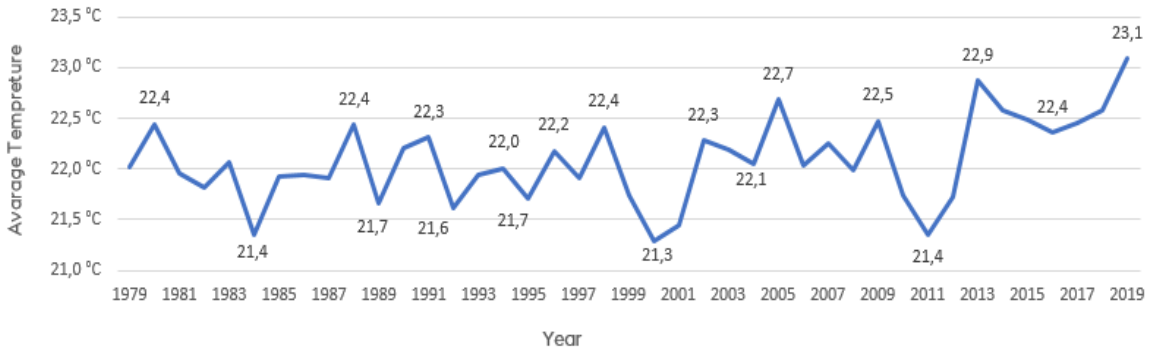


Figure 10 - Mean annual temperature in Australia since 1979 to 2019

For calculation of the total fire occurrence, the GEE's FIRMS dataset is used. Fire Information for Resource Management System (FIRMS) distributes satellite-derived near real-time data within 3 hours of satellite observation. FIRMS is part of NASA's Land, Atmosphere Near real-time Capability (LANCE) for EOS and provides both the Moderate Resolution Imaging Spectroradiometer (MODIS) with Terra and Aqua EOS and the Visible Infrared Imaging Radiometer Suite (VIIRS) data [41].

The active fires shown in figures bellow are presented as pixels covering 1 km^2 on the ground. Therefore, this pixel may contain one or more fire locations within a 500 m radius. Furthermore, the minimum detectable fire size depends on many variables, e.g. scan angle, land surface temperature, amount of smoke, etc. Generally, MODIS satellites can detect both flaming and smouldering fires in 1000 m^2 size but under extremely clean observing conditions smaller flaming fires can be noticed (50 m^2) [41]. Besides, the thermal anomalies, e.g., volcanoes, can be identified as active fires.

The GEE's FIRMS dataset includes the T21 band that shows the active fire locations, where the pixel value determinate the temperature of the surface [42]. This band is measured in Kelvin [41].

Figure 11 presents the total number of fires in Australia each year from 2001 to 2019. The last year, 2019, compared to the previous 18 years does not present outstanding numbers. Both 2011 and 2012 stands for the worst years in terms of fire activity. Recorded active fires in 2017 and 2018 had both approximately 200 000 fires more than in 2019.

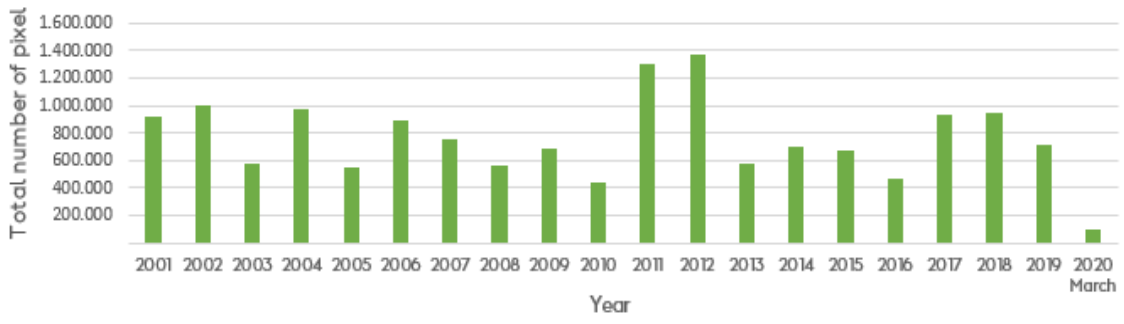


Figure 11 - Total number of pixels presenting active fire annually (1st January 2001 to 1st March 2020)

The detailed overview showing the fire activity over a year is required to uncover anomalies over months. Thus, Figure 12 shows active fires over a year from 2010 to March 2020 in Australia. The 2011 and 2012 years have a significant number of active fires compared to other years. However, the satellite-derived fire data reveals that the most active fires during December and January throughout the last decade happened in 2019 and 2020 respectively. MODIS recorded about 400,000 active fire indicators over Australia between December 2019 and February 2020.

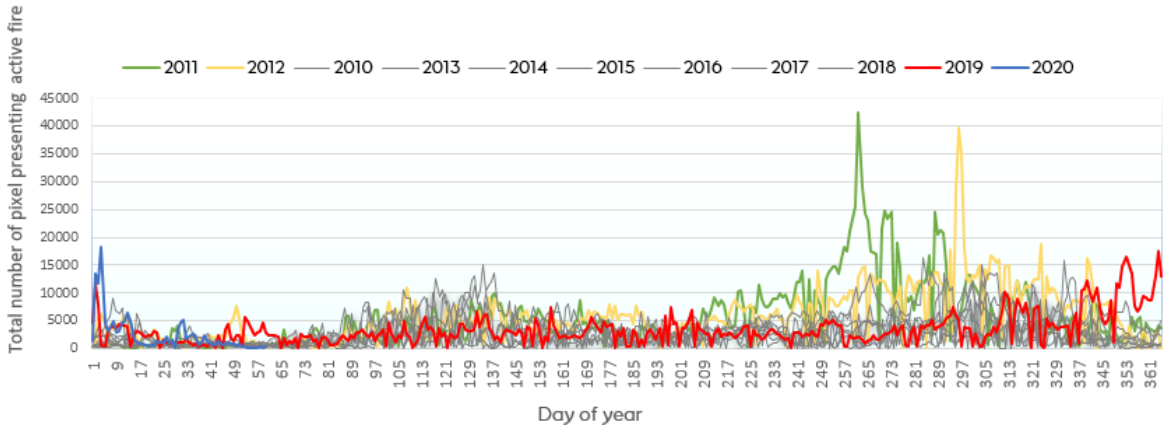


Figure 12 - Total number of fire locations over a year for nearly one decade (1st January 2010 to 1st March 2020), 1km pixel contains one or more fire locations within a 500 m radius

Plotting fire events on a map can present spatial distributions and patterns. Figure 13 shows a spatial distribution map of active fire locations from January 2019 to February 2020. The shown fire locations were remarkably occurring in the north and east coast of Australia while the south and west Australia were slightly fewer fire events. The inland territory was less affected than the coastal area.

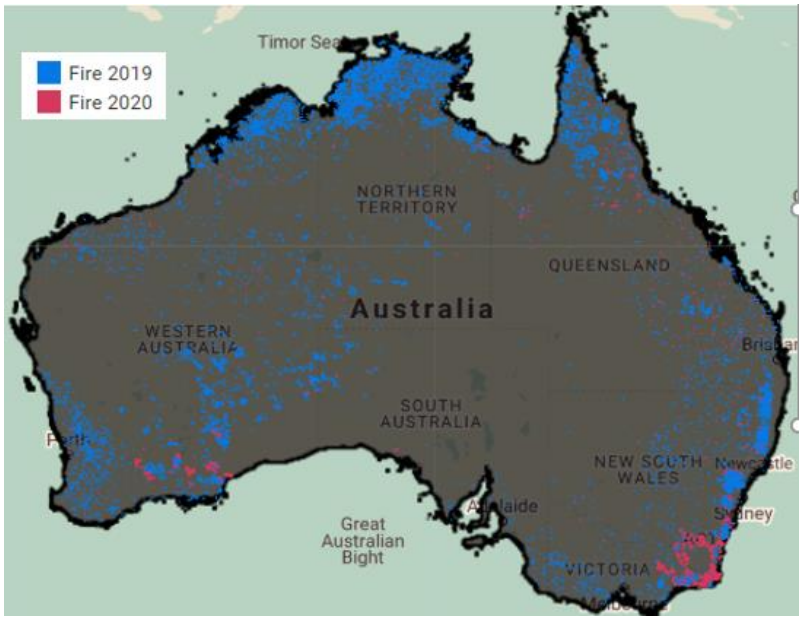


Figure 13 - Distribution of fire events based on the FIRMS dataset from January 2019 to February 2020

4. Methodology

The following chapter describes the methodology used for fulfilling the two objectives of this study, such as fire occurrence probability across Australia and identify the driving factors of wildfires. The entire structure is divided into three parts, such as data mining and pre-processing, classification and validation. This structure is presented in the flowchart in Figure 14 and intends to summarize the essential processes employed in this study.

The first step of the flowchart is creating the training dataset consisting of the previously occurred wildfires (a dependent variable) and the fire main factors, namely topographic, meteorological, anthropological and vegetation factor (independent variables). Subsequently, this set is divided into data sub-sets, called a training and testing dataset. The training dataset is applied in the ML models to train the model and then the trained model is validated by the testing dataset. The best performance from the selected ML models is used for the spatial prediction of wildfire susceptibility. All processes are described in detail in the following sub-chapters.

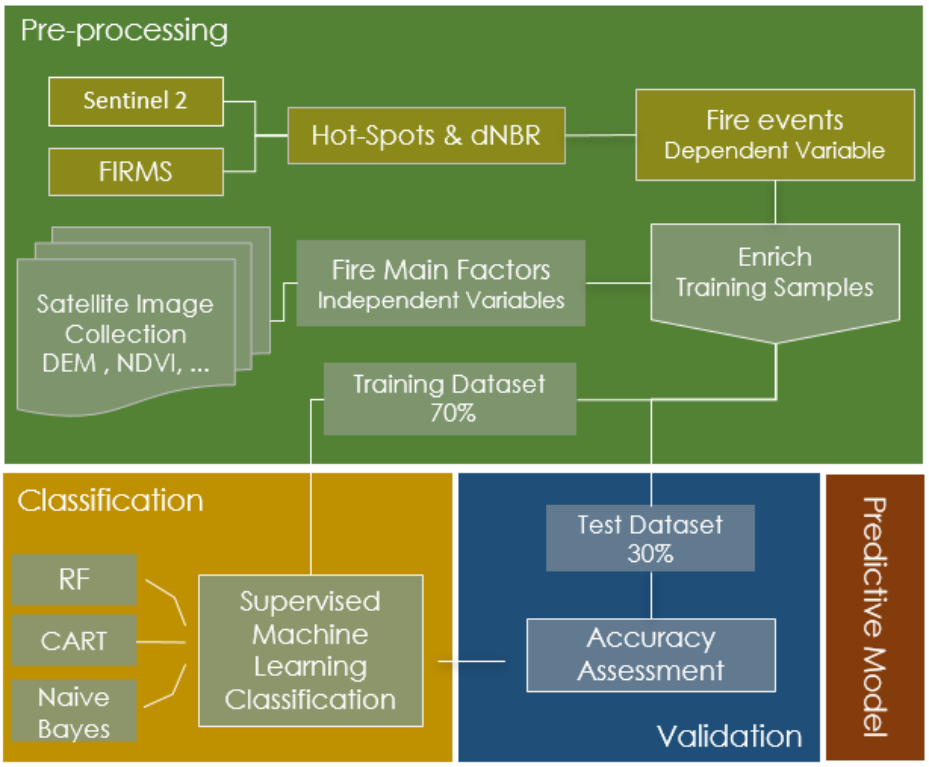


Figure 14 - The flowchart of processes employed in the study for generating the predictive model in GEE

The analysis was conducted in GEE cloud environmental analysis platform using JavaScript, as this enables the global-scale analysis to be completed more efficiently in regard to computing time cost compared to desktop computing. Additionally, satellite images do not need to be downloaded which leads to saving the processing time. The complete code can be obtained from the GitHub repository.

4.1 Data mining and pre-processing

The data mining and pre-processing part are important steps to generate the training dataset as an input for the ML models. The training dataset consists of independent variables also referred to as the predictors (land cover, temperature, etc.) and dependent variables also known as the responding variables (fire, no-fire).

Most of the ML algorithms use the training datasets created manually. In this study, the area of interest is at the continental level and the timeframe covers six months, thus leading to an overwhelming amount of data. Therefore, it is important to automize the process of generating the training dataset. This also brings a benefit to feed the selected models with more samples of training data to improve the models' performances.

4.1.1 Dependent variable

The dependent variable in this study is *fire* and *non-fire* occurrence locations. Thus, mapping susceptibility of fire occurrence can be considered from the ML perspective as a binary classification problem with two classes: fire and no-fire. However, the dataset of recently occurred fire locations with high resolution is not available from the Australian official sources. Therefore, collecting fire and no-fire occurrence locations is developed in this study as an automated workflow presented in Figure 15.

This automated workflow is applied to each month of the fire season (specified in section 2.9) as a consequence of changes in vegetation, which might bias the output results. Additionally, the Australian mainland is split into 3 areas based on state boundaries due to the large size of the Australian mainland that leads to the computational limitation. The workflow is being executed in total 18 times (6 months x 3 parts).

The automated workflow uses two satellite missions, FIRMS and Sentinel-2, which are pre-processed in the interest of obtaining the fire occurrence locations. The FIRMS image collections aggregate active fire locations over the period of one month from the daily observations across Australia with a 1 km^2 employed bounding box. Subsequently, the areas of FIRMS fire locations are vectorized.

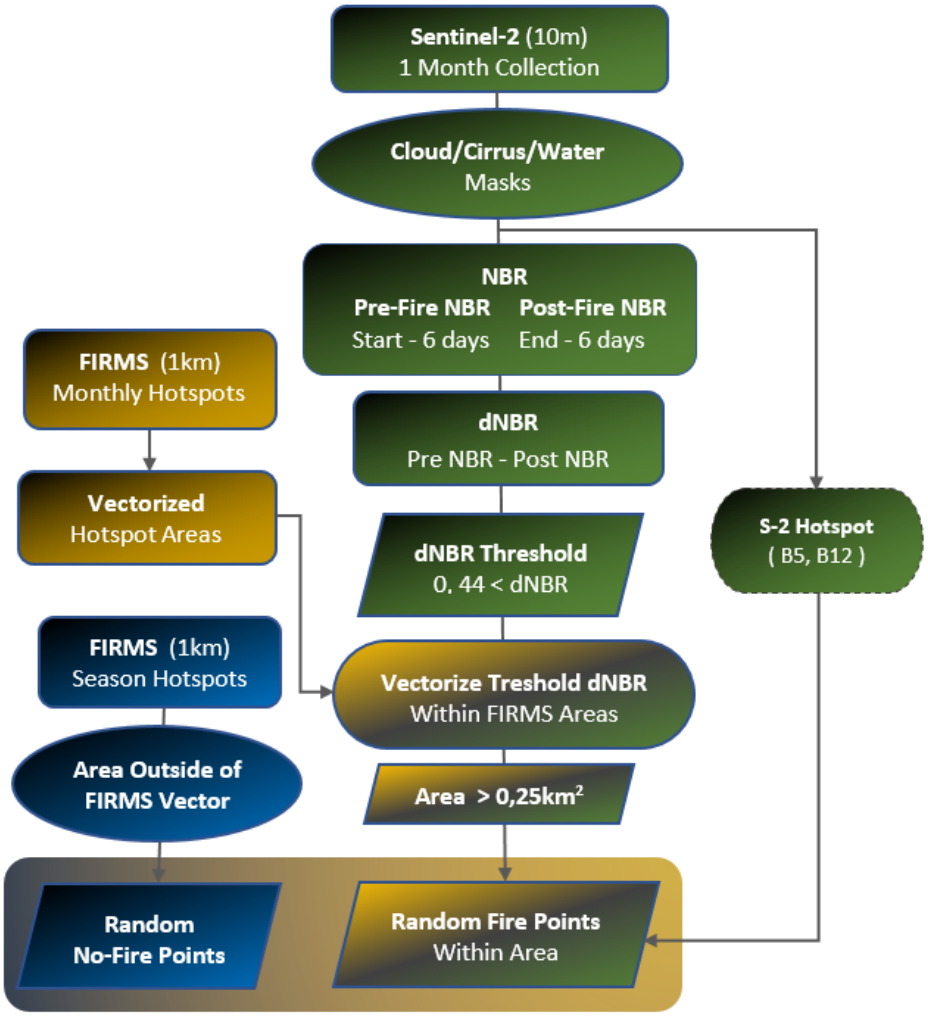


Figure 15 - The flowchart of fire occurrence locations applied in methodology

The Sentinel-2 mission is employed in the second step due to its high spatial resolution. This mission produces cloud and cirrus masks created as a product of the atmospheric correction. These masks are applied with the aim to provide cloudless images and avoid misleading results in the analyses of the surface. Subsequently, the Normalized Difference Water Index (NDWI) calculated from the green (B3) and

shortwave-infrared bands (B11) is applied to remove the water areas from the analysis. Figure 16 shows a difference between the general image and cloud and water-free image.

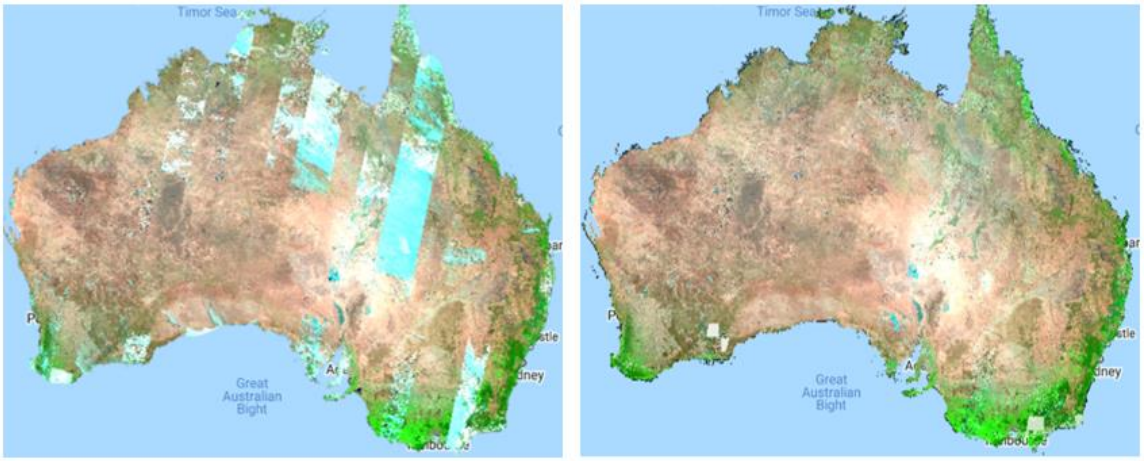


Figure 16 - The non-cloud-masked composite (left),
The cloud and water-free masked composite (right)

The next step is to compute dNBR, see chapter 2.3 for more information on dNBR. The pre-fire NBR is calculated from the time interval <6 days before the start of the month, start month> and post-fire NBR is calculated from the time interval <end month, 6 days after month>. The dNBR calculation highlights the burnt areas and gets an initial assessment of burn severity.

However, there is a dNBR obstruction referring to a change detection process. This means, the dNBR equation consists of deduction of the pre-fire NBR and the post-fire NBR, where changes in natural vegetation, e.g., deforestation, harvest, may be included as well. In other words, non-fire-related changes can be detected as wildfire damage. Despite the short-implemented period (one month), there is set up a dNBR threshold value of 0,44, which classifies the moderate-high severity or high severity burnt area. The threshold is applied only within active fire vector areas from the FIRMS dataset. The aim is to eliminate the small natural vegetation changes and increase the computational power, as the calculation is performed inside the FIRMS fire vector areas. The combination of both features, burnt and fire areas, is applied for creating the balance as the burnt areas tend to underestimate the results while active fire data may overestimate the results.

The selected burnt areas inside the fire location boundary boxes are vectorized and afterwards, the size of selected burnt areas is calculated. The area bigger than

0,25 km^2 (500 m x 500 m raster) is selected for generating the random points. The minimum size criteria mean that the random points are located in selected larger areas as they represent a pixel which covers this particular area.

The no-fire point selection is performed using a random point function where points are randomly placed outside of the FIRMS vector areas.

Figure 17 presents an example of one wildfire that occurred in September 2019 close to the West Coast of Australia. This figure presents the step-by-step results from the previously described processing.

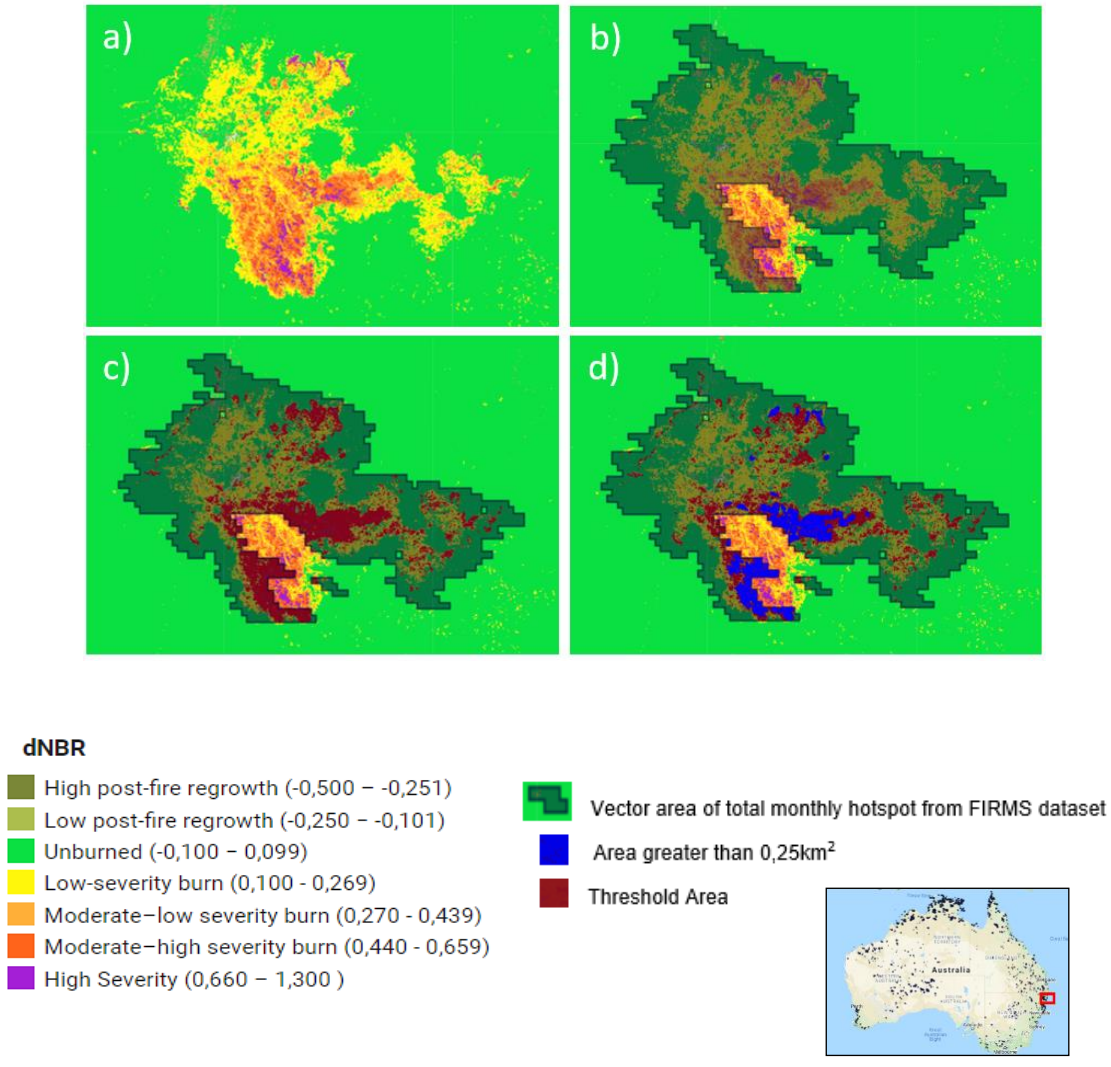


Figure 17 - An example of one wildfire used to illustrate the results from processes of burnt area selection. a) dNBR, b) dNBR with FIRMS vector fire area, c) dNBR with FIRM fire vector area and threshold areas d) dNBR with *FIRMS vector fire area* and *threshold areas and selected areas bigger than 0,25 km^2*

The random point function used in the processing places randomly generates 300 fire points and 300 no-fire points for each selected part (3) for each month (6), which results into 18 CSV files consisting of 600 points per each file. These CSV files are merged into the final file using the JavaScript code stored in the GitHub repository. Each fire and no-fire record in the final file has the "Fire" property name and stored value in the integer type where 1 represents a fire occurrence and 0 presents a no-fire occurrence.

4.1.2 Independent Variables

Selecting independent variables, which are also known as predictors or conditioning factors, is a critical step in predictive modelling. For this study, 15 conditioning factors are selected based on both the field observation found in different studies and available satellite data on the GEE platform. These applied wildfire conditioning factors can be divided into five categories, such as topography, vegetation type, infrastructure, meteorology and socio-economic factors. Table 3 summarizes each of the datasets used in this study.

Topographic category (Figure 18) consists of elevation, slope and aspect. The elevation is obtained from the digital elevation model (DEM) with 30 m spatial resolution. The model is generated from the dataset gathered from the Shuttle Radar Topography Mission (SRTM) provided by NASA. The slope or the gradient of the land expressed as an angle and aspect, also known as the direction in which the slope faces, are derived from DEM.

Category	Data Layers	Source of Data	Data Type	Spatial Resolution
Topography	Elevation	Digital Elevation Data SRTM	Raster	30 m
	Slope			30 m
	Aspect			30 m
Environment	Soil Depth	CSIRO SLGA	Raster	3 arc seconds ≈ 90 m
	Soil Moisture	Terra Climate	Raster	0.25 arc deg ≈ 4 km
	Land Cover	Copernicus CGLS-LC100	Raster	100 m
	NDVI	MODIS NDVI	Raster	250 m
	Drought Severity Index	Terra Climate	Raster	2.5 arc minutes ≈ 4 km
Climate	Precipitation	Terra Climate	Raster	2.5 arc minutes ≈ 4 km
	Maximum Temperature	Terra Climate	Raster	2.5 arc minutes ≈ 4 km
	Wind Power	Terra Climate	Raster	2.5 arc minutes ≈ 4 km
Socio - Economic	Human Population Distributions	World Population	Raster	3 arc second ≈ 85 m
	Global Human Modification	CSP gHM5	Raster	1 km
	Electric Line	OSM	Vector	500 m
	Road Network	OSM	Vector	500 m

Table 3 - The list and description of variable datasets included in the study

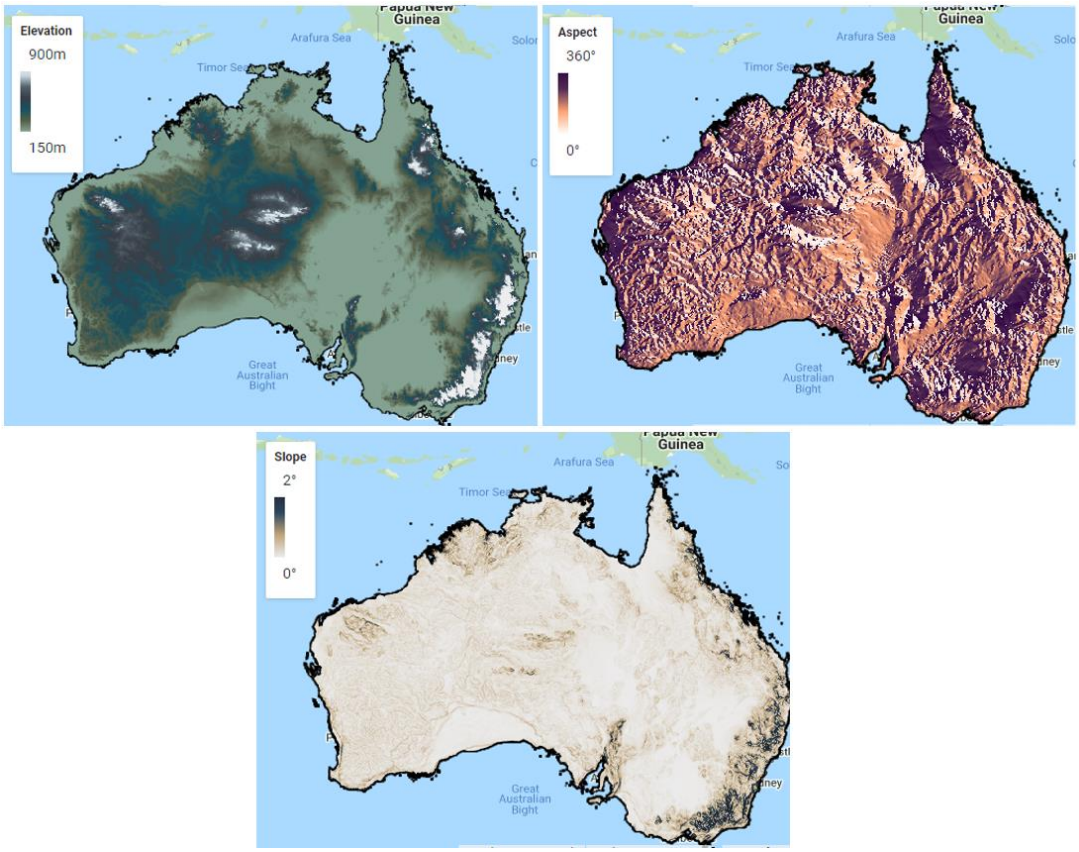


Figure 18 - Topographical factors: elevation, aspect and slope

Environmental category (Figure 19) includes the land cover, soil depth, the soil moisture, the drought severity index and the Normalized Difference Vegetation Index (NDVI). The Copernicus Global Land Service (CGLS) provides the evaluation of land cover at 100 m spatial resolution for the 2015 reference year. The land cover grid has the discrete classes shown in Appendix A. The soil depth gathered from the comprehensive Soil and Landscape Grid of Australia dataset describes the spatial distribution of the soil dept. The soil moisture raster and drought severity index are obtained from the Terra Climate 2019 dataset.

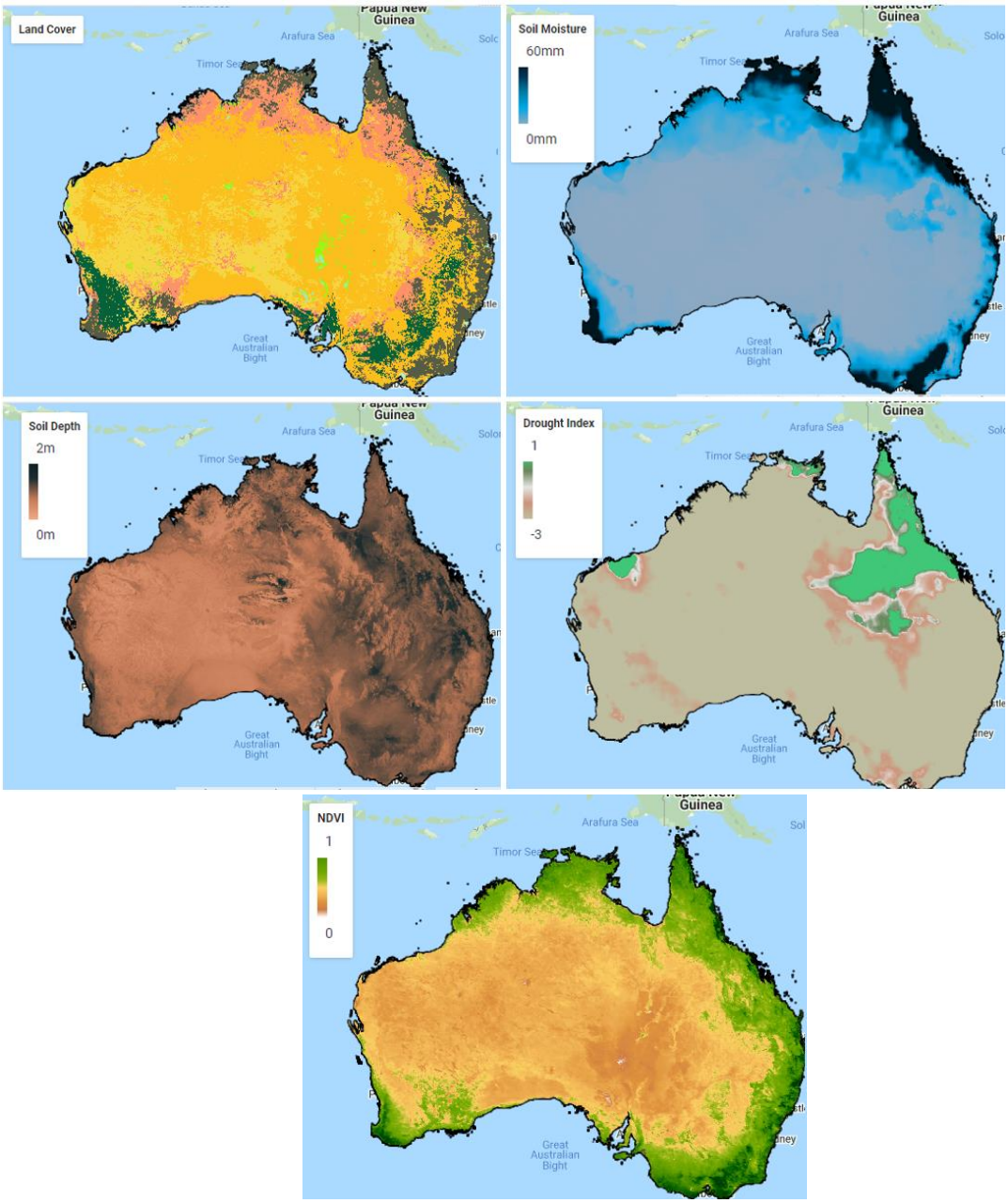


Figure 19 - Environmental factors: land cover (the legend is in Appendix), soil depth, soil moisture, drought severity index and NDVI

These rasters are generated from the image collection obtained from September 2019 to December 2019, where the mean statistic function is implemented (Figure 20). This function takes the mean value of a given pixel over the period. Ideally, the final rasters of both variables should be calculated for the entire fire season; however, the Terra Climate is available only for the 2019 year. The MOD13Q1 product directly provides the vegetation layer, i.e., NDVI, with the 250 m spatial resolution. The NDVI image is generated from the image collection collected during the entire fire season using the mean statistic function value.

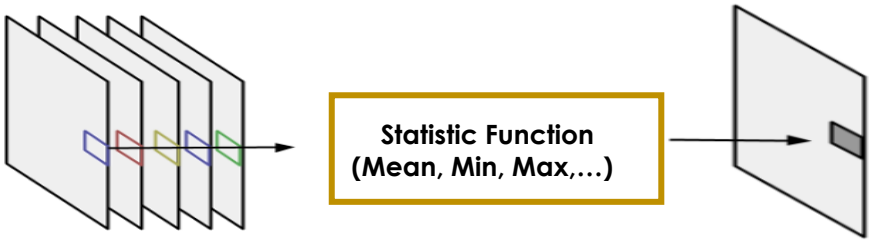


Figure 20 - The example of an image obtained based on statistic function over image collection

Climate category (Figure 21) includes precipitation accumulation, maximum temperature and wind speed. These variables are gathered from the Terra Climate dataset and are processed in the same manner as the previously used data from this dataset; e.g., the drought severity index.

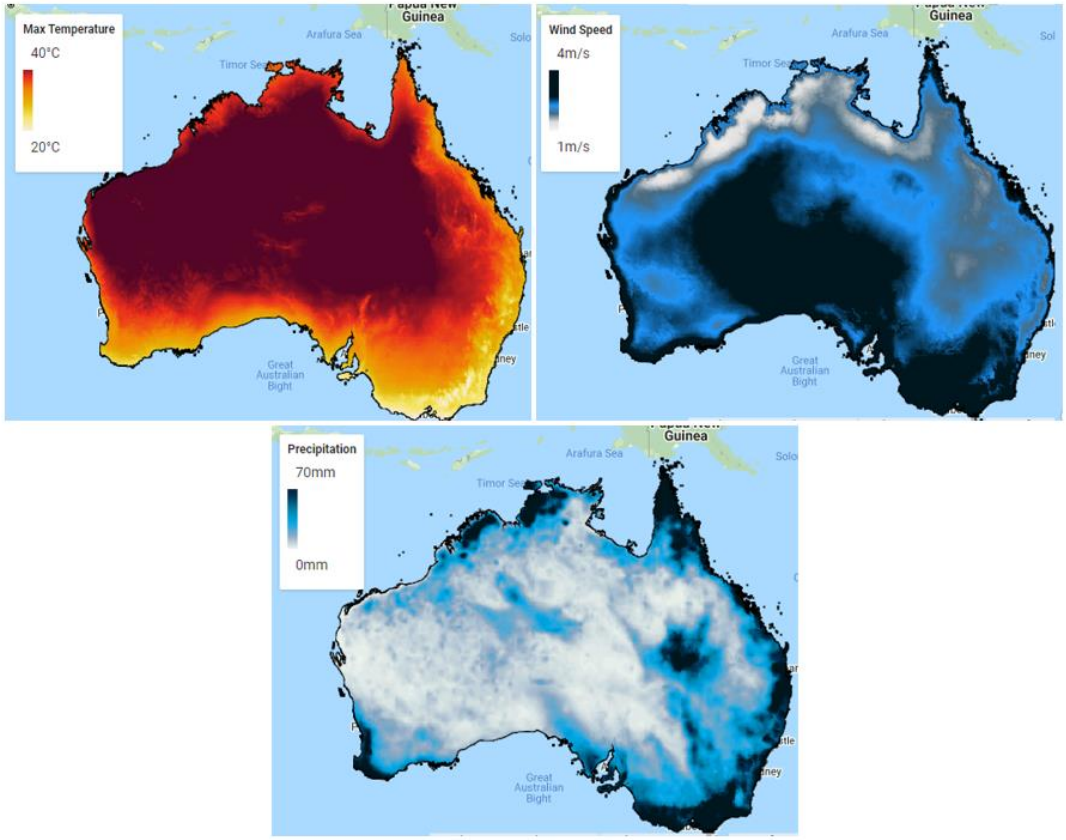


Figure 21 - Climate factors: precipitation, maximum temperature and wind speed

Socio-economic category (Figure 22) includes the Global Human Modification (GHM), population, electric lines and distance from roads. The GHM dataset delivers a cumulative measure of human modification of terrestrial lands over the globe with 1 km spatial resolution. The GHM values vary from 0 to 1 and are associated with a given type of human modification also known as a stressor. The major anthropogenic stressors are included, e.g., human settlement, transportation, mining and energy production. The population from the WorldPop dataset estimated number of people residing in ≈ 85 m grid cells. The vector data, the electric lines and road network, are obtained from the Open Street Map (OSM) and loaded into the GEE platform. Data are converted to the raster format with 500 m resolution where for the road distance the GEE' cumulative coast function is applied.

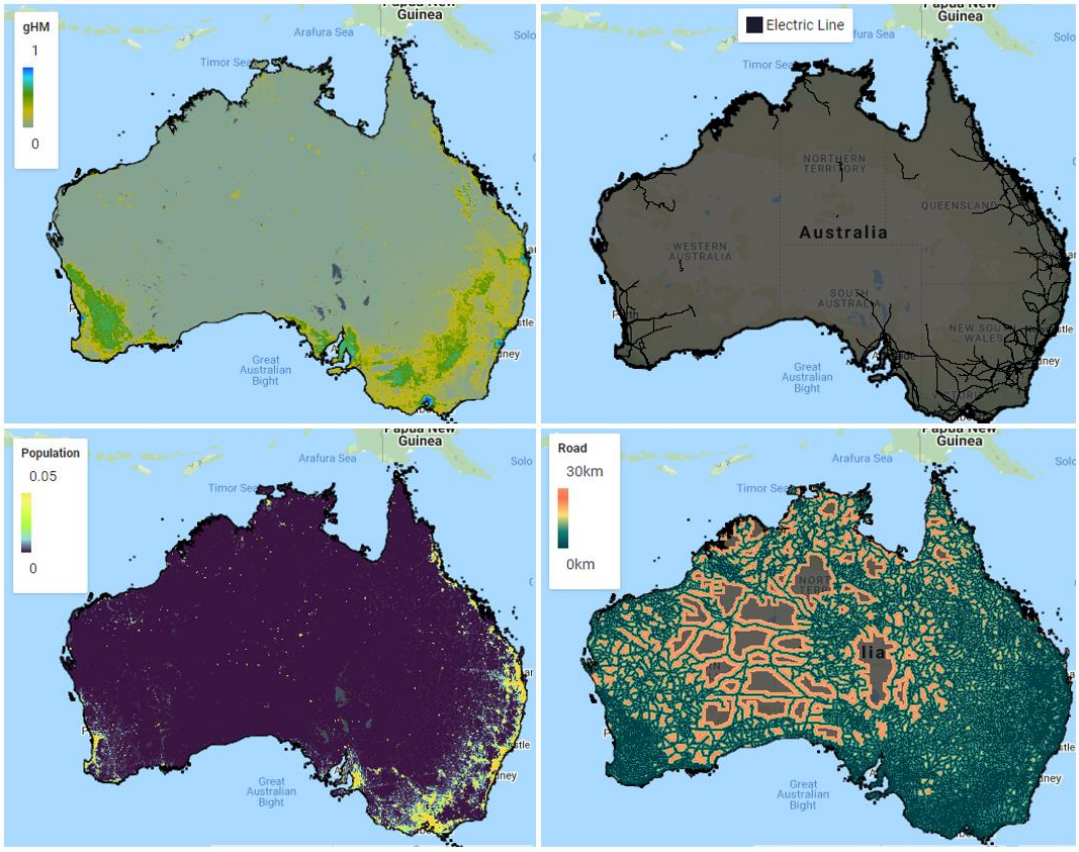


Figure 22 - Socio-economic factors: GHM, population, electric lines and distance from roads

4.2 Classification

When the fire and no-fire training points are created and conditional factors pre-processed, the next step is to create the training dataset which is enriched by predictor values. Firstly, the 15 independent variables are merged to create a composite image with 15 bands (Figure 23).

```
var merge = LandCover.addBands(elevation).addBands(slope).addBands(aspect).addBands(GHM_index)
      .addBands(pop_100m).addBands(Soil_Moisture).addBands(Cost_road_1km).addBands(Ele_Line)
      .addBands(ndvi).addBands(soilDepth).addBands(vs).addBands(temp_max).addBands(Drought_Index)
      .addBands(Precipitation)

var merge = merge.select(
['discrete_classification', 'elevation', 'slope', 'aspect', 'gHM',
'population', 'soil_mean', 'cumulative_cost','constant','NDVI', 'DES_000_200_EV',
'vs_mean', 'tmmx_mean', 'pdsi_mean', 'pr_mean'],
['Land Cover', 'Elevation', 'Slope', 'Aspect', 'Global Human Modification',
'Population', 'Soil Moisture', 'Distance From Road', 'Electric Network', 'NDVI',
'Soil Depth', 'Wind Speed', 'Temperature', 'Drought','Precipitation'])
```

Figure 23 - Merging all predictor variables into the final image (JavaScript GEE script)

Afterwards, the *sampleRegions* function is applied to get the value of predictors into the table and generate training samples as shown in Figure 24. Thus, the fire and no fire- points are overlayed by the composite image to get predictor variables along with labels. A nominal scale for sampling is 100 m.

```
// Sample the input imagery to get a FeatureCollection of training data.
var classifierTraining = merge.sampleRegions(
                {collection: point,
                properties: ['fire'],
                scale: 100});
```

Figure 24 - Creating the training sample

Once the training set is created, the next step is to examine the classifications. The performance of each classification model will be described in the *Results* chapter. These supervised pixel-based classifications rely heavily on the input training samples. The example of applying the ML supervised classification using JavaScript is presented in Figure 25.

```
// Make a Random Forest classifier and train it.
var RF_classifier = ee.Classifier.smileRandomForest(10).train(
                    {features:classifierTraining,
                    classProperty:'fire',inputProperties: bands});

var classification = merge.classify(RF_classifier);
```

Figure 25 - ML supervised classification, namely RF, applied in the GEE interface

Additionally, the GEE classifiers have still a limitation to analyse the variable importance. Even though this study compares three ML algorithms, only one model, precisely RF can observe the link between fire conditioning factors and the fire occurrence, i.e., variable importance. Moreover, the only RF classifier in GEE provides the probability function as shown in Figure 26.

```
var classifier_Pro = ee.Classifier.smileRandomForest(100).setOutputMode('PROBABILITY')
                     .train(classifierTraining,"fire");
var classification_Pro = merge.classify(classifier_Pro);
```

Figure 26 - The probability function in GEE for mapping of fire probability.

4.3 Validation

The trained ML models can predict the fire location; however, it is important to evaluate the performance of these models. For this reason, the accuracy assessment is conducted.

The sample dataset of fire and no-fire location is divided into training and test datasets for model validation. This is conducted by applying the *randomColumn* function which adds a column to the sample dataset and values into a column by default. The points are split with ratio 70:30 meaning that 70 % is used as a training dataset and 30 % as testing dataset. The accuracy assessment is applied to the testing dataset which assesses accuracy based on the confusion matrix. From the confusion matrix, the overall accuracy and the kappa are derived, as can be seen in Figure 27. All results of the validation are presented in the *Results* chapter.

```
var split = 0.7;  //  70% training, 30% testing.
var classifierTraining= classifierTraining.randomColumn();
var trained = classifierTraining.filter(ee.Filter.lt('random', split));
var test = classifierTraining.filter(ee.Filter.gte('random', split));
print('Number of training dataset: ', trained.size())
print('Number of test dataset: ', test.size())

var classifier_trained = ee.Classifier.smileRandomForest(10).train
                         ({features:trained,
                         classProperty:'fire',
                         inputProperties: bands});

var test_classification = test.classify(classifier_trained)
var confusionMatrix =test_classification.errorMatrix('fire','classification');
var confusionMatrixArray = ee.Feature(null, {matrix: confusionMatrix.array()});
print('Confusion Matrix:', confusionMatrixArray);
var overAccuracy = ee.Feature(null, {matrix: confusionMatrix.accuracy()});
print('Overal Accuracy:', overAccuracy)
var kappa = ee.Feature(null, {matrix: confusionMatrix.kappa()});
print(' kappa:', kappa)
```

Figure 27 - Accuracy assessment

5. Results

This chapter summarizes the findings of this study based on the applied methodology. The first section provides the results of the fire occurrence locations gathered from the Sentinel-2 and FIRMS missions. The second part of this chapter reveals the achieved results from the different ML algorithms, where employed prediction variables are gathered from the Earth observation, except for the roads and electric network data. The result of ML algorithms, the fire probability map, is presented in the third sub-chapter of this chapter. Finally, the last sub-chapter provides the results of the variable importance analysis where results are derived from the ML algorithm.

5.1 Fire occurrence location

The fire occurrence points represent a location of individual fires that occurred during the fire season 2019-2020, precisely defined in chapter 2.9. The flowchart presented in Figure 15 in chapter 4.1.1 identifies the fire locations with 10 m accuracy automatically for the ML algorithms. The results show the distribution of fire and no-fire points locations and they are presented in Figure 28. All these locations are a part of the sample training dataset, comprising of 10 800 training points across the Australian mainland.

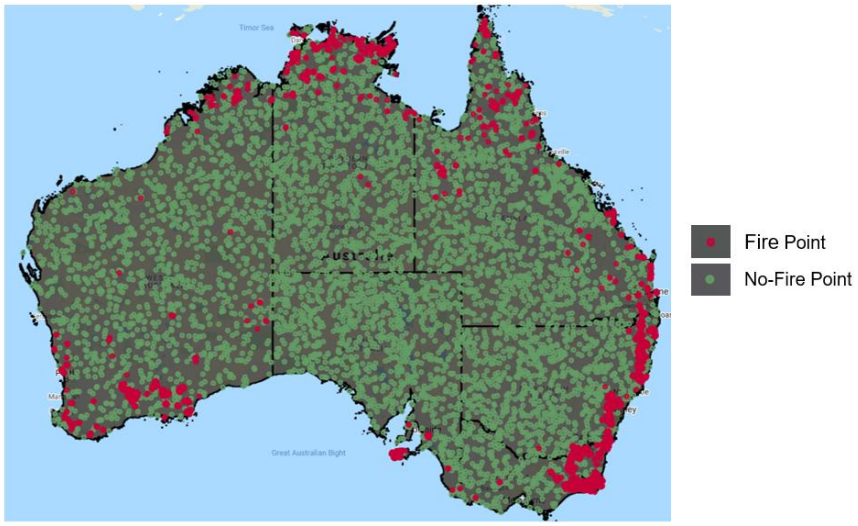


Figure 28 - The distribution of fire and no-fire points from the automated process

The fire location is being visually verified by active fire alerts calculated from the Sentinel-2 data. Figure 29 presents an example of the verification of fire-points. Firstly, the

pre-fire and post-fire area is visualized in the RGB image. The monthly active fire alerts are calculated using B5 and B12 bands and verify fire-points inside the area where the Sentinel-2 fire alert is located.

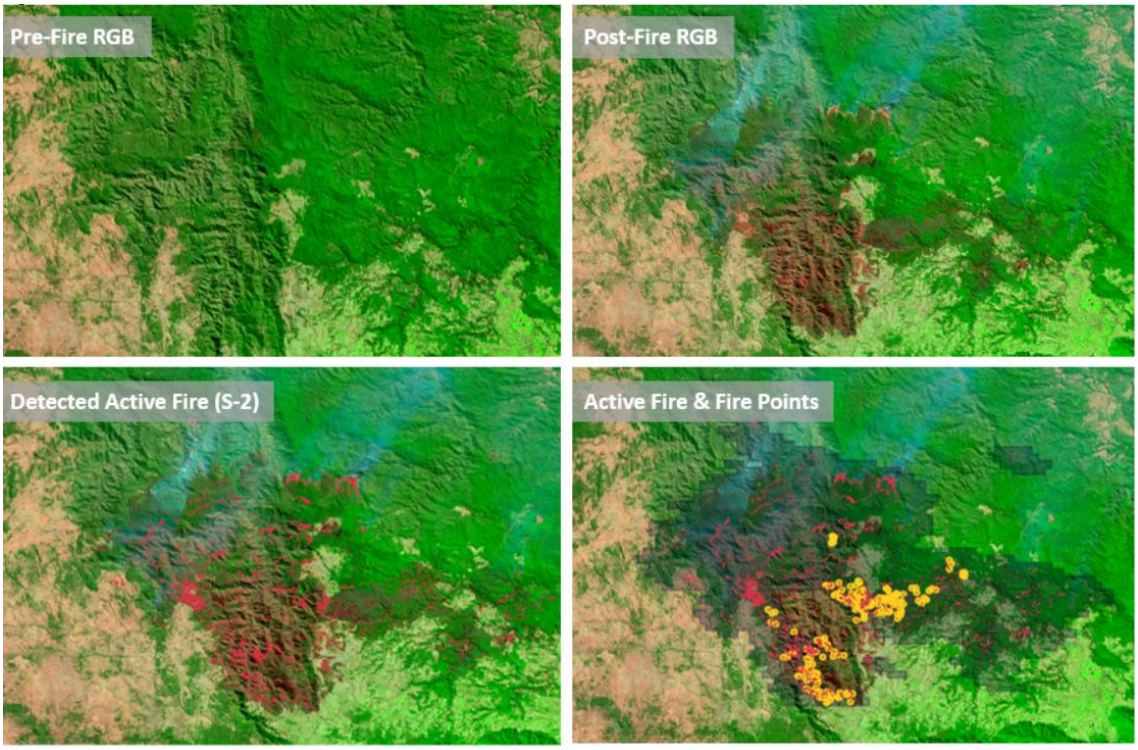


Figure 29 - An example of wildfire in pre-fire and post-fire RGB imagery and monthly active fire from the Sentinel-2 mission for visual verification of fire points.

5.2 Accuracy assessment of ML algorithms

The performance of each classification model is examined in this sub-chapter. The widely used accuracy assessment method is used to evaluate the performance of the ML models. This is calculated on the GEE platform using the characteristics specified in chapter 2.5.

The accuracy assessment is calculated based on the independent testing datasets gathered from the sample dataset. This sample dataset is split in the 70:30 ratio, meaning the 70% of the dataset is used for training the model and 30% is applied for testing. Thus, the selected pixel-based supervised ML algorithms, namely, RF, CART and NB, are trained using a 70% training dataset representing 3250 test samples. The samples contain 1633 fire class and 1617 no-fire class. Table 4 captures the results of ML models' accuracy. The best overall accuracy is shown by the RF model (96%)

while the lowest performance is represented by the NB model (64%). The CART results (93%) are not as accurate as of the RF results but they show better performance than the NB model.

The confusion matrix reveals that these 3 algorithms generally predict well for the no-fire class compared to the prediction of the fire class. The RF model classified correctly the 1593 fire testing samples of 1633 which means that only 40 fire testing samples were predicted incorrectly. The 1540 no-fire samples were predicted properly and only 77 were classified inaccurately.

The NB and CART models cannot handle the classification with missing values. This might occur when processing different predictive factors represented in raster format. These rasters might have a few missing cells representing the absence of data. Therefore, the number of testing samples is less in CART and NB although the input testing dataset is the same as for the RF model.

	Confusion Matrix				Overall Accuracy	Kappa
		Predicted No-Fire	Predicted Fire	Σ		
Naive Bayes	Actual No - Fire	524	1087	*1611*	64%	27%
	Actual Fire	75	1515	*1590*		
	Σ	*599*	*2602*	***3201***		
CART (300)	Actual No - Fire	1494	117	*1611*	93%	88%
	Actual Fire	77	1513	*1590*		
	Σ	*1571*	*1630*	***3201***		
Random Forest (300)	Actual No - Fire	1540	77	*1617*	96%	93%
	Actual Fire	40	1593	*1633*		
	Σ	*1580*	*1670*	***3250***		

Table 4 - Overall statistics of the accuracy assessment results of ML algorithms

The accuracy assessment script with the RF and CART algorithms was executed multiple times to find the proper number of the maximum trees for the RF model and maximum leaf nodes for the CART model. This is an essential step as these numbers have a direct impact on the accuracy of the model. Additionally, it can also reveal how many leaf nodes it is important to implement when two classes are classified.

As seen below in Figure 30, the accuracy of the CART model increases with the number of leaf nodes until the number of 300 leaf nodes is reached. From more than 300 leaf nodes, the accuracy of the model is almost constant. The results of the RF model shown in Figure 31 reveals that with the increasing number of trees, the accuracy is increased as well. Thus, the optimal number of trees applied in the RF model in this study is 300 trees.

Figure 30 - The accuracy of CART models with a different number of leaf nodes applied

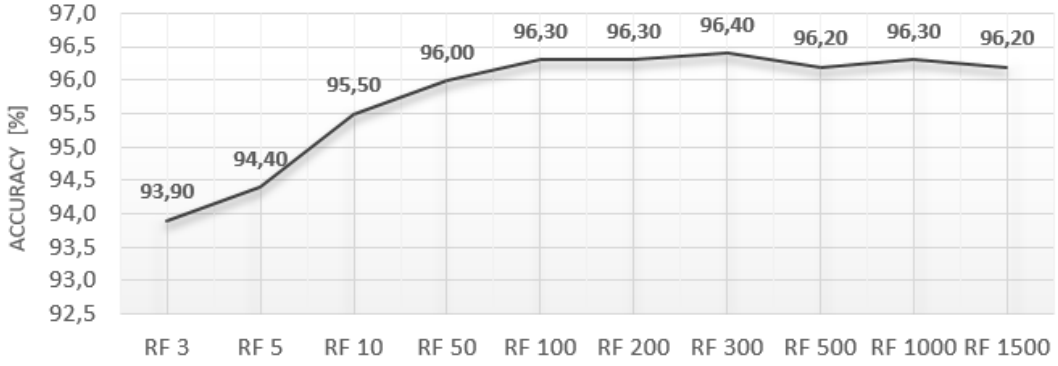


Figure 31 - The accuracy of RD models with a different number of trees applied

5.3 Importance of conditioning factors

The RF model achieves higher accuracy in comparison with other ML models such as NB and CART. Therefore, it is chosen to be the most appropriate and suitable ML model for wildfires prediction. This model enables a quantitative measurement of each variable's contribution to the classification output, which is useful in evaluating the importance of each variable. The variable importance was calculated based on the training dataset.

Figure 32 presents the most important conditioning factors of the wildfires in the 2019-2020 season using the RF model. The most important variables considered as 'key drivers' are the soil moisture and temperature along with drought. The lowest important factors are aspect, land cover and the electric network.

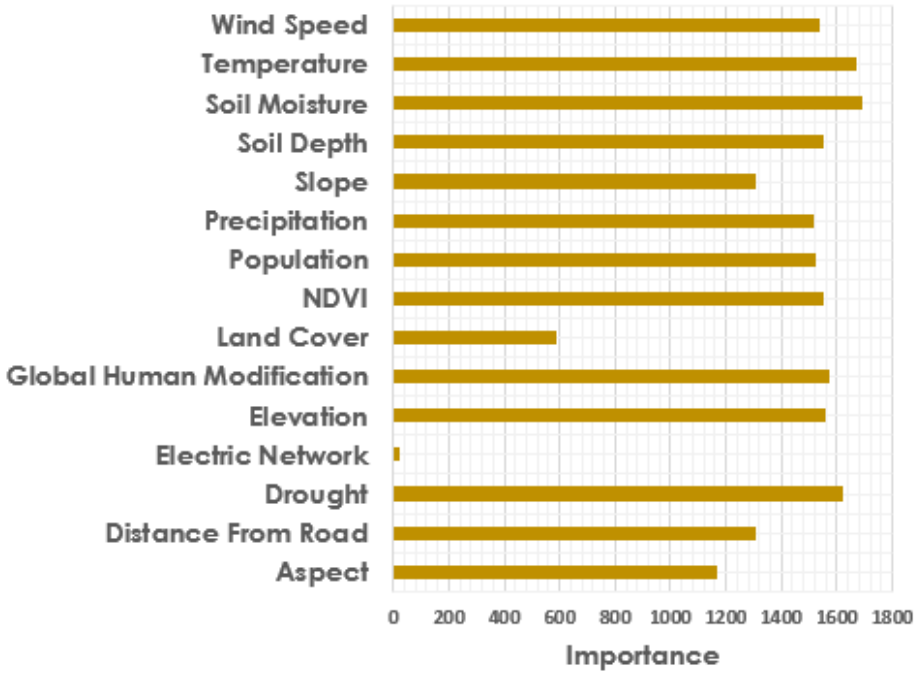


Figure 32 - The variable importance analysis based on the RF model

5.4 Predictive model

Predictive modelling is the overall concept of building an ML model that is capable of making predictions. In this study, the RF model and the training dataset present the wildfires in Australia during the 2019-2020 season. The probability map is shown in Figure 33 where a low value presented by the green colour is an area with the least probability of forest fire occurrence, while the very high value presented by the red colour depicts areas with the highest probability of forest fire susceptibility. The fire risk classes shown in Figure 34 are divided into five classes.

These maps reveal a high risk of fire occurrence concentrated in the coastal area and mainly in the south-west areas in Australia. They also display fire-prone zones distributed throughout the northern coastal regions.

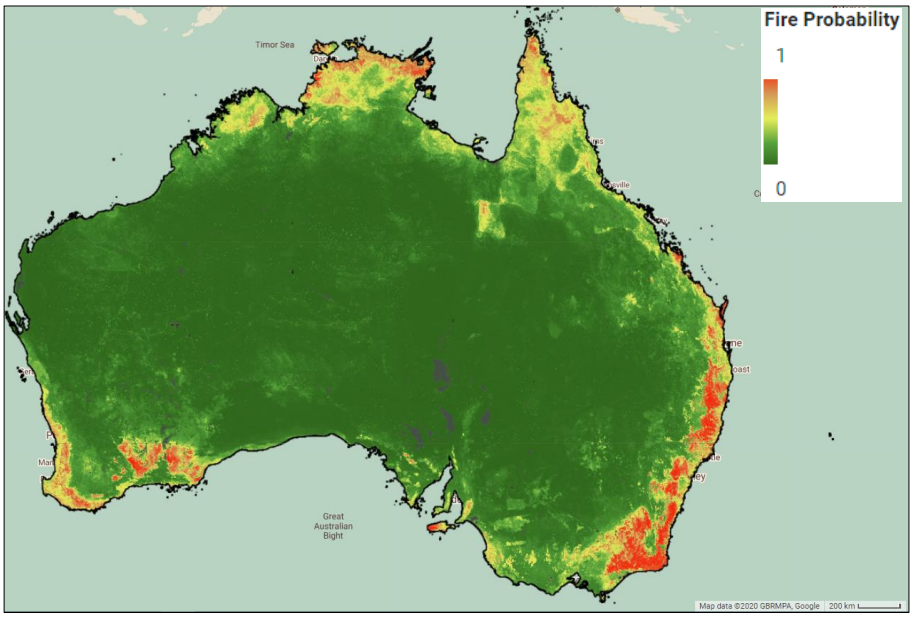


Figure 33 - The fire susceptibility map using the RF model

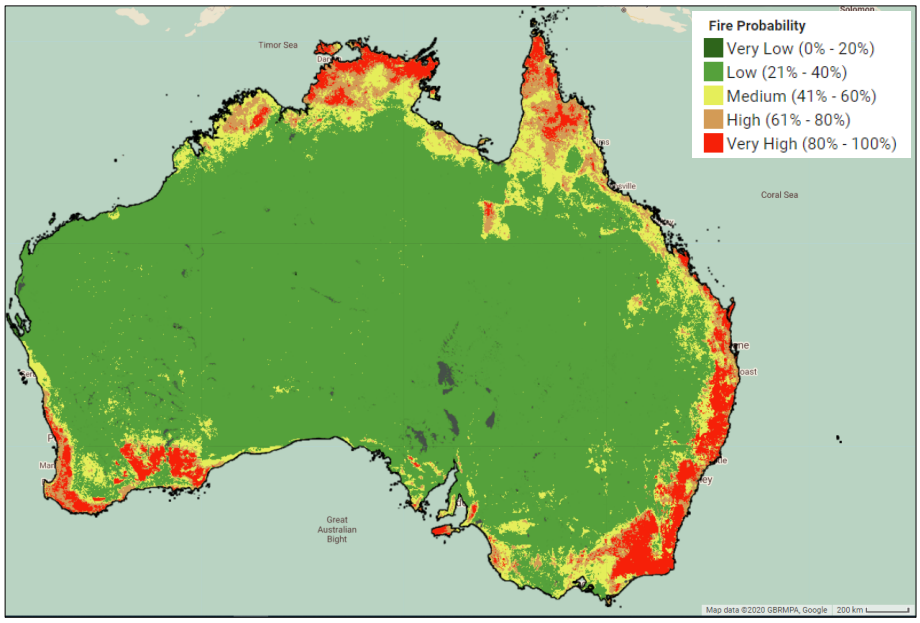


Figure 34 - The fire susceptibility map with classes using the RF model

6. Discussion

The discussion chapter presents findings on the conducted study and evaluates the potential strengths and weaknesses of the implemented methods.

This study is focused firstly on a deep understanding of how the fire occurrence dataset can be obtained in order to be used for the ML algorithms and predict the fire occurrence probability. Many studies used 1 km FIRMS datasets gathered from the Earth observation and showing the active fires. However, this approach of mapping the fire occurrence provides false detections, and the spatial resolution can be also enhanced.

Thus, this study introduces an innovative and automated approach for gathering the samples of fire occurrence locations across the Australian mainland with 10 m spatial precision. The active fire FIRMS locations with 1 km resolution are used as the area of interest where dNBR can be calculated using the Sentinel-2 satellite data. This improves the spatial resolution of the FIRMS active fire locations as Sentinel-2 provides 10 m spatial resolution and reduces the computational time due to chosen FIRMS areas where the dNBR is calculated. Moreover, using these two datasets can decrease the number of false detections of active fires as the dNBR can reveal the burn severity areas.

A limitation of this workflow is circumvented by short time frames, not bigger than 1 month because it can bring the biased results as the burnt severity areas would be influenced by the natural vegetation changes.

Additionally, this workflow as a JavaScript code can be executed on the GEE cloud-based platform, which makes easy access for a potential user. Additionally, the user can modify a custom period (start and end date of the fire season) and add the study area in the vector format or create a spatial boundary defined as a polygon through the drawing tool in GEE. The CSV output of the training dataset is exported to Google Drive and it can be imported into the ML code for further calculations.

The second aim of this study is an attempt to compare different ML approaches where the best model performance is used to map the fire occurrence probability. The three ML algorithms were applied and validated by the testing dataset. The results depicted that the RF model has the best performance while the worse performance showed in the NB model.

The number of trees in the RF model was tested in order to increase accuracy. It turns out that the model with 300 trees can achieve the best performance. However, this number of trees in the model might increase when more predictive variables would be implemented into the current model. Generally, the ML algorithms in GEE can be processed without identifying the numbers of trees of leaf nodes for the CART model due to the implemented default values.

One advantage of RF is its capability of handling categorical variables, such as soil moisture, NDVI, precipitation, etc.. This leads to analyse the variable importance of the 15 variables to show the contribution of each variable. The results show, that the most important fire driver factor in wildfire modelling is soil moisture. The second most important was temperature and then drought, GHM and elevation. The variable which was ranked lowest on the variable importance plot was the electric network.

The predictive performance of the RF models implemented in the present study is suitable as the confusion matrix showed only 117 samples of 3250 were detected incorrectly. Therefore, this model was used to show the susceptibility map displaying the spatial probability of an area to burn. In other words, the map shows the probability for each pixel to burn under the assumptions which are based on conditioning variables and are therefore specific to it. Nonetheless, the wildfires are structurally complex and vary widely in their physical attributes. Thus, the integration of other key factors might increase the complexity model and increase accuracy. The advantage of this model is that it can incorporate different causal factors readily.

It is always essential to validate the stability of ML models. This study used the most common validation, the train/test split technique. This approach brings the benefit that the model responds to previously unseen data can be seen. Moreover, the testing sample was produced via random numbers, which should mitigate the risk of sampling bias.

This development presents the great opportunities of GEE platforms used for the research due to the free availability of datasets and processing the algorithms in the cloud environment. For these reasons, there is no need to download, store, process and analyze the great amount of data on a local computer, however, the internet connection is required. Thus, the entire scope of the study, from generating a training dataset and pre-processing satellite data and trained ML model was conducted in

the powerful GEE cloud-based tool across the massive area of interest. This analysis with spacious datasets would not be possible to undertake on a local computer.

On the other hand, there are also limitations such as exporting the raster data with a good resolution across entire Australia, even when the area was split into multiple grid areas. Also, this platform is ultimately not optimal due to the lack of access to statistics regarding the classification. Even though the numerous satellite missions are presented in the GEE library, most of them provide data for America or Europe. It would be better to use more conditioning variables referring to the wildfires by applying different satellite missions that cover Australia.

This study combines remote sensing, big data, and data mining algorithms and machine learning models to handle data collected from satellite images over large areas and retrieve insights from them to predict the occurrence of wildfires. This was conducted to avoid similar disasters by better planning of infrastructure in disaster-prone areas. The current decision support systems can use this predictive model with the input variables substituted with daily information from the earth observations. An accurate knowledge of the spatial distribution of fire-prone areas can be essential for forest fire risk management.

7. Conclusion

In this chapter, the proposed research questions are answered.

1 ***Research question:*** *What are the main characteristics of the last decade's Australian wildfires obvious from freely available satellite data?*

Among wildfire domains, it is important to illustrate the fire exploratory analysis. The analysis of Australian wildfires discloses that both 2011 and 2012 stands for the worst years in terms of fire activity from 2001 to 2019. However, the most active fires during December and January months from the last 10 years occurred in the 2019-2020 season. The satellite-derived fire data also reveal that approximately 200 000 fewer fires occurred in 2019 than in 2017 and 2018.

Additionally, Australia is becoming a warmer place based on satellite data from the atmosphere dataset named ERA5, which is due to climate change. Thus, if no mitigation and preparedness actions are taken, Australia will witness more wildfires and more severe wildfires in the future.

2 ***Research question:*** *Which ML algorithm outperforms other existing models available in GEE for prediction of future fire occurrences?*

The study compares the chosen ML classifiers available in the GEE platform and recommended based on the literature review. The CART, NB, and RF models were applied and cross-compared. The accuracy assessment analysis using the independent testing dataset shown that the RF model reached the best performance. It had the highest overall accuracy (96%) along with the highest kappa statistics (93%). The other models performed with a lower overall accuracy, where the overall accuracy was 93% and 64% for CART and NB models, respectively.

3 ***Research question:*** *To what extent are the various causal factors associated with the fire locations?*

The best performing model, the RF model, allows the determination of variable importance analysis. The results of variable importance analysis present that the most important variables are soil moisture, temperature and drought which is in line with

other studies where these factors play a major role as well. On the other hand, the lowest influence had the electric network.

In this study, a data-driven model has been set up on the cloud with the massive datasets and accessible to everyone and executable by any dummy user. This would hardly be possible on a local machine. Furthermore, the application can be turned into a decision support system or warning system for alerting decision-makers and stakeholders in case of severe climatic conditions.

7.1 Sustainable development goals

These large-scale and more intense wildfires are becoming an increasing concern as in unfavourable meteorological conditions they are becoming more extreme. As a result, they endanger both human life and property but also release the harmful pollutant particles and gases contributing to the global climate change. All these wildfire challenges are related to some of the sustainable development goals (SDGs). The SDGs adopted in 2015 aim to balance the economic, environmental and social needs [43].

The enhanced technology helps to achieve the SDGs in many ways. Thus, this study combined the remote sensing, big data, data mining algorithms and machine learning models to collect data from satellite images over large areas and retrieve insights from them to predict the incidence of wildfires. This can support to avoid similar disasters by enhanced planning of infrastructure in fire-prone areas.

This study supports sustainable development in three goals. Firstly, goal number 3 *Good health and well-being* as wildfire smoke contributes to air pollution and irritates the human respiratory system. Secondly, goal number 13, namely *Climate action,* is considered due to the emitting carbon dioxide from wildfires along with other greenhouse gasses which accelerate global warming. Lastly, the goal 15 presents *Life on land* which is referred to by a massive impact of wildfires on land which can lead to a short-term economic decline.

8. Future work

There is remarkable potential to predict natural disasters based on machine learning models with enormous amounts of good quality datasets from remote sensing data. This study shows application on fire disaster occurrence using the GEE for the academic purpose, but the concept of prediction can be applied to different natural disasters. The prediction model might substitute the traditional methods which are used nowadays.

There are still several parts that could be improved in the future. Machine learning models use the training dataset to learn how to recognize patterns and apply technologies. This study compared only three ML algorithms which are suitable in GEE, but it would be interesting to compare other models such as neural networks, where each neuron is represented as circles that are connected. This model can learn, create complex relationships, and make accurate predictions when later presented with new data.

Additionally, the model can be tuned by removing the lowest-ranked conditioning variables and see how the model would be influenced. On the other hand, bringing more relevant condition factors might influence the model. Thus, testing influence by the new independent variable is also suggested as future work.

The ML validation processes can be undertaken through different techniques. This study applied the most common train/test split approach. However, different validation techniques can be likewise applied and bring different assessments of the model effectiveness. Thus, the different validation approaches can be implemented in this study, such as the stratified k-fold cross-validation or holdout sets techniques.

Last but not least, the trained RF model can be incorporated with more training samples but from the historic fire events and not just from the recently occurred wildfires. This might help to tune the model and improvement of its current accuracy.

9. Bibliography

[1] Derek Weber, Mehwish Nasim, Lucia Falzon, Lewis Mitchell, "Arson Emergency and Australia's "Black Summer": Polarisation and misinformation on social media," 2020.

[2] CDP, "Center for Disaster Philanthropy, Australian Bushfires 2019-2020,," *Febrary 17,* pp. https://disasterphilanthropy.org/disaster/2019-australian-wildfires/, 2020.

[3] Pei Yu, Shanshan Li, "Bushfires in Australia: a serious health emergency under climate change," 10 January 2020. [Online].

[4] Bareau of Meteorology, "Annual climate statement 2019," Australia, http://www.bom.gov.au/climate/current/annual/aus/, 2020.

[5] I. Gomez-Jimenez, Raul Romero-Calcerrada, C. J. Novillo, J. D. A. Millington, "GIS analysis of spatial patterns of human-caused wildfire ignition risk in the SW of Madrid (Central Spain)," S. S. B. M. B. 20, Ed., Landscape Ecol, 2007, p. 14.

[6] Nathalie Pettoreli, "Satellite remote sensing for conservation," WWF, 2009, p. 125.

[7] Omid Ghorbanzadeh, Khalil Valizadeh Kamran, Thomas Blaschke, Jagannath Aryal, Spatial Prediction of Wildfire Susceptibility Using Field Survey GPS Data and Machine Learning Approaches, July 2019.

[8] Smaranda Belciug, Florin Gorunescu, "Intelligent Decision Support Systems - A Journey to Smarter Healthcare," Pitesti, Romania, Springer, 2020, p. 157.

[9] F. Sunar, C. Ozkan, "Forest fire analysis with remote sensing data," *Remote Sensing,* p. 13, 14 March 2000.

[10] Guang Xu, Xu Zhong, "Real-time wildfire detection and tracking in Australia using geostacionary satellite: Himawari-8," *Australia,* p. 11, 17 July 2017.

[11] S. Jones, K. Reinke, S. Mitchell, F. Mc Conachie and C. Holland, "Advances in the remote sensing of active fires. Detection, mapping and monitoring v1.0," in *RMIT University*, Australia, Detection, mapping and monitoring v1.0, 2017, p. 40.

[12] Yifang Ban, Puzhao Zhang, Andrea Nascetti, Alexandre R. Bevington, Michael A. Wulder, "Near Real-Time Wildfire Progression Monitoring with Sentinel-1 SAR Time Series and Deep Learning," Nature Research, www.nature.com/scientificreports, 2020.

[13] ESA, "European Space Agency (ESA)," 2020. [Online]. Available: https://sentinel.esa.int/web/sentinel/missions/sentinel-2/overview.

[14] M. Majidi Nezhad, A. Heydari, L. Fusilli, G. Laneve, "Land Cover Classification by using Sentinel-2 Images: A case study in the city of Rome," in *World Congress on Civil, Structural, and Environmental Engineering (CSEE'19)*, Italy, Department of Astronautics, Electrical and Energy Engineering (DIAEE), Sapienza University of Rome Rome, 2019, p. 8.

[15] Carl H. Key, Nathan C. Benson, "Remote sensing of severity, the Normalized Burn Ratio," in *Landscape Assessment (LA), Sampling and Analysis Methods*, 2006, p. 56.

[16] A. E. Cocke, P. Z. Fulé, J. E. Crouse, "Comparison of burn severity assessments using Differenced Normalized Burn Ratio and ground data," *Northern Arizona University,* p. 11, 2005.

[17] Allison E. Snyder, Peter Z. Fulé, Joseph E. Crouse, "Comparison of burn severity assessments using Differenced Normalized Burn Ratio and ground data," *Northern Arizona University,* p. 11, 2005.

[18] FIREMON BR Cheat Sheet, The Normalized Burn Ratio (NBR) - Brief Outline of Processing Steps, June 2004. [Online]. Available: https://burnseverity.cr.usgs.gov/pdfs/lav4_br_cheatsheet.pdf.

[19] A. Kato, L. M. Moskal, J. Batchelor, A T. Hudak, "A.T. Relationships between Satellite-Based Spectral Burned Ratios and Terrestrial Laser Scanning," *Forests,* p. 10, 2019.

[20] B. Huang, T. J.Cova, M. H. Tsou, "Comprehensive Geographic Information Systems," in *GIS methods and techniques*, Hong Kong, Elsevier, 2018.

[21] Mohd Hasmadi, Pakhriazad HZ, Shahrin MF, Evaluating supervised and unsupervised techniques for land, Malaysian Journal of Society and Space, 2009.

[22] M. J. Canty, "Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for Python," in *fourth edition*, United State, 2019.

[23] GEE, "Supervised Classification," Google, 19 Feb 2020. [Online]. Available: https://developers.google.com/earth-engine/classification.

[24] L. Pekelis, "Classification And Regression Trees : A Practical Guide for Describing a Datase," Stanford University, 2013.

[25] Heidi Spratt, Hyunsu Ju, Allan R. Brasierb, "A structured approach to predictive modeling of a two-class problem using multidimensional data sets," https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3661737/, 2013.

[26] Ljubomir Gigović, Hamid Reza Pourghasemi, Siniša Drobnjak, Shibiao Bai , Testing a New Ensemble Model Based on SVM and Random Forest in Forest Fire Susceptibility Assessment and Its Mapping in Serbia's Tara National Park, Remote Sensing Technology Applications in Forestry and REDD, 2019.

[27] Andy Liaw, Matthew Wiener, Classification and regression by random forest, R News, December 2002.

[28] Allen Downey, Anders Gorm, Anna Lincoln, Arauzo, "Naive Bayes classifier".

[29] D. Berrar, "Bayes' Theorem and Naive Bayes Classifier," in *Tokyo Institute of Technology*, Tokyo, Japan, 2019.

[30] Sofia Visa, Brian Ramsay, Anca Ralescu, Easther Knaap, Confusion Matrix-based Feature Selection, The 22nd Midwest Artificial Intelligence and Cognitive Science, USA, 2011.

[31] Maria Antonia Brovelli, Monia Elisa Molinari, Eman Hussein, Jun Chen, Ran Li, The First Comprehensive Accuracy Assessment of Globe Land 30 at National Level: Methodology and Results, www.mdpi.com/journal/remotesensing, March 2015.

[32] Anthony J. Viera, Joanne M. Garrett, Understanding Interobserver Agreement: The Kappa Statistic, Research Series, May 2005.

[33] Kim Calders, Inge Jonckheere, Joanne Nightingale, Mikko Vastaranta, "Remote Sensing Technology Applications in Forestry and REDD+," Basel, 2020.

[34] Onisimo Mutanga, Lalit Kumar, "Google Earth Engine Applications," *Remote Sensing,* p. 4, 12 March 2019.

[35] GEE, "https://earthengine.google.com/faq/," 2019. [Online]. Available: https://earthengine.google.com/faq/.

[36] Lalit Kumar, Onisimo Mutanga, "Google Earth Engine Applications," in *Printed Edition of the Special Issue Published in Remote Sensing*, fireBasel, Switzerland, ISBN 978-3-03897-884-8, 2019, p. 422.

[37] Li Huan, Wei Wan, Yu Fang, Siyu Zhu, Xi Chen, Baojian Liu, Yang Hong, "A Google Earth Engine enabled software for efficiently generating high-quality user ready Landsat mosaic images," p. 7, 6 November 2018.

[38] Williams Robyn, Vandenbeld John, Nature of Australia : a portrait of the island continent, 1988.

[39] Tufa Dinku, Chris Funk, Pete Peterson, Ross Maidment, Tsegaye Tadesse, "Validation of the CHIRPS Satellite Rainfall Estimates over Eastern of Africa: Validation of the CHIRPS Satellite Rainfall Estimates," *ADVANCES IN REMOTE SENSING OF RAINFALL AND SNOWFALL,* vol. nternational Research Institute for Climate, p. 23, 2018.

[40] M. Tarek, François P. Brissette and Richard Arsenault, "Evaluation of the ERA5 reanalysis as a potential reference dataset," in *Hydrology and Earth System Science, EGU*, Canada, 2019.

[41] FIRMS, "https://earthdata.nasa.gov," 25 February 2020. [Online]. Available: https://earthdata.nasa.gov/faq/firms-faq#ed-modis-fire-size.

[42] Louis Giglio, Wilfrid Schroeder, Joanne V. Hall, Christopher O. Justice, "MODIS Collection 6 Active Fire Product User's Guide," December 2018.

[43] Sustainable Development Goals, "TRANSFORMING OUR WORLD: THE 2030 AGENDA FOR Sustainable Development Goals," UNITED NATIONS, sustainabledevelopment.un.org.

10. Appendix

A. Land Cover Description

Value	Colour	Colour HEX	Description
0		282828	Unknown.
20		FFBB22	Shrubs.
30		FFFF4C	Herbaceous vegetation.
40		F096FF	Cultivated and managed vegetation/agriculture.
50		FA0000	Urban / built up.
60		B4B4B4	Bare / sparse vegetation.
70		F0F0F0	Snow and ice.
80		0032C8	Permanent water bodies.
90		0096A0	Herbaceous wetland.
100		FAE6A0	Moss and lichen.
111		58481F	Closed forest, evergreen needle leaf.
112		009900	Closed forest, evergreen broadleaf.
113		70663E	Closed forest, deciduous needle leaf.
114		00CC00	Closed forest, deciduous broadleaf.
115		4E751F	Closed forest, mixed.
116		007800	Closed forest, not matching any of the other definitions.
121		666000	Open forest, evergreen needle leaf.
122		8DB400	Open forest, evergreen broadleaf.
123		8D7400	Open forest, deciduous needle leaf.
124		A0DC00	Open forest, deciduous broadleaf.
125		929900	Open forest, mixed.
126		648C00	Open forest, not matching any of the other definitions.
200		000080	Oceans, seas.

B. Random Forest Model

```
//Author: Andrea Sulova;
//Date: Feb 2020 - May 2020;

//_____Import SHP Australian's States_____

var Australia = ee.FeatureCollection("users/sulovaandrea/Australia_Polygon");

//_____Import States in Australia_____

var Capital_AUS = Australia.filterMetadata("name","equals","Australian Capital Territory");
var Northen_AUS = Australia.filterMetadata("name","equals","Northern Territory")
var Queensland = Australia.filterMetadata("name","equals","Queensland")
var South_AUS = Australia.filterMetadata("name","equals","South Australia")
var Tasmania = Australia.filterMetadata("name","equals","Tasmania")
var Victoria = Australia.filterMetadata("name","equals","Victoria")
var Western_AUS = Australia.filterMetadata("name","equals","Western Australia")
var NewSouthWales = Australia.filterMetadata("name","equals","New South Wales")

var Australia_Mainland= Capital_AUS.merge(Victoria).merge(NewSouthWales)
                        .merge(South_AUS).merge(Queensland).merge(Western_AUS).merge(Northen_AUS)

Map.addLayer(Australia_Mainland,{pallete: '666e64', strokeWidth: 1}, 'Australia Mainland',1);

var Australia = ee.FeatureCollection("USDOS/LSIB/2013").filterMetadata("cc","equals","AS")
Map.centerObject(Australia,4.5);

//_____VARIABLES_____

// 1 LandCover
// COPERNICUS LAND COVER forest_type Class Table:
var LandCover =ee.ImageCollection("COPERNICUS/Landcover/100m/Proba-V/Global")
var LandCover = LandCover.select('discrete_classification').mosaic().clip(Australia);
var Classes ='<RasterSymbolizer>'+
    '<ColorMap type = "intervals" extended="false" >' +
    '<ColorMapEntry color="#0779e4" quantity="11" label="11 - Irrigated croplands"/>' +
    '<ColorMapEntry color="#f6d743" quantity="20" label="20 - Mosaic Croplands/Vegetation"/>' +
    '<ColorMapEntry color="#fcbf1e" quantity="30" label="30 - Mosaic Vegetation/Croplands"/>' +
    '<ColorMapEntry color="#f6f578" quantity="14" label="14 - Rainfed croplands"/>' +
    '<ColorMapEntry color="#06623b" quantity="40" label="40 - Closed to open broadleavedevergreen
    or semi-deciduous forest"/>' +
    '<ColorMapEntry color="#b7efcd" quantity="50" label="50 - Closed broadleaved deciduous
    forest"/>'+
    '<ColorMapEntry color="#94fc13" quantity="60" label="60 - Open broadleaved deciduous
    forest"/>' +
    '<ColorMapEntry color="#75b79e" quantity="70" label="70 - Closed needleleaved evergreen
    forest"/>' +
    '<ColorMapEntry color="#a7e9af" quantity="90" label="90 - Open neepdleleaved deciduous or
    evergreen forest"/>' +
    '<ColorMapEntry color="#698474" quantity="100" label="100 - Closed to open mixed broadleaved
    and needleleaved forest "/>' +
    '<ColorMapEntry color="#00bdaa" quantity="110" label="110 - Mosaic
    Forest-Shrubland/Grassland"/>' +
    '<ColorMapEntry color="#565d47" quantity="120" label="120 - Mosaic
    Grassland/Forest-Shrubland"/>' +
    '<ColorMapEntry color="#ff926b" quantity="130" label="130 - Closed to open shrubland"/>' +
    '<ColorMapEntry color="#ffc38b" quantity="140" label="140 - Closed to open grassland"/>' +
    '<ColorMapEntry color="#fff3cd" quantity="150" label="150 - Sparse vegetation"/>' +
    '<ColorMapEntry color="#4cbbb9" quantity="160" label="160 - Closed to open broadleaved forest
    regularly flooded (fresh-brackish water)"/>' +
    '<ColorMapEntry color="#bbded6" quantity="170" label="170 - Closed broadleaved forest
    permanently flooded (saline-brackish water)"/>' +
    '<ColorMapEntry color="#30e3ca" quantity="180" label="180 - Closed to open vegetation
    regularly flooded"/>' +
    '<ColorMapEntry color="#e84545" quantity="190" label="190 - Artificial areas "/>' +
    '<ColorMapEntry color="#e3fdfd" quantity="200" label="200 - Bare areas"/>' +
    '<ColorMapEntry color="#3f72af" quantity="210" label="210 - Water bodies"/>' +
    '<ColorMapEntry color="#f5f5f5" quantity="220" label="220 - Permanent snow and ice "/>' +
    '<ColorMapEntry color="#252a34" quantity="230" label=" No data"/>' +
    '</ColorMap>' +
  '</RasterSymbolizer>';
Map.addLayer(LandCover.sldStyle(Classes), {}, 'Land Cover',0);
```

```
// 2  30mTographical data processing for land cover classification and RF modelling
var srtm = ee.Image('USGS/SRTMGL1_003');
var srtm = srtm.clip(Australia)
var elevation = srtm.select('elevation');
var slope = ee.Terrain.slope(elevation);
var aspect = ee.Terrain.aspect(elevation);

var palette = ['85a392','565d47','155263','393e46','52616b','c9d6df','eeeeee']
Map.addLayer(elevation, {min: 150, max: 900,palette: palette}, 'SRTM 30m elevation',0);

var palette = ['ececec','c1a57b','30475e','222831']
Map.addLayer(slope, {min: 0, max: 2,palette: palette}, 'SRTM 30m slope',0);

var palette = ['ffffff','ffa372','512b58','2c003e']
Map.addLayer(aspect, {min: 0, max: 360, palette: palette}, 'SRTM 30m aspect',0);

// 3 Population WorldPop Global Project Population Data 100m
var dataset = ee.ImageCollection("WorldPop/GP/100m/pop").filterDate('2019');
var pop_100m = dataset.select('population');
var populationVis = { min: 0.0, max: 0.05,
    palette: ['3C1642','92dce5','affc41','d4ff50', 'f6f578','f6d743','f6f578']}
var pop_100m = pop_100m.mosaic().clip(Australia)
Map.addLayer(pop_100m,populationVis, 'Population 100m',0);

// 4 Road
var road_shp = ee.FeatureCollection("users/sulovaandrea/AUS_roads");
var road_img = ee.Image().toByte().paint(road_shp, 1);
var road_no_img = road_img.unmask(0).gt(0);
var cumulativeCost_road = ee.Image(1).cumulativeCost({source: road_no_img, maxDistance: 30000 });

var cumulativeCost_road_clip = cumulativeCost_road.clip(Australia)
var palette1 = ['024249','16817a','fa744f','ffa372']
Map.addLayer(cumulativeCost_road_clip, {min: 0, max: 50000,
            palette: palette1}, 'Roads cost 50 km', 0);
Map.addLayer(road_img,{min: 0, max: 1, palette: '222831'},'Roads',0);

var Cost_road_1km = ee.Image(1).cumulativeCost({source: road_no_img, maxDistance: 30000})
    .reproject(ee.Projection('EPSG:4326').atScale(500));
var Cost_road_1km = Cost_road_1km.unmask(1000000).clip(Australia)
Map.addLayer(Cost_road_1km, {min: 0, max: 30000, palette: palette1}, 'Roads Coast 30km Raster', 0
);

// 5 Electric Line
var ele_line = ee.FeatureCollection("users/sulovaandrea/Aus_Electric_Line");
var ele_img = ee.Image().toByte().paint(ele_line, 1).clip(Australia);
var ele_no_img = ele_img.unmask(0).gt(0).clip(Australia);
var palette2 = ['06623b', 'black']
Map.addLayer(ele_img,{min: 0, max: 1, palette: palette2},'Electric Line',0);
var Ele_Line = ele_no_img.reproject(ee.Projection('EPSG:4326').atScale(500)).clip(Australia);
Map.addLayer(Ele_Line, {min: 0, max: 1,
            palette: palette2}, 'Electric Line Raster', 0);

// 7 Human Modification - 1km
var GHM = ee.ImageCollection("CSP/HM/GlobalHumanModification")
var GHM_index = GHM.mean().clip(Australia)
var palette_GHM = ['85a392','#C7B808','#4E8E07','26D5F6','DDCC09','#16089C']
Map.addLayer(GHM_index, {min:0, max:1, palette:palette_GHM}, 'Global Human Modification',0);

// 8 MODIS NDVI 250m
var dataset = ee.ImageCollection('MODIS/006/MOD13Q1')
                .filter(ee.Filter.date('2019-09-01', '2020-02-22'));
var ndvi = dataset.select('NDVI').mean().clip(Australia);;
var ndviVis = { min: 0.0, max: 8000.0,
            palette: ['FFFFFF', 'CE7E45', 'DF923D', 'F1B555', 'FCD163', '99B718',
            '74A901','66A000', '529400', '3E8601', '207401', '056201', '004C00',
            '023B01', '012E01', '011D01', '011301'],};
Map.addLayer(ndvi, ndviVis, 'NDVI 250 MODIS',0);

// 9 Soil Depth SLGA: Soil and Landscape Grid of Australia (Soil Attributes)
// 90m Depth of soil profile (A & B horizons)
```

```
var dataset = ee.ImageCollection('CSIRO/SLGA').
                filter(ee.Filter.eq('attribute_code', 'DES'));
var soilDepth = dataset.select('DES_000_200_EV').mosaic().clip(Australia);
var soilDepthVis = {min: 0, max: 2,
        palette: ['252525', 'f1ab86', 'c57b57', '1E2D2F', '041F1E'],};
Map.addLayer(soilDepth, soilDepthVis, 'Soil Depth',0);

// 11 Climate - WIND SPEED 2.5 arc minutes more then 2km
var vs = ee.ImageCollection('IDAHO_EPSCOR/TERRACLIMATE')
                .filter(ee.Filter.date('2019-09-01', '2019-12-31'));
var vs = vs.select('vs').reduce(ee.Reducer.mean()).clip(Australia);
var vsVis = { min:100,max: 400,palette: ['F7F3F0','DFDFDF','496A81','1E96FC','00171F'],};
Map.addLayer(vs, vsVis, 'Wind-speed at 10m Scale 0,01',0);

// 12 Maximum temperature  2.5 arc minutes more then 2km
var temp_max = ee.ImageCollection('IDAHO_EPSCOR/TERRACLIMATE')
                .filter(ee.Filter.date('2019-09-01', '2019-12-31'));
var temp_max = temp_max.select('tmmx').reduce(ee.Reducer.mean()).clip(Australia);
var vsVis = { min: 200,max: 400,
            palette: ['F9EBE0','F5E663','E3B505','F18805','EA2B1F','550527'],};
Map.addLayer(temp_max, vsVis, 'Maximum temperature  Scale 0,1',0);

// 13 Drought Severity Index
var Drought_Palmer= ee.ImageCollection('IDAHO_EPSCOR/TERRACLIMATE')
                    .filter(ee.Filter.date('2019-09-01', '2019-12-31'));
var Drought_Index = Drought_Palmer.select('pdsi').reduce(ee.Reducer.mean())
                    .clip(Australia);
var DroughVis = { min:-300,max: 100,
                palette: ['40C778','6D855D','C0BEA0','CD947B','E5E6E4'],};
Map.addLayer(Drought_Index, DroughVis, 'Palmer Drought Severity Index', 0);

// 14 Precipitation accumulation
var Precipitation= ee.ImageCollection('IDAHO_EPSCOR/TERRACLIMATE')
            .filter(ee.Filter.date('2019-09-01', '2019-12-31'));
var Precipitation = Precipitation.select('pr').reduce(ee.Reducer.mean()).clip(Australia);
var PrecipVIS = { min:0 ,max: 70,
                palette: ['EEF4ED','8DA9C4','00A8E8','007EA7','003459','00171F'],};
Map.addLayer(Precipitation, PrecipVIS, 'Precipitation accumulation mm',0);

// 15 Soil Moisture
var Soil_Moisture= ee.ImageCollection('IDAHO_EPSCOR/TERRACLIMATE')
            .filter(ee.Filter.date('2019-09-01', '2019-12-31'));
var Soil_Moisture = Soil_Moisture.select('soil').reduce(ee.Reducer.mean()).clip(Australia)
var PrecipVIS = { min:0 ,max:600,
        palette: ['EEF4ED','8DA9C4','00A8E8','007EA7','003459','00171F'],};
Map.addLayer(Soil_Moisture, PrecipVIS, 'Soil moisture Scale 0.1',0);

//_____M E R G E ____ A L L__ V A R I A B L E S
_____________________________________________

var merge = LandCover.addBands(elevation).addBands(slope).addBands(aspect)
        .addBands(GHM_index).addBands(pop_100m).addBands(Soil_Moisture)
        .addBands(Cost_road_1km).addBands(Ele_Line).addBands(ndvi)
        .addBands(soilDepth).addBands(vs).addBands(temp_max).addBands(Drought_Index)
        .addBands(Precipitation)

var merge = merge.select(['discrete_classification', 'elevation', 'slope',
        'aspect', 'gHM','population', 'soil_mean', 'cumulative_cost','constant',
        'NDVI', 'DES_000_200_EV','vs_mean', 'tmmx_mean', 'pdsi_mean', 'pr_mean'],
        ['Land Cover', 'Elevation', 'Slope', 'Aspect', 'Global Human Modification',
        'Population', 'Soil Moisture', 'Distance From Road', 'Electric Network',
        'NDVI','Soil Depth', 'Wind Speed', 'Temperature', 'Drought','Precipitation'])

var bands= ['Land Cover', 'Elevation', 'Slope', 'Aspect', 'Global Human Modification',
            'Population', 'Soil Moisture', 'Distance From Road','Electric Network',
            'NDVI', 'Soil Depth', 'Wind Speed', 'Temperature', 'Drought','Precipitation']
```

```
//____R A N D O M __F O R E S T__C L A S S I F I C A T I O N
_____________________________________________

var point = ee.FeatureCollection("users/sulovaandrea/TrainingDataset")

var active_fire_point = point.filterMetadata("fire","equals",1)
Map.addLayer(active_fire_point, {color:'orange',size:0.1}, 'Active Fire Points',0);
var No_fire_point = point.filterMetadata("fire","equals",0)
Map.addLayer(No_fire_point, {color:'black',size:0.1}, 'No-Fire Points',0);

print(point.size())

// Sample the input imagery to get a FeatureCollection of training data.
var classifierTraining = merge.sampleRegions(
            {collection: point, properties: ['fire'], scale: 500});

// Make a Random Forest classifier and train it.
var RF_classifier = ee.Classifier.smileRandomForest(300).train(
                    {features:classifierTraining,
                    classProperty:'fire',inputProperties: bands});

var classification = merge.classify(RF_classifier);

Map.addLayer(classification, {min: 0, max: 1,
            palette: ['green', 'red']},'classification', 0);

var RF_Classsifier = ee.Classifier.smileRandomForest(300)
                .setOutputMode('PROBABILITY').train(classifierTraining,"fire");

var RF_Classs_Pro= merge.classify(RF_Classsifier);

Map.addLayer(RF_Classs_Pro, {min: 0, max: 1,
            palette: ['green', 'red']},'classification_Pro', 0);

print(classifierTraining.size())

var classification_REG = merge.classify(classifier_REG);
Map.addLayer(classification_REG, {min: 0, max: 1,
            palette: ['green', 'red']},'classification_REG', 0);

var test_classification = classifierTraining.classify(RF_classifier)

var confusionMatrix =test_classification.errorMatrix('fire','classification');

var confusionMatrixArray = ee.Feature(null, {matrix: confusionMatrix.array()});
print('Internal Confusion Matrix:', confusionMatrixArray);

var overAccuracy = ee.Feature(null, {matrix: confusionMatrix.accuracy()});
print('Internal Overal Accuracy:', overAccuracy)

var kappa = ee.Feature(null, {matrix: confusionMatrix.kappa()});
print('Internal kappa:', kappa)

//_____ A C C U R A C Y___ A S S E S S M E N T_____________________________

var split = 0.7;  //  70% training, 30% testing.
var classifierTraining= classifierTraining.randomColumn();
var trained = classifierTraining.filter(ee.Filter.lt('random', split));
var test = classifierTraining.filter(ee.Filter.gte('random', split));
print('Number of training dataset: ', trained.size())
print('Number of test dataset: ', test.size())

var classifier_trained = ee.Classifier.smileRandomForest(300).train
                        ({features:trained,
                        classProperty:'fire',
                        inputProperties: bands});

var test_classification = test.classify(classifier_trained)
var confusionMatrix =test_classification.errorMatrix('fire','classification');

var confusionMatrixArray = ee.Feature(null, {matrix: confusionMatrix.array()});
print('Confusion Matrix:', confusionMatrixArray);
```

```
var overAccuracy = ee.Feature(null, {matrix: confusionMatrix.accuracy()});
print('Overal Accuracy:', overAccuracy)

var kappa = ee.Feature(null, {matrix: confusionMatrix.kappa()});
print(' kappa:', kappa)

//___V A R I A B L E_____ I M P O R T A N C E_____________________________________

var RF_Classsifier_explain = RF_Classsifier.explain();
print('Explain:',RF_Classsifier_explain);
var variable_importance = ee.Feature(null, ee.Dictionary(RF_Classsifier_explain)
                          .get('importance'));

var chart = ui.Chart.feature.byProperty(variable_importance).setChartType('ColumnCha
            .setOptions({title: 'Random Forest Variable Importance',
              legend: {position: 'none'},
              hAxis: {title: 'Bands'},
              vAxis: {title: 'Importance'},
              colors: ['2e651b']});

print(chart);


///____PRO_MAP_____________________________
var Pro ='<RasterSymbolizer>'+
      '<ColorMap type = "intervals" extended="true" >' +
      '<ColorMapEntry color="#2e651b" quantity="0" label="0.20"/>' +
      '<ColorMapEntry color="#55a13b" quantity="0.21" label="0.40"/>' +
      '<ColorMapEntry color="#e5ee5a" quantity="0.41" label="0.60"/>' +
      '<ColorMapEntry color="#d39b56" quantity="0.61" label="0.80"/>' +
      '<ColorMapEntry color="#f62008" quantity="0.81" label="0.9999"/>' +
      '<ColorMapEntry color="#f62008" quantity="0.99991" label="0.99999"/>' +
    '</ColorMap>' +
  '</RasterSymbolizer>';
Map.addLayer(RF_Classs_Pro.sldStyle(Pro), {}, 'classifier_legend',1);

var legend = ui.Panel({style: {position: 'middle-right',padding: '8px 15px'}});
var legendTitle = ui.Label({value: 'Fire Probability',style: {fontWeight: 'bold' ,
                  fontSize: '12px', margin: '2px',padding: '2px'}});
legend.add(legendTitle);

//_______Creates and styles 1 row of the legend_____________________

var Row = function(color, name)
      {var colorBox = ui.Label({style: { backgroundColor: color,
                            padding: '8px', margin: '0 0 4px 0'}});
      var description = ui.Label({value: name,style: {margin: '0 0 2px 3px'}});
      return ui.Panel({widgets: [colorBox, description],
                layout: ui.Panel.Layout.Flow('horizontal')})};

var palette =['#2e651b', '#55a13b', '#e5ee5a', '#d39b56', '#f62008'];

var names = ['Very Low (0% - 20%)',' Low (21% - 40%)',
            'Medium (41% - 60%)', 'High (61% - 80%)',
            'Very High (80% - 100%)']

for (var i = 0; i <5; i++) {
  legend.add(Row(palette[i], names[i]));}
Map.add(legend);

//________Basemap_____________________________
var mapStyle = [
  {elementType: 'geometry', stylers: [{color: '#ebe3cd'}]},
  {elementType: 'labels.text.fill', stylers: [{color: '#523735'}]},
  {elementType: 'labels.text.stroke', stylers: [{color: '#f5f1e6'}]},
  {featureType: 'administrative',elementType: 'geometry.stroke',
                stylers: [{color: '#c9b2a6'}] },
  {featureType: 'administrative.land_parcel',elementType: 'geometry.stroke',
                stylers: [{color: '#dcd2be'}]},
  {featureType: 'administrative.land_parcel',elementType: 'labels.text.fill',
                stylers: [{color: '#ae9e90'}]},
  {featureType: 'administrative.land_parcel',  elementType: 'labels.text.stroke',
```

```
                stylers: [{color: '#000040'}, {visibility: 'simplified'}]},
  {featureType: 'administrative.neighborhood',elementType: 'labels.text.fill',
                stylers: [{color: '#408080'}]},
  {featureType: 'landscape.man_made',elementType: 'geometry.fill',
                stylers: [{color: '#800040'}]},
  {featureType: 'landscape.natural',  elementType: 'geometry',
                stylers: [{color: 'blue'}]},
  {featureType: 'landscape.natural',elementType: 'geometry.fill',
                stylers: [{color: 'blue'}]},
  {featureType: 'landscape.natural.terrain',elementType: 'geometry.fill',
                stylers: [{color: 'blue'}]},
  {featureType: 'road',elementType: 'geometry',
                stylers: [{color: '#f5f1e6'}]},
  {featureType: 'road.highway',elementType: 'geometry',
                stylers: [{color: '#f8c967'}]},
  {featureType: 'road.highway',elementType: 'geometry.stroke',
                stylers: [{color: '#e9bc62'}]},
  {featureType: 'road.local',elementType: 'labels.text.fill',
                stylers: [{color: '#806b63'}]},
  {featureType: 'water', elementType: 'geometry.fill',
                stylers: [{color: '#b9d3c2'}] },
  {featureType: 'water',elementType: 'labels.text.fill',
                stylers: [{color: 'blue'}]}];

Map.setOptions('mapStyle', {mapStyle: mapStyle});

```