ERASMUS UNIVERSITEIT ROTTERDAM
ERASMUS SCHOOL OF ECONOMICS

Wildfire risk modelling in Spain using context-sensitive Deep Learning models

Author

J.M.W. Scheltema (577074)

Date

February 7, 2023

Supervisor

dr. P.C. Bouman

Second assessor

prof. dr. ir. Rommert Dekker

Master Thesis Econometrics & Management Science
Business Analytics & Quantitative Marketing

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

Abstract

In many countries, climate change has resulted in more extreme weather conditions, such as intense heat and drought. This has caused a global increase in the frequency of wildfire occurrences as well as their destructive potential. Within Europe, Spain has seen the most significant rise in wildfire damage. There is a strong need for predictive models that can accurately assess future wildfire risk in order to more effectively combat this problem. In this thesis, the suitability of context-sensitive Deep Learning models to this problem is examined. First, a spatio-temporal, gridded data set of Spain was created consisting of various wildfire influencing factors, such as weather, geography and human activity. Then, three different Deep Learning architectures, sensitive to respectively spatial, temporal and spatio-temporal context, were trained on this data to predict the probability of wildfire occurrence and the total area burned of actual wildfires. Strong evidence was found in favor of the suitability of context-sensitive Deep Learning models, although the optimal context differed per classification task. Firstly, the temporal model achieved an accuracy of 74% in the binary classification task of next-day wildfire occurrence. Secondly, the spatio-temporal model achieved an accuracy of 81% in the multi-class classification problem of burned area prediction.

Contents

1 Introduction

Forests are one of the Earth's most important natural resources due to the many ways in which they protect the environment and purify nature. Most importantly, forests regulate the Earth's climate and carbon cycle, which is crucial for maintaining our environment's ecological equilibrium. They do this by extracting and storing carbon from the air, while emitting oxygen and water vapor, which purifies and cools the air. Forests provide many other benefits to humans and all other living creatures on earth. For example, they play an important role in the mitigation of natural disasters, such as floods and landslides, they host and protect 90% of the Earth's terrestrial biodiversity and play an important economic, social and cultural role to mankind (Pawar & Rothkar, 2015). In recent years, however, the health and even survival of forests across the globe has come under threat due to an increase in frequency and destructiveness of wildfires (Wang et al., 2021). Wildfires are unplanned and uncontrolled forest fires that burn in forested areas, grass, or alpine/tundra vegetation (Center, 2003). The increase in frequency and severity of wildfires is often attributed to human-made climate change and a higher rate of industrialization, since these two factors have caused higher temperatures, more irregular rainfall and more wildland-urban interface (Flannigan et al., 2006).

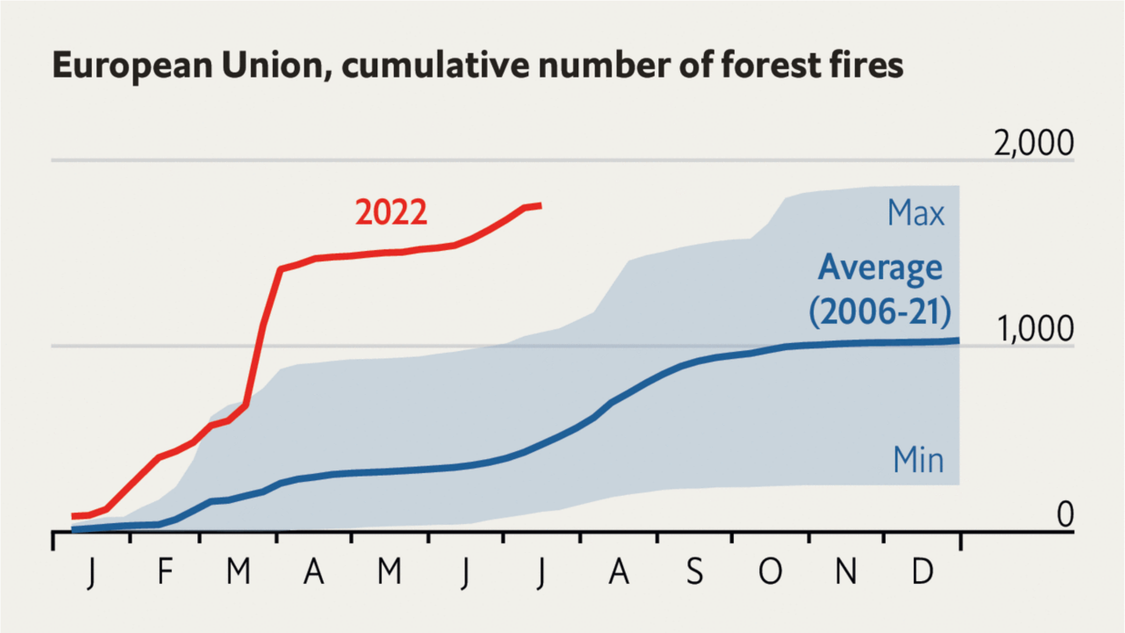


Figure 1. A comparison in the cumulative number of forest fires between the average of 2006 to 2021 and 2022. Source: The Economist, July 18th 2022

Especially in populated areas, wildfires have become a devastating threat to people and the environment. Many regions that naturally experience a hot and dry climate have seen record-breaking numbers of wildfire occurrences and total area burned, as can be seen in Figure 1. For example, in 2022, Spain endured the most wildfires and area burned since data collection by the European Forest Fire Information System (EFFIS) began in 2006. In order to prevent and combat wildfires more effectively, various studies into the key drivers of wildfires have been conducted in recent years (Jain et al., 2020). A specific area of interest within this field is the creation of spatial wildfire susceptibility maps that predict the probability and severity of wildfires in a certain region, often based on a combination of meteorological, geographical and sociological data. Gaining a deeper understanding into spatial wildfire susceptibility and improving our current predictive capabilities has a clear societal relevance and impact. Due to the size of natural lands, it is a difficult task for forestry management groups to continuously monitor the entire area in their jurisdiction and respond quickly to emerging wildfires. Spatial wildfire susceptibility maps allow forestry management groups to anticipate the occurrence of wildfires in high-risk areas according to the model and mobilize their resources in a timely manner. More efficiently combatting emerging wildfires has many potential benefits, such as less deforestation and destruction of nature, less pollution due to forests burning and an earlier evacuation of the high-risk area, which could save many human and animal lives.

Wildfire susceptibility maps were usually created using machine learning methods such as Logistic Regression and later Random Forest classification (Vilar et al., 2010; Su et al., 2021; Reyes-Bueno & Loján-Córdova, 2022). However, recent studies have argued for the suitability of Deep Learning (DL) models, due the unique ability of artificial neural networks to account for spatial and temporal context (Reichstein et al., 2019). At first, the suitability of Convolutional Neural Networks (CNNs) was examined, as this type of Deep Learning architecture is sensitive to spatial-context (Zhang et al., 2019; Huot et al., 2020). Recently, other studies have looked into Recurrent Neural Networks (RNNs), which are sensitive to temporal context, and combinations of CNNs and RNNs in order to create a model sensitive to spatio-temporal context (Prapas et al., 2021; Bjånes et al., 2021). While these studies have introduced an interesting new type of DL architecture to the field of wildfire susceptibility modelling, a few important limitations of the current research should be noted. Firstly, the current generation of context-sensitive wildfire risk models have only been trained to predict the probability of wildfire occurrences, not their severity. It is arguably more important for forestry management groups to be able to accurately anticipate severe fires, than to be able to anticipate all fires without any indication of their potential destructiveness. As such, this is an important improvement that needs to be made before such models can be put into practice. Secondly, the current generation of spatio-temporal models have not outperformed less complex temporal models, so there is currently no evidence for the benefit of taking spatial context into account. Thirdly, these models have not yet been tested in many regions, such as the Spanish biome, which means more research is needed in order to examine their suitability and improve their performance (Prapas et al., 2021).

Therefore, the main aim of this thesis is to create a Deep Learning model that predicts both the probability and severity of wildfires in Spain, taking spatial and temporal context into account. This leads to the following main and secondary question:

1. **How can context-sensitive Deep Learning models be used to predict wildfire occurrence in Spain?**

2. **To what extent are context-sensitive Deep Learning models suitable for wildfire severity prediction?**

These questions have a strong societal relevance, as a deeper understanding of wildfire probability and severity gives forestry management groups more tools to save human and animal lives and to minimize the damage that wildfires cause to nature and the economy. Additionally, this study is highly scientifically relevant as it examines and extends existing Deep Learning methods by applying them to wildfire severity prediction. The application of such models to severity prediction is a novel approach and the first attempt at creating a holistic risk model that looks at both probability of occurrence and expected damage. Moreover, an extensive and highly useful data set is created for this study that combines data gathered by many different satellites and other sources. Through a complex data engineering process, these data sources are harmonized into a spatio-temporal grid map of Spain containing both positive and negative samples that can be used to study many wildfire-related questions. As such, the intended contributions of this thesis are threefold. Firstly, to examine and improve the performance of context-sensitive Deep Learning models for wildfire occurrence prediction in the Spanish biome. Secondly, to introduce a novel approach to the field of burn area prediction and create the first ever holistic wildfire risk model based on context-sensitive DL models. Thirdly, to create and publish a unique data set generated through a complex data engineering procedure that can be used for future research. Within Europe, Spain is the country with the most frequent wildfires and prior studies have focused on other countries in the Mediterranean Basin, which is why Spain has been selected as the study area.

The structure in this thesis is as follows. In Section 2, the current state of wildfire risk modelling is explored and potential extensions highlighted. In Section 3, relevant background information on Spain is discussed as well as the wildfire influencing factors used as features in the model. Furthermore, the complex data engineering procedure used for creating the spatio-temporal grid map of Spain is described, which includes the harmonization of various satellite sources through map reprojection, the inclusion of negative samples and various data cleaning steps. In Section 4, the Deep Learning models used to predict wildfire occurrence will be explained. Moreover, this section details the unique approach used for the creation of predictive wildfire severity models. In Section 5 the results of this study will be discussed using a quantitative approach, however a qualitative comparison to current wildfire risk models will also be made. In Sections 6 and 7, the conclusion and discussion of this study will be provided.

2 Related work

In this section, the current state of academic research on wildfire susceptibility mapping and burn area prediction are discussed. The first subsection is focused on what factors influence wildfires. In this subsection a distinction is made between variables that influence the probability of wildfires and variables that influence the severity of wildfires. The second subsection discusses the methods that have been used in prior studies for wildfire susceptibility mapping. The third subsection discusses the various methods that have been used to predict the total burn area of wildfires. In the last subsection, the intended contribution to the literature made by this research is presented.

2.1 Wildfire influencing factors

Wildfire influencing factors can broadly be put into four categories: meteorological, vegetation-related, topographical and anthropogenic variables (Bjånes et al., 2021; Mhawej et al., 2015). Firstly, meteorological factors are fundamental drivers of wildfire. Extreme weather conditions, such as intense heat, low humidity and high wind speed, can increase the probability of wildfires occurring and help them spread faster, leading to more severe fires. Additionally, precipitation, evapotransporation and climate water deficit (CWD) can play an important role as well (Mann et al., 2016). Secondly, variables related to vegetation are also crucial for wildfire occurrence and severity. Specifically, factors such as vegetation type, density, condition and moisture content are important. These characteristics of the vegetation in a particular area are often quantified by various indices. Land cover data can provide information on the type of vegetation in a particular area, which functions as the fuel of a fire. Studies have shown that the probability and severity of wildfires vary with the type of vegetation. Small fires are more selective in their land cover preferences than large fires. In addition, wildfires have a clear preference for shrublands and grasslands, followed by other forest cover types, while agriculture is avoided (Nunes et al., 2005; Oliveira et al., 2013). The Enhanced Vegetation Index (EVI), Normalized Difference Vegetation Index (NDVI) and Leaf Area Index (LAI) provide information on the moisture content in leaves and health condition as well as the distribution and density of vegetation (Mhawej et al., 2015; Prapas et al., 2021). Other variables that are often used to gauge the moisture content of vegetation are the distance to the nearest stream and the Fine Fuel Moisture Code (FFMC) (Bjånes et al., 2021). Thirdly, topographical factors impact the vegetation, climatic conditions and the rate of wildfire spread of a region, which in turn influence the probability of ignition and total area burned. Elevation, aspect and slope are seen as the most important topographical factors (Oliveira et al., 2013). Oliveira et al. (2013) showed that areas with a slope of more than 25% and a Northern aspect were less susceptible to wildfires. Lastly, anthropogenic factors, such as distance to roads and settlements, land use and wildland-urban interface are crucial to understanding wildfires, as various studies show human activity to be the direct cause of the majority of wildfires (Doerr et al., 2013).

All in all, different studies have identified different specific variables as the key drivers, however they all stem from one of these four categories. The most exhaustive list of wildfire influencing factors was made by Mhawej et al. (2015), who extracted a total of 28 different factors from the literature. An important note is that a fifth category was added, which only contained the variable 'Probability of occurrence of wildfire', defined as the empirical probability when looking at historical data of a wildfire occurring in a certain area.

2.2 Wildfire susceptibility mapping

Wildfire susceptibility mapping has been a topic of academic research since the late twentieth century, with the development of fire danger indices such as the Canadian forest fire weather index system, Keetch-Byram drought index and the McArthur forest fire danger index (Van Wagner, 1987; Keetch & Byram, 1968; McArthur et al., 1967). These and similar fire danger indices were the first attempts at calculating how the risk of wildfires varied over time and place, influenced by natural variables such as temperature, humidity and wind. These methods intended to capture physical variables in a formula that gave a score and associated categorical fire danger label as output. In terms of complexity and granularity, these models have been surpassed by more recent developments, however they are still widely used by wildfire management groups across the globe.

Martell et al. (1987) showed that logistic regression can be used for calculating the probability of wildfire occurrence in Ontario. Two years later, he extended on his own paper by adding a seasonal component to his model in order to account for more frequent wildfires during fire season. In this study of an Ontarian forest, the region of interest was divided into nine areas, each serving as a prediction unit. Subsequently, for each area, various wildfire predictors were gathered and used as input data for the logistic regression. This was the first example of a classification method being used in the field of wildfire prediction (Martell et al., 1989). Vega-Garcia et al. (1995) provided more evidence for the suitability of logistic regression in wildfire risk modelling by using it in a study of Alberta. Ever since then, the research framework of dividing the region of interest into multiple areas that serve as prediction units has become a popular technique in the field. A major benefit of using logistic regression techniques for wildfire risk modelling, compared to the traditional technique at the time, was that the output of such models could be directly interpreted as the probability of wildfire occurrence. This interpretation in combination with the division of the region of interest into smaller tiles makes it possible to create a spatially varied wildfire occurrence probability map, which can be an advantage for the responsible wildfire management group (Martell et al., 1989).

Logistic Regression and related methods have remained a popular tool for wildfire susceptibility mapping. Vilar et al. (2010) used a logistic generalised additive model to estimate daily human-caused wildfire ignition probability at a 1 km^2 resolution focused on an area surrounding Madrid. Su et al. (2021) used logistic regression in addition to geographically weighted logistic regression to find the most important wildfire drivers in two tropical Chinese forests. Reyes-Bueno & Loján-Córdova (2022) employed three methods in their research: logistic regression, logistic decision tree, and multivariate adaptive regression spline. They found the logistic decision tree method to be most suitable and achieved an accuracy of 83% on a balanced data set. In recent years, however, other Machine Learning methods have also started to emerge in this field. Guo et al. (2016) trained a Logistic Regression and Random Forest (RF) model to predict wildfire occurrence in Chinese forests. They found their Random Forest model to perform better. Similar results were found by Milanovic et al. (2020), who also compared these two models and found the RF model to achieve better performance. He et al. (2021) compared three ensemble methods: Random Forest, AdaBoost and Gradient Boosted Decision Trees (GBDT). Their RF model achieved the best performance with an AUC of 0.91. These and other studies showed the suitability of Machine Learning for wildfire susceptibility prediction, however Reichstein et al. (2019) called attention to wildfire ignition and spread as a sub-domain of the earth sciences that stands to benefit from models that are sensitive to spatial and temporal context, specifically Deep Learning models.

Until recently, there have been few studies that have taken a Deep Learning approach (Jain et al., 2020). Zhang et al. (2019) trained a Convolutional Neural Network on weather and remote-sensing data to predict wildfire locations in Yunnan, in order to take advantage of the unique architecture of such neural networks, which makes it sensitive to spatial context (O'Shea & Nash, 2015). Their method outperformed Random Forests, support vector machine (SVM), multilayer perceptron neural network (MLP), and kernel logistic regression benchmark classifiers in terms of AUC. Zhang et al. (2021) extended their own research by comparing two CNN architectures and two Multi-Layer Perceptron architectures for wildfire prediction at a global scale. Their best model was a grid-based CNN-2D model that took a range of 25 by 25 pixels surrounding the pixel of interest to make its prediction, which indicates the advantage of incorporating spatial context into wildfire prediction models. This model was composed of two convolutional layers and three fully connected layers and achieved AUC scores between 0.956 and 0.982. While this study achieved remarkable results, their models did not exploit the temporal context of wildfires. A similar CNN-based method was used by Huot et al. (2020), who approached this problem as an object segmentation task, rather than classification. They attempted to predict a daily or weekly aggregated fire mask based on historical wildfire data from MODIS. Their main U-Net algorithm, which is sensitive to spatial-context, did not perform well on its own, but combining this model with a Long Short-Term Memory Neural Network, which is sensitive to temporal context, significantly improved the model's overall performance. However, since their target outcome was a fire mask rather than a classification algorithm, their predicted outcomes are not probabilities and as such could not be used to create a spatial wildfire probability map. Le et al. (2021) implemented a three-layer deep neural network, which achieved an AUC of 0.894. This model, however, was not context-sensitive as it was not fed multiple time steps or pixels surrounding the pixel of interest. Additionally, Bjånes et al. (2021) criticized the process that was used for sampling no-fire data points, as this was not restricted to forest areas. In their view, this could lead to biased results, as the model might learn to distinguish forest from non-forest ground instead of fire from no-fire data points. In their own paper on this topic, Bjånes et al. (2021) combined two promising Deep Learning architectures to create an ensemble model. Their ensemble model was composed of the spatial context-sensitive CNN architecture proposed by Zhang et al. (2019) and a CNN architecture with multiple convolution heads. The benefit of using multiple convolution heads is that groups of input variables can be processed by separate CNNs before they are combined and processed by a final CNN, which allows for enhanced feature extraction. After experimenting with multiple sample sizes, their model showed the best performance with an area of 25 x 25 surrounding the pixel of interest, which is the same as Zhang et al. (2019) found. Their ensemble model achieved an AUC of 0.953, which implies an exceptional ability to discriminate fire from no-fire samples.

The first attempt at designing a Deep Learning classification algorithm that considers both spatial and temporal context was made by Prapas et al. (2021). Their innovative paper compared the performance of four different classification models on four different data structures, extracted from a data cube consisting of various wildfire predictors over a ten year period across the whole of Greece. The four models they compared were a Random Forest, an LSTM, a CNN and a ConvLSTM. They showed the LSTM model to have the highest Recall and F1-score, while the ConvLSTM had the best AUROC-score. The main contribution of this study was two-fold. Firstly, they used a larger region than any research before them in the field of wildfire prediction. Secondly, they showed the suitability of ConvLSTMs and the benefit of using a model that is sensitive to spatio-temporal context. A limitation of this study is that the models were presented as a prototype and as such were not subjected to a rigorous hyper-parameter tuning procedure, which might indicate that their performance can still be improved. While Prapas et al. (2021) have shown such context-sensitive models to be promising, there has not yet been research that has attempted to improve and extend this method. For this reason, context-sensitive Deep Learning models will be the main estimation method used in this thesis in an attempt to extend their capabilities, improve their performance and show their efficacy in other biomes than Greece.

2.3 Burned area prediction

While post-fire burned area mapping has been studied extensively, studies on pre-fire burned area prediction have been sparse. Additionally, compared to other wildfire domains, the use of Machine Learning methods in studies of burned area prediction has been a relatively recent development (Jain et al., 2020). Cheng & Wang (2008) performed one of the first studies on this topic, however they used a Recurrent Neural Network to predict annual average area burned in Canada, which is quite different than predicting the area burned by one specific fire. Castelli et al. (2015) compared Semantic Segmentation Genetic Programming (SS-GP) to other ML methods using the well-known Montesinho National Park data set. In this study, the prediction was framed as a regression problem of total hectares burned. The SS-GP model achieved the best performance with a Mean Absolute Error of 12.9, while their neural network benchmark performed poorly. In their study, however, they note that the Montesinho data set consists of only 517 samples, which might be too few for a neural network approach. Mayr et al. (2018) compared five statistical and ML methods to predict burned area in Namibia. These methods were: GLM, MARS, regression trees, RF and SVM for regression. The Random Forest model performed best.

To the best of our knowledge, there have only been two studies that have attempted to predict burned area using Deep Learning. Firstly, Liang et al. (2019) compared a Back-Propagation Neural

Network, a Recurrent Neural Network and a Long Short-Term Memory model to predict burned area of specific fires in Alberta, Canada. This study was framed as a multi-label classification problem that used regression as an auxiliary step. Their three models provided an exact amount of hectares burned as the output, which were then binned in five classes ranging from low to severe. The LSTM model performed best and achieved a remarkable accuracy of 90.9% and AUC of 0.942. While this is an excellent performance, the only input data used in their models were weather-based, which leaves room for improvement. Secondly, Lai et al. (2022) used a sparse auto-encoder Deep Neural Network on the Montesinho data set. They framed the study as a regression problem and reduced the MAE and RMSE that was achieved by other state-of-the-art methods.

All in all, there seems to be a relative lack of research on burned area prediction, especially with regards to Deep Learning. As such, there is room for improvement and further research on this topic. Based on the promising results achieved by Liang et al. (2019) with an LSTM architecture, temporal context may be a highly relevant factor for burn area prediction.

All in all, this thesis will combine and extend elements of prior studies in order to introduce a novel approach to the field of burn area prediction by applying a spatio-temporal DL architecture to perform multi-class classification of wildfires into bins of severity.

2.4 Intended contribution to the literature

The work in this thesis builds upon prior research and intends to improve and extend the methods proposed in these studies as well as introduce a novel approach. The intended contribution to the field of wildfire occurrence and severity prediction is five-fold:

1. The suitability of context-sensitive Deep Learning models for wildfire occurrence prediction is examined for the Spanish biome and the performance of these models is improved by further research into the optimal model architecture and wildfire influencing features.

2. Deep Learning multi-class classification models sensitive to spatio-temporal context are introduced to the field of burned area prediction.

3. Two unique and highly useful data sets containing burned areas and their associated wildfire influencing factors are created and published for the period 2010 - 2021 in Spain, which can be used for future research.

3 Data

In this section, the data sets used to answer the first and second research questions are presented. Firstly, the study area and its relevant characteristics will be introduced. Secondly, a description of the target variable and predictive variables is provided for both research questions. Thirdly, a detailed explanation of how the data set is constructed is given, including an overview of how the remote-sensing data was harmonized through map reprojection, an explanation of the grid framework and the manner in which negative data points are sampled. The data engineering procedure is an important and complex component of this research and is thus given due attention. Fourthly, the data is briefly explored to provide some information on the distribution and characteristics of the data sets. Lastly, the last step in the data processing procedure will be described, which includes data cleaning and data imputation, as well as necessary data transformations, such as one-hot encoding categorical variables and scaling.

3.1 Study area

The country of Spain is situated in Southern Europe with small territories across the Mediterranean sea and in the Atlantic Ocean. The study area in this thesis is the main landmass of Spain, otherwise known as Peninsular Spain, which covers an area of 493,514 square kilometres. The landscape of Peninsular Spain mostly consists of highland plateaus surrounded and dissected by mountain ranges. The average altitude is 600 meters above sea level, ranging from 0 to 3,477 meters from its lowest to its highest point. The climatic conditions in Peninsular Spain are extremely varied, as the country can be divided into three major climatic types, namely Oceanic, Continental, and Mediterranean. The majority of Spain is covered by a Continental climatic zone, which is characterized by warm to hot summers, irregular and little precipitation and a hot, dry wind. The northern part of Spain experiences an Oceanic climate, which is associated with lower temperatures and no discernable seasonal variability in rainfall. The south-eastern part of Spain experiences a Mediterranean climate, which is characterized by high diurnal and seasonal variability in temperature, low rainfall and high evaporation, which leaves the land arid. While the summers can be extremely hot and dry, with average daytime temperatures upwards of 30° Celsius, winters are cold with strong wind and low rainfall, but high humidity (Carr, 2023).

Forests and other natural lands cover approximately 30% of the land, which corresponds to 15 million hectares. Vegetation is mainly made up of leafy evergreens, however in the dry parts of Spain scrublands and coniferous forests can also be found. Compared with other countries in Europe, the frequency of wildfires is high. It has a strong seasonal and spatial pattern, as wildfire hotspots are mainly concentrated in the dry southern part of Spain during fire season, which typically ranges between mid-July and September (Moreno et al., 2011).

The total population of Spain is 47.2 million as of January 1st 2022, which includes the population of non-mainland Spain. The average population density is lower than other Western-European countries with 91.4 inhabitants per square kilometer, however it has the highest real population

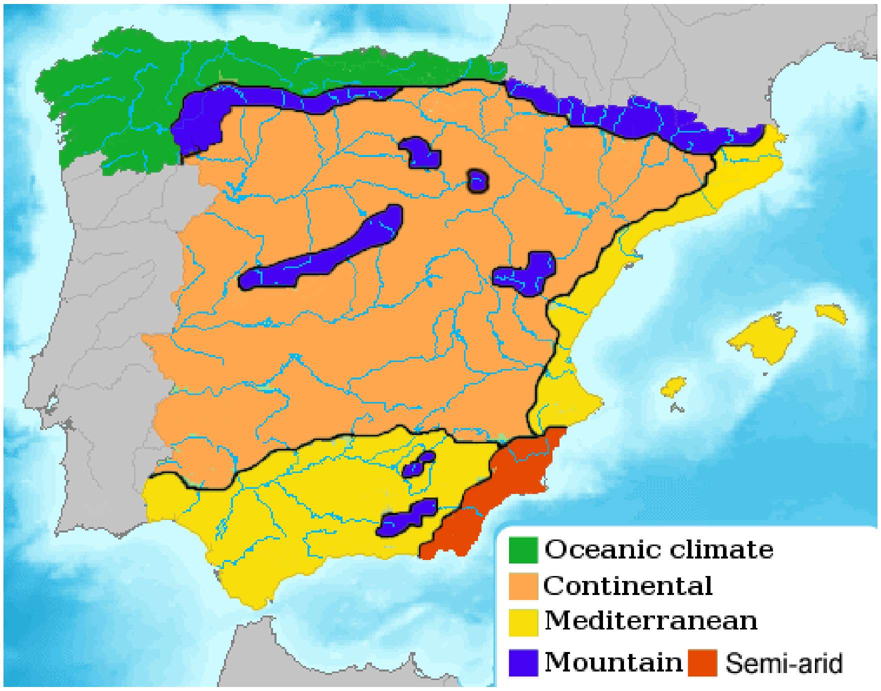


Figure 2. The climatic zones of mainland Spain, including the semi-arid sub-climate. Source: FDV CC 4.0 International, no changes made

density, which only takes inhabited areas into account. By nominal GDP, Spain is the fifth-largest economy in Europe and the fourteenth-largest in the world, yet it has the highest unemployment rate in Europe with 12.6 % as of January 1st 2022 (Carr, 2023).

3.2 Variables

3.2.1 Target variable

As the first research objective of this thesis is to create a spatial wildfire susceptibility map, the target variable is the location and date of wildfire occurrences. This data is gathered from the NASA Fire Information for Resource Management System (FIRMS), which provides open-source data on active and historical fires across the globe. The data source used in this thesis is the Moderate Resolution Imaging Spectroradiometer (MODIS), which is a key instrument aboard the NASA satellites Terra and Aqua that acquire data in 36 spectral bands. Specifically, the historical data set of active fires in Spain from 2010 to 2021, gathered by the MCD14ML Active Fires Product, is used. The MCD14ML Active Fires product records the latitude and longitude, acquisition date and time, confidence measure and Fire Radiation Power (FRP), which is a measure of the intensity of the fire, in addition to a few auxiliary variables. The confidence value ranges between 0% and 100%, where 0 to 30 means low confidence, 30 to 80 nominal confidence and 80 to 100 high confidence. The temporal resolution is approximately 12 hours, which means the acquisition date can generally be regarded as the start date of the fire. The detection of active fires is automated, using a proprietary Machine Learning algorithm that exploits sudden changes in reflected radiation from vegetation and the strong emission of mid-infrared radiation from fires (Giglio et al., 2016). Although there are some known issues regarding cloud obscuration, lack of coverage or misclassification of land and sea, the MCD14ML Active Fires product is the main source of information on wildfire location used in the majority of (recent) studies. (Zhang et al., 2019; Prapas et al., 2021; Ghorbanzadeh et al., 2019; Huot et al., 2020)

The extension on wildfire susceptibility maps proposed in this thesis is an algorithm for burn area prediction, which means for the second research question the target variable is hectares burned. This data is extracted from a data set that contains a collection of all fires in Spain between 2001 and 2015 that burned 1 hectare of land or more. This data set was created by non-profit data journalism organization Civio. Civio harmonized and processed a data set created by the Spanish General Statistic of Forest Fires (EGIF) and the Coordination Centre of National Information about Forest Fires (CCINIF). They performed some data cleaning and outlier detection and filtered only the largest 35% of fires, which were responsible for 98% of area burned.

3.2.2 Predictive variables

As described in Section 2, wildfire influencing factors can be put in four general categories: meteorological, vegetation-related, topographical and anthropogenic variables. Based on the literature review and availability of data sources, 19 variables have been selected.

The meteorological variables are: average temperature of the air at two meters above ground level, maximum temperature of the air at two meters above ground level, wind speed in eastward direction at ten meters above ground level, wind speed in northward direction at ten meters above ground level, thermal radiation emitted by the surface of the Earth, total evaporation and total precipitation. These variables are gathered from the ERA5-Land data set, which was produced by the Copernicus Climate Change Service (C3S) of the European Commission. The spatial resolution is 0.1° x 0.1°, which is approximately 9 km^2. The temporal resolution is hourly (Muñoz Sabater et al., 2021). The vegetation-related variables are: Build-Up Index, Fine Fuel Moisture Code, Drought Code, Normalized Difference Vegetation Index, Enhanced Vegetation Index and land cover. The Build-Up Index, Fine Fuel Moisture Code and Drought Code are also gathered from the ERA5-Land data set and come in the same hourly 0.1° x 0.1° resolution as the meteorological data. The NDVI and EVI are collected from the MODIS Vegetation Index Products MOD13C2 and the land cover data from the MODIS MCD12Q1. Both of these MODIS products have a spatial resolution of 500 m x 500 m. The category of topographical variables consists only of elevation. This variable is gathered from the Copernicus Land Monitoring Service Digital Elevation Model. The resolution is 25m. Lastly, the anthropogenic variables are: population, unemployment rate, road density and a dummy variable that signifies whether the tile of interest contains a human settlement. Population can be gathered from WorldPops with a spatial resolution of 0.1° x 0.1°. Unemployment rate can be gathered yearly from the International Labour Organisation (ILOSTAT) database. Road density and Wildland-Urban Interface are gathered from the OpenStreetMap API. An important note regarding the data gathered from the OSM API is that no historical archived data sets could be recovered, so instead the data at the time of writing has to be used across all dates. Consequently, this could introduce errors in the data if roads or settlements have been built in the meantime, however this data quality issue will not introduce enough noise to significantly disturb the training of the models.

Category	*Variable*	*Unit*	*Resolution*	*Source*
Meteorological	Temperature of the air	Degrees Kelvin	0.1°	Copernicus ERA5-Land
	Maximum temperature of the air	Degrees Kelvin	0.1°	Copernicus ERA5-Land
	Eastward wind speed	Meters per second	0.1°	Copernicus ERA5-Land
	Northward wind speed	Meters per second	0.1°	Copernicus ERA5-Land
	Land Surface Temperature	Degrees Kelvin	0.1°	Copernicus ERA5-Land
	Evaporation	Meter water equivalent	0.1°	Copernicus ERA5-Land
	Precipitation	Millimeter	0.1°	Copernicus ERA5-Land
Vegetation	Build-Up Index	None	0.1°	Copernicus ERA5-Land
	Fine Fuel Moisture Code	None	0.1°	Copernicus ERA5-Land
	Drought Code	None	0.1°	Copernicus ERA5-Land
	Ignition Code	None	0.1°	Copernicus ERA5-Land
	NDVI	None	500m	MODIS MOD13C2
	EVI	None	500m	MODIS MOD13C2
	Land cover	Categorical	500m	MODIS MCD12Q1
Topographical	Elevation	Scaled meters	25m	EU-DEM
Anthropogenic	Population	People	0.1°	WorldPops
	Unemployment rate	Percentage	Nationwide	ILOSTAT
	Road density	Number of roads	Approx. 10m	OpenStreetMap
	Presence of Wildland-Urban Interface	None	Approx. 10m	OpenStreetMap

Table 1: All variables used to create the two data sets and their characteristics

3.3 Building the data set

3.3.1 Grid framework

As will be further explained in Section 4, four different models will be used to answer the two research questions, namely a baseline Random Forest model and three different Deep Learning architectures. These four models require four different data modalities dependent on which context they are sensitive to. The creation of the data set that can accommodate these four different data modalities is a complex and time-consuming process. The baseline model uses no context, which means the data it requires is simply the data associated with the tile in which the fire occurred one day prior to the day it occurred. The Deep Learning models all use different contexts, which means they are fed the data of tiles in different times and locations in addition to the tile in which the fire occurred. The temporally-sensitive model uses a time-series of data from the tile of interest for a period of 7 days as input. The spatially-sensitive model uses the data from all tiles in a 7 by 7 grid surrounding the tile of interest as input. The spatio-temporal model uses a time-series of the data from all tiles in a 7 by 7 grid surrounding the tile of interest for a period of 7 days. For all predictive variables, the data points within the cell are harmonized and re-sampled into (approximately) 1 km × 1 km × 1 day cells. With regards to the continuous variables, such as total precipitation and population, all data points that fall within the cell are averaged for that date, except for the maximum temperature. With regards to the categorical variables, such as land cover class, the mode is taken within the cell for that date. Figure 3 shows an illustration of the grid framework used to build the data set.

An important note is that the first research question is concerned with next-day predictions, which means all variables with a daily temporal resolution should be shifted back one day in order to reflect this data not yet being available. However, the only variables that have a daily temporal resolution are the meteorological variables and in this study the assumption is made that in practice the actual weather values can be substituted for day-ahead forecasts, which means no data needs to be shifted back. As such, the date of each sample in the instance-based and spatial data modality and the seventh date of each sample in the temporal and spatio-temporal data modality is the same date for which the prediction is made.

While the gathering of predictive variables is identical across the first and second research questions, the gathering of the target value differs. For every wildfire occurrence in the target data set for the first research question, a target value of 0 is assigned to the cell if no fire occurred within the boundaries of the cell on that date and 1 if one or more fires did occur. For every wildfire occurrence in the target data set for the second research question, a class in the range from 1 (low) to 5 (high) is assigned based on the total hectares burned by that fire. If a fire's total hectares burned falls below the 20$^{\text{th}}$ percentile it is assigned a 1, if it falls in between the 20$^{\text{th}}$ and 40$^{\text{th}}$ percentile it is assigned a 2 and so on. This results in ranges that are not equal in length and may not have a direct policy interpretation, however it guarantees an equal division of the samples into bins. This is important, because it guarantees that there are enough samples in each class for the models to learn from and it ensures a balanced data set, which makes evaluating the models' performances through metrics such as accuracy more straight-forward.

3.3.2 Map projections

All variables, except for unemployment rate, are in a gridded data-format, which means every sample is indexed to a specific time and geographical location. However, the various data sources use different map projections to geographically index their data, which means an important intermediary step in building the data set is reprojecting all data into a single coordinate system.

The raw data can be extracted from the different data sources in one of the following three coordinate systems: WGS84 projection, sinusoidal projection and EPSG:3035 projection. The WGS84 projection is the well-known latitude-longitude projection that is used in most GPS-systems. On the contrary, the sinusoidal projection is a far less common map projection, which preserves the correct relative surface area of regions. Shapes, angles and distances are generally distorted in this projection. Figure 4 shows this map projection. EPSG:3035 is a coordinate system used mainly in Europe, which uses meters as unit rather than degrees. In order to harmonize these three coordinate systems, Python library Rasterio was used, which is able to easily map geographical data from one coordinate system to another. Except for the data that was already indexed with this

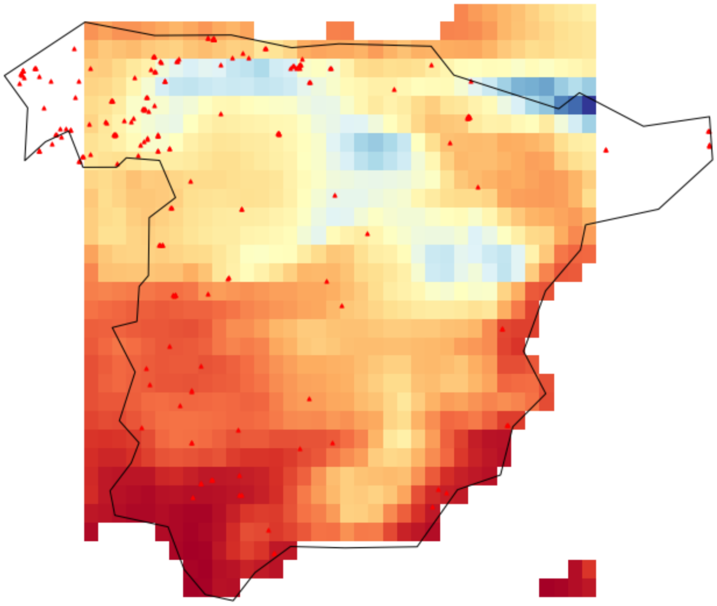


Figure 3. An example of the grid structure for a specific date and variable. The data plotted is total precipitation, where red signifies a high value and blue a low value. The red markings signify forest fires that occurred on that date. The size of the cells is increased for illustrative purposes.

coordinate system, all data was re-projected to WGS84.

3.3.3 Sampling no-fire samples

The second research question is concerned with predicting the severity of fires using data up to the day they ignited. As such, no negative examples are needed to train this model. On the other hand, answering the first research question does require no-fire samples, which means these must be randomly sampled. Prior studies have generally not aligned in their approach to sampling no-fire data. Overall, the goal is to reduce the likelihood of the model learning trivial mappings between input and output and to prevent biases in the composition of the data set. From the perspective of these two goals, there are three main points of contention in the way prior studies sampled no-fire data.

The first point of contention is the spatial distribution of the negative samples. Some studies simply randomly sample no-fire pixels from anywhere within their study area (Huot et al., 2020; Zhang et al., 2019). However, as Prapas et al. (2021) notes, this may lead to trivial mappings if many negative samples are not even forest or other natural land. Based on this input, a model could learn the trivial mapping between a high population and few wildfires when many of the negative samples are taken from urban areas, which has both of these features. In the literature, two ways to deal with this problem have been described. The first is stratifying the set of negative samples to have the same proportion of land cover classes as the set of positive samples. The second is to sample each negative sample from an area in close proximity to a positive sample (Bjånes et

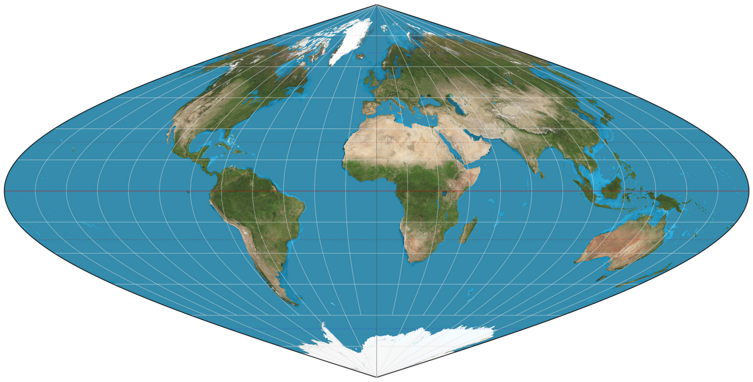


Figure 4. The Sinusoidal projection is used by the MOD12Q1 data set, from which land cover data is retrieved for each tile.

al., 2021; Prapas et al., 2021). Both of these solutions achieve the same goal of preventing biases in the data set and making it harder for the model to differentiate positive and negative samples based on trivial mappings, which forces it to extract more complex features from the data.

The second point of contention is the temporal distribution of the negative samples. Randomly sampling from the entire time-frame of your available data, which is the approach taken by Huot et al. (2020), could overestimate the performance of the model in the real world as information on changing climatic trends are implicitly used to predict wildfires on dates when this data was not yet available (Prapas et al., 2021). For this reason, most papers view wildfire occurrence prediction as a forecasting task, which means that all samples in the test set are taken from a later moment in time than those in the training set. This could be achieved by, for example, constructing the training set with all data from 2010 to 2018 and the test set with all data later than 2018.

The third and last point of contention on the topic of sampling is the ratio between positive and negative samples. As there are many more pixels on a given date that do not experience a wildfire than there are pixels that do, all studies have faced the same problem of a highly imbalanced data set. One solution presented in the literature is oversampling positive data points by assigning pixels in the close proximity of a positive data point a 1 for their fire-values, even if there was no fire recorded in that pixel on that date (Zhang et al., 2019). However, this procedure is not a popular choice, since most studies simply limit the number of negative data points sampled to a ratio of 1 to 1 or 1 to 2. (Le et al., 2021; Bjånes et al., 2021; Prapas et al., 2021)

In this study, these three problems will be solved in the following way. Firstly, the no-fire pixels will be sampled by selecting a random location from the target variable data set and adding or subtracting a random number between 0° and 0.1°. This ensures that the no-fire samples will generally lie in the same area as the fire samples, which means they will have a similar stratification of land cover classes. Secondly, the first research question will be framed as a forecasting task, which means the training set must chronologically strictly precede the testing set. Thirdly, no use will be made of oversampling and the ratio of fire to no-fire samples will be 1 to 1.

3.4 Data exploration

In the following subsection, the context-less data sets used to answer the first and second research question are explored. In this data set every sample consists of only one tile, whereas each sample in the other three data set use respectively 7, 49 and 343 tiles. A few notable elements of the summary statistics and the distributions of the variables are commented on, while an example of a few samples, the full table of summary statistics and the distributions for all the remaining variables can be found in Appendix A for both research questions.

Wildfire occurrence prediction Table 5 in the appendix shows the mean, standard deviation, minimum, median and maximum of the continuous variables, which means the categorical variable landcover and the binary variable wildland-urban interface have been excluded. A few observations can be made from these summary statistics. Firstly, the mean value for the binary target variable is approximately 0.5, which shows it is a balanced data set. Secondly, the mean and median precipitation are almost zero, as there are many days without any rainfall, which stems from Spain's dry climatic conditions. Thirdly, the anthropogenic variables road density and population have a low median and an extremely high maximum, which reflects the fact that Spain consists of large areas of desolate, rural land interspersed with extremely highly populated urban areas, such as Madrid, Barcelona and Sevilla. Fourthly, the Drought Code, Build-Up Index and Ignition Code show a similar low mean and median combined with a high maximum, which suggests that while the overall wildfire risk based on these factors might be low, it can certainly reach critical levels in highly localized areas.

In order to gain insight into the difference in distributions for the predictive variables between fire and no-fire samples, Figure 5 shows the distribution of the fire and no-fire samples for some notable continuous variables. The first two plots show that the fire samples more often have a higher average and maximum temperature, which is reflected in the next three plots, as the fire samples show more evaporation and thus a lower NDVI and EVI, which are both associated with vegetation moisture. This leads to the fire samples having a slightly higher average drought code and ignition code, as can be seen in the last two plots.

Wildfire severity prediction The context-less data set used to answer the second research question shows similar characteristics as the first data set. However, an important difference is that this data set does not contain negative samples. Intuitively, one would expect the second data set to contain values more associated with wildfire risk, as it contains only positive examples. However, while this difference is reflected in some variables, such as a lower average NDVI and EVI and a higher Fine-Fuel Moisture Code, it is not reflected in others, such as a slightly lower average temperature and a slightly higher average precipitation. This is further evidence for the notion that wildfires are complex systems that require modelling the interaction between variables, rather than simply looking at, for example, high temperature, drought and wind speeds.

Figure 6 shows the distribution of the target variables into the five classes based on the range

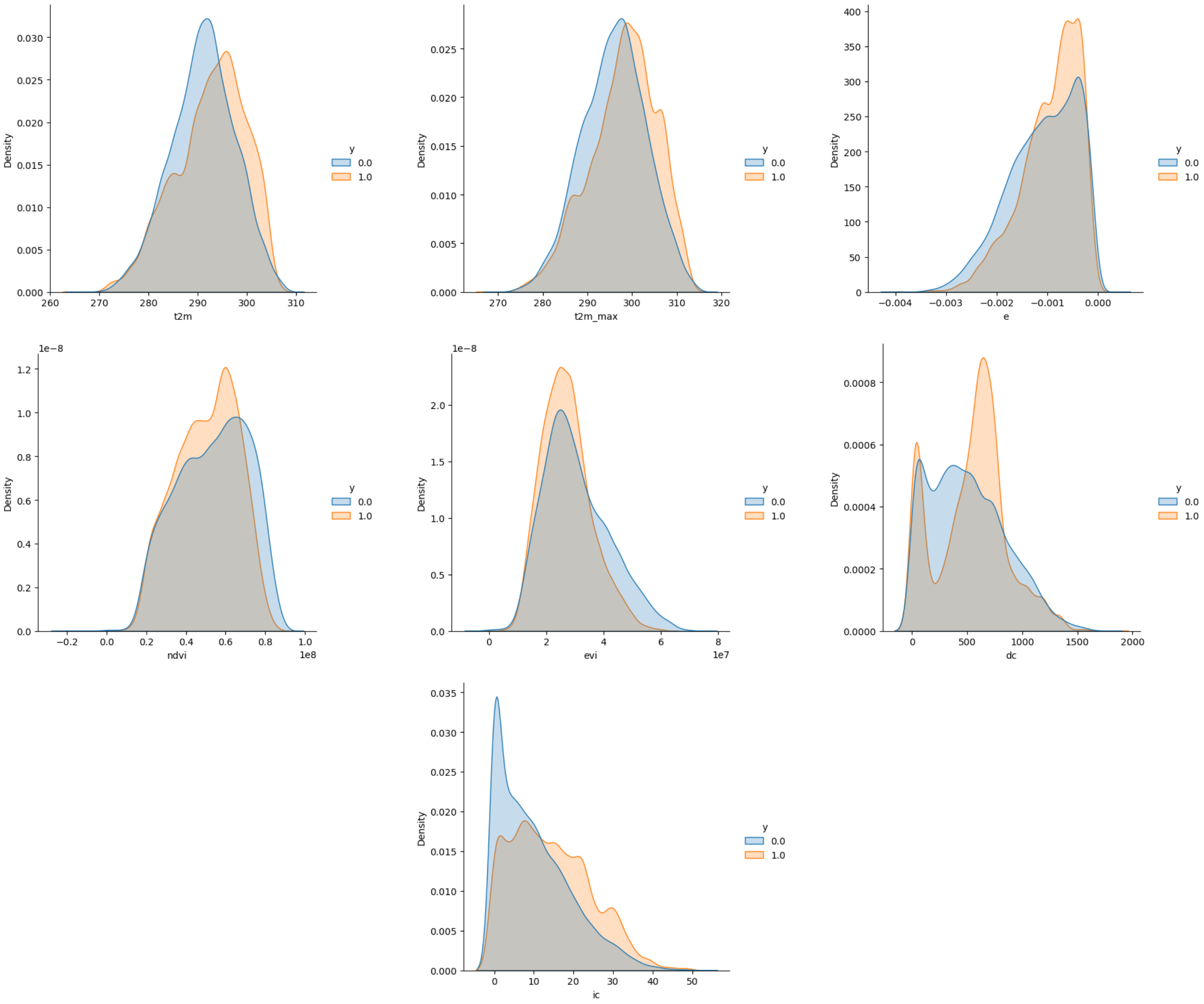


Figure 5. The distribution for fire and no-fire data points for respectively temperature, maximum temperature, evaporation, NDVI, EVI, drought code and ignition code in the baseline data set for the first research question.

of area burned. In the left plot the fifth class is capped at 20 hectares burned, while the right plot shows the remaining samples that exceed this limit, which is about half of the total samples in this class. As can be seen in the figure, there are more small wildfires than large wildfires, which results in the first few classes having a smaller range, as each class contains approximately 20% of the total number of samples. Another remarkable element of this data set is the extremely large value and deviation from the mean that some rare wildfires exhibit. While the lower bound for the highest class lies at approximately 10 hectares burned, there are many examples of wildfires that resulted in multiple thousands of hectares burned.

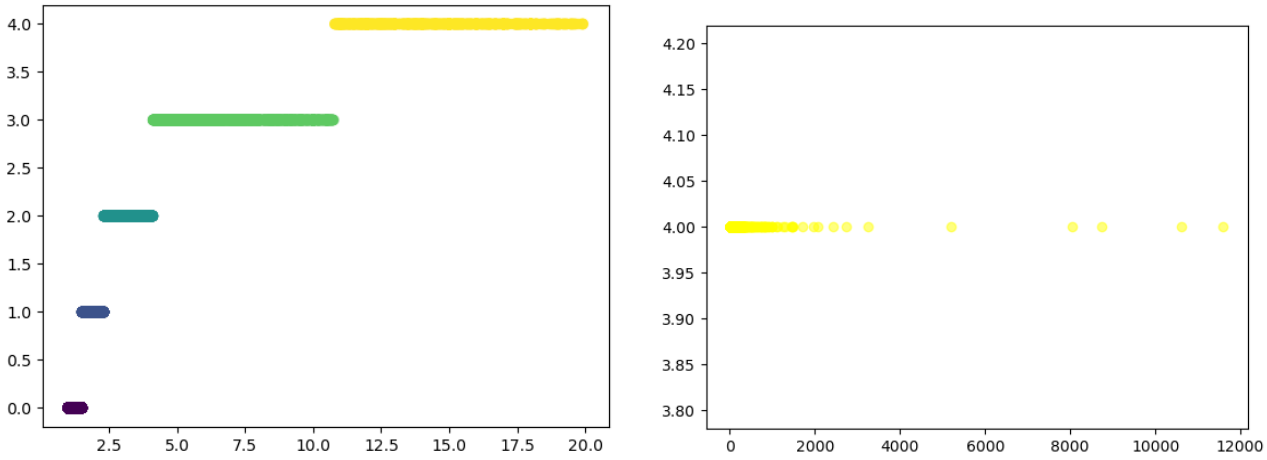


Figure 6. The left plot shows the value of the target variable for all samples in the data set, limited at 20, used for the second research question. The right plot shows the remaining samples that exceed 20 hectares burned. The color shows to which class each sample belongs, which corresponds to a certain range in hectares burned. Each class contains approximately 20% of the total number of samples.

3.5 Data Processing

After the data set was built, three final data processing steps were completed: imputation, scaling and one-hot encoding categorical variables. In the following three paragraphs, these processing steps are discussed individually.

Imputation Neural networks and Random Forests both do not have the capability of dealing with missing values in training or testing data. However, missing values in data is a common problem, especially when data is gathered through physical measurement systems, such as those aboard the satellites that were used in this thesis. For this reason, a necessary data processing step is to perform data imputation. The purpose of data imputation is to replace missing values in the data set with plausible values. This allows the data point to be used for training or testing, rather than having to be discarded. The data imputation method used was to substitute the missing value by the closest non-missing value. If a missing value was surrounded by non-missing values on both sides, it was substituted by the average value of these two non-missing values.

Scaling While neural networks are theoretically scale invariant, there are some practical benefits to scaling your data set. The main issue with using the raw data set stems from the fact that large input values for a specific variable may result in a model that has learned large weight values. A model that uses large weight values is extremely sensitive to small differences in input values, which can result in a highly unstable model. Additionally, the instability can cause the model to have difficulty finding the global optimum, making it potentially slow to optimize during the training phase. For this reason, scaling methods, such as standardization and normalization, shift and rescale the data in order to make the input variables consist of small values. The scaling method used in this thesis is standardization, which substitutes each value for its Z-score:

$$z_{ij} = \frac{(x_{ij} - \mu_j)}{\sigma_j}$$

In this formula, z_{ij} and x_{ij} denote the Z-score and raw value for the ith element of column j. Moreover, μ_j and σ_j denote the mean and standard deviation for column j. This transformation was applied to all continuous variables in the data set.

Categorical variables The last data processing step is to one-hot encode the categorical variables in the data set. This means a binary variable is created for every category of every categorical variable present in the data set. In the data set used in this thesis, the only categorical variable is land cover, which had 11 distinct categories. If a row in the raw data set had a value of, for example, 6 in the column land cover, that row in the processed data set now has a 1 in the column that signifies land cover category 6 and a 0 in all other land cover columns.

4 Methodology

In this section, the methods used to answer the main and secondary research questions will be described. Firstly, a detailed problem description for both questions will be provided, accompanied by an explanation of the general approach suitable to these problems. Secondly, the models used in this research will be introduced. For each model, the motivation for choosing this model will be provided in addition to an explanation of its theory and the specific algorithm structure used in this research. Thirdly, the practical implementation of these models will be discussed, which includes the hyper-parameter tuning procedure for each model, cross validation and the relevant metrics used to evaluate the models' performance. Additionally, a brief overview will be given of the code and libraries that were used to gather the data and implement the models.

4.1 Problem description

Due to the often enormous scale of natural lands that are susceptible to wildfires, wildfire respondents have to efficiently allocate their resources. Both in terms of monitoring capabilities as well as ensuring a timely mobilization of fire extinguishing resources, wildfire respondents would stand to benefit massively from knowing the future location of imminent wildfires. Firstly, this would allow them to focus their monitoring capabilities on a specific sub-region of the area they are responsible for and catch fires early. Secondly, this would allow them to anticipate on wildfires igniting in a specific area and preemptively allocate their resources, such as fire fighting personnel and water tanks, in a more efficient manner. Additionally, being able to predict the severity of ongoing wildfires could also be highly useful, since priority can be given to wildfires that have high destructive potential, especially given the exponential growth over time that wildfires exhibit (Juang et al., 2022). In short, information on the probability and severity of future wildfires could form a key component in improving existing wildfire response procedures.

Wildfire occurrence prediction Quantitative wildfire risk models currently in use are often based on purely meteorological predictors. For example, the Fire Weather Index, which is used globally, takes temperature, humidity, precipitation and wind speed as input to provide a categorical level of fire danger as output. Other indices, such as the U.S. Forest Service National Fire-Danger Rating System and the McArthur Mk5 Forest Fire Danger Meter function similarly. In recent years, various studies have shown promising results for the use of Machine Learning and specifically Deep Learning for wildfire occurrence prediction. The main benefit as opposed to traditional methods is a higher accuracy due to the possibility to include other categories of predictive variables and the high capacity for interaction effects between predictors. An additional benefit is that the output of such predictive models is a probability between 0 and 1, which can be interpreted as the level of fire danger directly. The aim in this thesis is to find whether context-sensitive Deep Learning models can become the next generation of quantitative wildfire risk models. Specifically, an answer will be sought to whether this technique shows promise and what type of context and neural network architecture are most suitable. This leads to the main research question in this thesis: **How can context-sensitive Deep Learning models be used to predict wildfire occurrence in Spain?**

Wildfire severity prediction Viewing wildfire risk as simply the probability of a wildfire occurring does not accurately reflect the real world. In reality, the overwhelming majority of land is burned by a small minority of forest fires, which suggests that a robust risk model should incorporate severity as well. For this reason, the secondary aim in this thesis is to study whether the proposed methods for the main research question can be extended to predict wildfire severity as well. Thus, the secondary research question is: **To what extent are context-sensitive Deep Learning models suitable for wildfire severity prediction?**

4.2 Approach

Wildfire occurrence prediction The general approach to answering the main research question is to create different context-sensitive Deep Learning models that can predict the probability of next-day wildfires in different times and locations. The output of these models can then be directly interpreted as the fire danger, which means they could substitute existing methods, such as the FWI.

As the goal is to predict wildfire probability in different places and times, the grid framework introduced in Section 3 is used, which discretizes the spatial and temporal dimension of Spain into a grid consisting of 0.1° x 0.1° x 1 day cells, which is approximately equal to 1 km x 1 km x 1 day. Each cell contains wildfire-influencing factors and a binary value of either 1 or 0, signifying whether a fire occurred anywhere within that cell on a specific day. The class of predictive models suitable to this problem are known as binary classifiers. Binary classifiers take input data and provide a discrete predicted outcome of either 1 or 0. For both of these classes, the classifier calculates a continuous value bounded between zero and one, which can be interpreted as the probability of the event associated with that class to occur, and then chooses the class with the highest probability is the final output. Mathematically, a model F is trained that takes $x_{t'<t}$ as input and gives $y_t = p(E_t|x_{t'<t}) \in (0,1)$ as output for any given cell. In this equation, $x_{t'<t}$ are the predictive variables prior to time t, y_t the probability of an event occurring within the cell at time t and E_t the event occurring at time t, which in this case is a wildfire. Note that in this specification the input data has a temporal element. While this is not universal for all classifiers, it does not change the fundamental properties of the algorithm. In this study, three context-sensitive models and one baseline model are trained and compared. This will allow us to examine if and to what extent context-sensitive Deep Learning models have added value for wildfire occurrence prediction and which context and type of architecture is most suitable. The baseline model in this thesis is a Random Forest Classifier and the context-sensitive models are respectively a Long Short-Term Memory Model, a Convolutional Neural Network and a ConvLSTM.

Wildfire severity prediction The general approach to answering the secondary research question is to extend the models used for the main research question to predict wildfire severity and subsequently compare their performance. The output of the models in this task will not be a probability between 0 and 1, but instead a number of total hectares burned. However, due to the large range in possible output values and the highly stochastic nature of wildfire spread, a multi-class classification approach will be used rather than a regression approach to predict in what range the total hectares burned per fire falls. Mathematically, a model F will be trained that takes the same x_i input as for the main question and gives $y_i = \arg\max_k P(k|x_i) \in [1, ..., k]$ as output. In this equation, y_i is the predicted output class of sample i and $P(k|x_i)$ the probability of sample i belonging to class k.

In order to adjust the context-sensitive Deep Learning models to provide a multi-class output, three major changes must be made to the architecture of the models. Firstly, output nodes must be added, because for any classification task using neural networks, the number of output nodes is the same as the number of output classes. Secondly, the loss function must be changed from binary cross-entropy loss to categorical cross-entropy loss. Thirdly, the activation function in the output layer must be changed from sigmoid activation to softmax activation. Additionally, other changes to architecture and hyper-parameters can be made, however these changes are beneficial to the model's performance rather than crucial for its function.

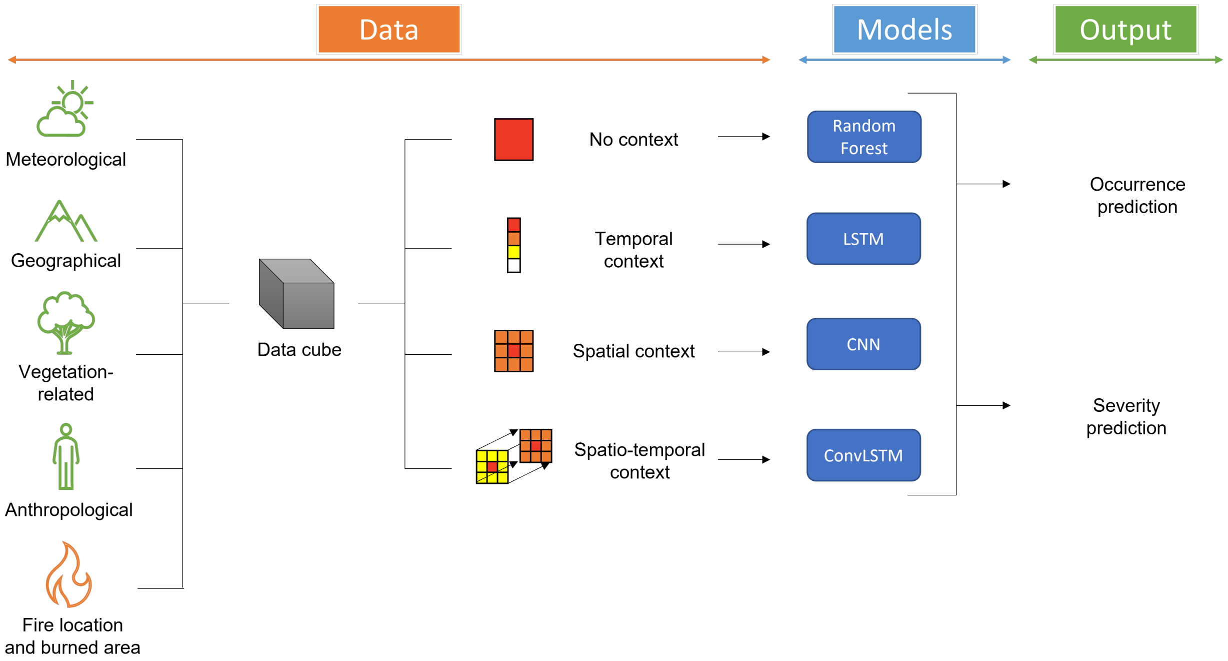


Figure 7. The general flow in this thesis from data to output. Four different data modalities are created from the four categories of variables. These are fed into three context-sensitive neural networks and one instance-based baseline Random Forest. All four models are trained in two different tasks and the performances are compared.

4.3 Models

In the following section, four different models will be discussed that will be used for both the occurrence and severity prediction. Firstly, a Random Forest as a baseline predictive model. Secondly, a Long Short-Term Memory, which is a specific type of Recurrent Neural Network. Thirdly, a Convolutional Neural Network. Lastly, a Convolutional Long Short-Term Neural Network. The three Neural Network approaches are sensitive to respectively temporal, spatial and spatio-temporal context, which potentially gives them an edge over the baseline Random Forest model. For each model, the motivation for choosing this model, its theory and the architecture chosen in this study will be explained. Within the explanation of the model architecture, a clear distinction will be made between the architecture used for occurrence prediction and the architecture used for severity prediction.

4.3.1 Baseline model: Random Forest

Motivation Random Forests are an ensemble method, based on combining the predictions of multiple decision trees. They can be applied to both classification and regression and are widely used to deal with tabular data. Their relative ease of use and generally good performance for many different tasks make them a logical choice to function as a baseline model.

Theoretical background A Random Forest is a supervised-learning algorithm that can be applied to tabular data. The model consists of a collection of individual Decision Trees. Decision trees operate through recursive binary splitting, which is where the a-cyclical, tree-like structure comes from. For every variable, various splits are possible, so the split that results in the lowest loss is chosen. Various loss functions can be used to calculate which split is optimal, however the most popular are the Gini-index and cross-entropy loss. Once the tree stops growing by recursively splitting, which can be controlled through hyper-parameters, the leaf nodes represent a mapping from the input space to the possible output values. Once a new data point is introduced, a single decision tree can provide a prediction for its accompanying target variable. In order to introduce robustness and combat overfitting, the Random Forest algorithm combines the predictions of multiple Decision Trees. The individual decision trees are initialized using an element of randomness, for example in the sub-set of variables it can use, after which the predictions of all Decision Trees are gathered and the final output is derived through majority voting.

Model architecture The Random Forest model architecture and hyper-parameters used for wildfire occurrence prediction is as follows. The number of estimators is 200, each with a maximum depth of 50 and a maximum number of features per tree equal to the square root of the total number of features. The same model architecture and hyper-parameters are used for wildfire severity prediction, since the model automatically deals with the difference in output values.

4.3.2 Context-sensitive Deep Learning models

Firstly, an explanation of the basic components that make up a Neural Network will be provided. Afterwards, the motivation, theory and architecture for each specific model will be discussed.

A Neural Network is a collection of nodes, connected by vertices, that maps an input to an output through the use of weights, activation functions and loss functions. It has been shown that Neural Networks are ‘universal approximators’, which means that they can represent any function $R^n \rightarrow R^m$ to an arbitrarily close degree when given the appropriate weights. This characteristic makes Neural Networks extremely useful for modelling systems in almost any domain.

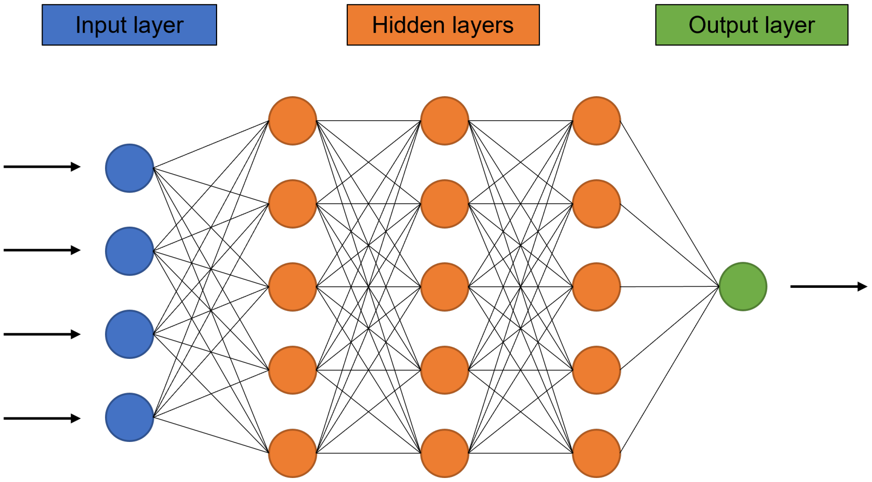


Figure 8. The typical lay-out of a fully-connected neural network, consisting of an input layer, three hidden layers and an output layer with a single node.

Typically, a neural network has layers of nodes. The input layer receives the external data, while the output layer produces the final result generated by the model. Between the input and output layers, the data travels through a sequence of hidden layers. These layers can be ’fully connected’ or ’dense’, which means that every node is connected to every node in the following layer. Additionally, these layers can be ’sparse’, which means that not every node is connected to every node in the following layer, or ’pooled’, which means that multiple nodes are connected to the same node in the following layer. Since such connection structures only allow nodes to be connected to nodes in the following layer, these neural networks are known as feed-forward networks. Networks that allow nodes to be connected to nodes in their own layer or previous layers are known as Recurrent Neural Networks, because in such networks connections between nodes can create a cycle.

As an input enters each node, it is mapped to an output by an activation function, which dictates if that specific node ’fires’, mimicking the workings of our brain. Among the most popular activation functions are linear, sigmoid and Rectified Linear Unit or ReLu, however there are many more examples.

$$\text{Linear: } f(x) = x$$

$$\text{ReLu: } f(x) = \begin{cases} 0 & x \leq 0 \\ x & x > 0 \end{cases}$$

$$\text{Sigmoid: } \frac{e^x}{e^x + 1}$$

The node the input just entered is connected to a different node, often in the next layer, through a vertex that is weighted. The input in the original node is thus not only altered by the activation function, but the output of the activation function is also multiplied by the weight of the vertex before it reaches the next node. As such, the weight of a vertex controls the strength of the signal from one node to the next. The weight of this vertex is updated each time step, which is what it means for a neural network to be 'learning'.

The way a neural network learns, is by minimizing its loss function. The loss function is a measure of how close the model's output is to the true values both in classification and regression problems. For binary classification problems, the most popular loss function is 'binary cross-entropy':

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \cdot \log \hat{y_i} + (1 - y_i) \cdot \log(1 - \hat{y_i})))$$

This loss function is identical to the more general multi-class cross entropy loss function specified for two classes:

$$L_{MCE} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k}^{K}(y_i^{(k)} \cdot \log \hat{y}_i^{(k)})))$$

After each time step, the partial derivative of the loss function with respect to each weight is calculated through back-propagation. Subsequently, the weight of each vertex is increased or decreased in order to minimize the total loss.

4.3.2.1 Long Short-Term Memory Model

Motivation A Long Short-Term Memory Model (LSTM) is a specific type of Recurrent Neural Network, which itself is a sub-set of Artificial Neural Networks. Recurrent Neural Networks have the unique feature that connections between nodes can create a cycle, which allows them to exhibit temporal dynamic behaviour. This characteristic makes RNNs uniquely suitable for analysing time-series in both regression and classification problems. The specific benefit of LSTMs is their ability to avoid the vanishing gradient problem, due to the use of forget gates, which allows LSTMs to learn patterns that require a memory of inputs that happened many time-steps in the past. This characteristic makes LSTMs a popular choice in many time-series regression or classification tasks.

Theory In principle, regular RNNs are capable of learning long term dependencies in the data, however due to the vanishing gradient problem this is not the case in practice. The vanishing gradient problem stems from the fact that certain activation functions map a large input space to a small output space. For example, the sigmoid activation function $e^x/(e^x+1)$ maps $R \to (0,1)$. For small and large x, the derivative of the sigmoid function approaches zero. Since neural networks use back-propagation to train its weights, the derivatives for all layers are multiplied together using the chain rule and may become extremely small, especially in deep neural networks. As the gradient becomes smaller, the weights of nodes can not be updated effectively, which hinders the training of the model. All in all, regular RNNs present a trade-off between being efficiently trainable with gradient descent and holding on to information for many time-steps. Long Short-Term Memory Models, however, deal with the vanishing gradient problem in a way that allows them store dependencies in the data for thousands of time-steps.

At a general level, an individual LSTM cell i can be seen as a black box that takes the current input $x_i^{(t)}$ and its own previous output $h_i^{(t-1)}$ and generates a new output $h_i^{(t)}$. Note that $x_i^{(t)}$ may be the 'pure' data from the input layer or transformed data from the previous hidden layer, dependent on where the cell is located in the network. The core concept that allow LSTMs to accumulate information over a longer time period is the use of cell states and gating units that control the flow of information within a node. The cell state can retain information from many time-steps in the past, which allows the model to identify long-term temporal dependencies in the data. It does so by training four separate neural network structures within each cell that learn what part of the old hidden state and the new data input should be incorporated into the cell state and what part should be forgotten.

A single LSTM node consists of the following components: a forget gate, an input gate, an output gate and a cell state. In 9, the first, second and third sigmoid signs from the left denote respectively the forget gate, input gate and output gate, while the arrow going from left to right along the top of the cell denotes the cell state.

First off, the current input vector $x^{(t)}$ and the cell's previous hidden state or output vector $h^{(t-1)}$ are passed through a forget gate $f_i^{(t)}$ for time-step t and cell i. The forget gate uses a sigmoid layer to output a vector of numbers between 0 and 1, according to the following formula, where b^f, U^f and W^f are respectively biases, input weights and recurrent weights for the forget gates:

$$f_i^{(t)} = \sigma(b_i^f + \sum_j U_{i,j}^f x_j^{(t)} + \sum_j W_{i,j}^f h_j^{(t-1)})$$

The numbers between 0 and 1 in the vector generated by this process represent whether the information in the cell state $C^{(t-1)}$ should be completely forgotten (0), completely remembered (1) or somewhere in between. The reason for this interpretation is the fact that the new cell state $C^{(t-1)}$ is in part calculated by multiplying the old cell state $C^{(t-1)}$ with the forget gate vector $f_i^{(t)}$, so cell state values associated with a forget gate value of 0 are essentially forgotten. This is the crucial self-loop of an LSTM cell, as it updates its cell state in the next time-step based on its own output. Since the weights in this self-loop are set by another hidden unit, the time scale of integrating new information changes dynamically.

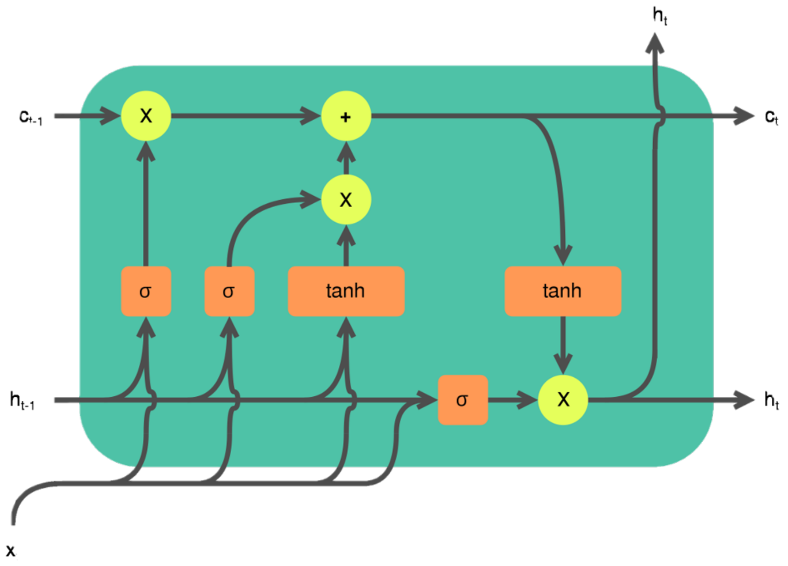


Figure 9. Structural lay-out of a single LSTM cell. Source: The LSTM cell, Guillaume Chevalier

Whereas the first step regulates what information is removed from the cell state or 'forgotten', the second step regulates what new information is added to the cell state. This process consists of two parts: the input gate regulates which state values are updated and a tanh layer generates new candidate values for the cell state. The input gate uses a sigmoid layer similar to the forget gate but with different parameters:

$$g_i^{(t)} = \sigma(b_i^g + \sum_j U_{i,j}^g x_j^{(t)} + \sum_j W_{i,j}^g h_j^{(t-1)})$$

The tanh layer generates candidate state values through the following equation:

$$\tilde{C}_i^{(t)} = \tanh(b_i^C + \sum_j U_{i,j}^C x_j^{(t)} + \sum_j W_{i,j}^C h_j^{(t-1)})$$

The cell state $C^{(t)}$ can now be updated by removing the information that needs to be forgotten and adding new information:

$$C_i^{(t)} = f_i^{(t)} \cdot C_i^{(t-1)} + g_i^{(t)} \cdot \tilde{C}_i^{(t)}$$

The third and last step generates an output $s^{(t)}$ that will function as the new hidden state of the cell. A third sigmoid layer is used to determine what part of the cell state is used, which uses a similar equation to the prior sigmoid layers with different parameters:

$$o_i^{(t)} = \sigma(b_i^o + \sum_j U_{i,j}^o x_j^{(t)} + \sum_j W_{i,j}^o h_j^{(t-1)})$$

The resulting vector is multiplied by $\tanh(C_i^{(t)})$, which results in the following equation for the output:

$$h_i^{(t)} = o_i^{(t)} \cdot \tanh(C_i^{(t)})$$

The new output $h_i^{(t)}$ will then be used as one of the inputs of the cell on time-step $t+1$.

Model architecture The architecture of the LSTM model used for wildfire occurrence prediction is as follows. It starts with two LSTM layers with 32 neurons and 16 neurons respectively using the ReLu activation function. The LSTM layers are followed by one dense fully-connected layer with 8 neurons, ReLu activation and 0.2 drop-out and an output layer with 1 neuron and sigmoid activation. The loss-function is binary cross-entropy. The model is trained over 50 epochs, using the Adam optimizer, a 10^{-4} learning rate and a batch size of 32.

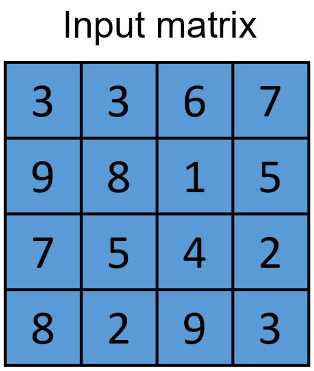
The architecture used for wildfire severity prediction differs slightly. The input and hidden layers are identical, however in order to accommodate multi-class classification, the output layer has five nodes rather than one. Additionally, the loss function has been changed to categorical cross-entropy loss and the activation function in the output nodes has been changed to SoftMax, which is the standard activation for multi-class classification.

4.3.2.2 Convolutional Neural Network

Motivation Convolutional neural networks are uniquely suited to processing data that are structured in a grid. Rather than flattening the grid of data points into a vector, it retains and uses the grid structure and with it the spatial dependencies and structure between data points. This allows CNNs to identify spatial patterns in the data, which makes it a logical starting point for developing a model that is sensitive to spatial context.

Theory Convolutional neural networks are a sub-class of neural networks that make use of kernel convolution in order to retain spatial structures in data ordered in a grid. Such a grid filled with data can be seen as a $W \cdot H \cdot c$ matrix, with W being the width of the matrix, H the height and c the number of 'channels' or variables per cell. Figure 10 shows an example of what a $5 \cdot 5 \cdot 1$ input matrix for a CNN could look like. Convolutional neural networks process such an input matrix using a combination of convolutional layers, pooling layers and fully-connected layers. The purpose of the convolutional and pooling layers is to extract features from the input matrix, while the fully-connected layers perform map these features to a regression or classification output.

The most important element of a convolutional layer is the filter. A filter is a collection of kernels, one for each channel of the input. If the input matrix is an image consisting of three channels (Red-Green-Blue), one filter consists of three kernels. The convolutional layer extracts features from the input matrix using these kernels. A kernel is a set of learnable weights in matrix form that are updated during the training of the CNN. The kernel slides across the height and



Input matrix

3	3	6	7
9	8	1	5
7	5	4	2
8	2	9	3

Figure 10. An example of data in grid from that a CNN can use as an input matrix.

width of the input matrix. At each step, the kernel matrix and the restricted part of the input matrix over which the kernel is hovering, known as the receptive field, are matrix multiplied and the result is stored, as illustrated in Figure 11.

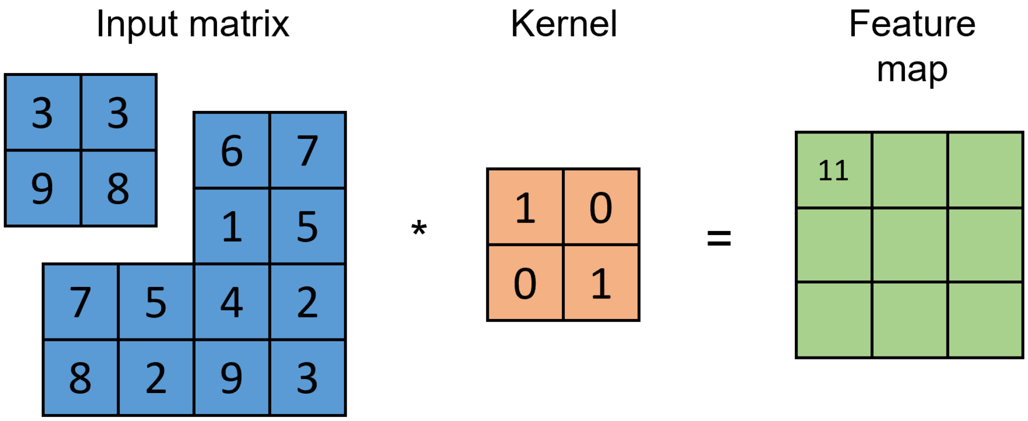


Figure 11. One step in the convolution between the input matrix and the kernel. The receptive field is the top-left quarter of the input matrix.

Once the convolution operation between the kernel and specific receptive field is completed, the kernel slides one or more steps to the right in order to perform the same operation with a different part of the input matrix. The amount of steps it slides over is called the stride, which is a hyper-parameter that can be adjusted. An important note to make is that for a kernel of size $N \times N$, the $N - 1$ outermost edges can not be analysed, since the receptive field would lie partly outside the input matrix. For this reason, an extra layer of data, called padding, can be added to make sure also the edges of the input matrix can be analysed, as illustrated in Figure 12. The data points generated for the padding procedure can be initialized on different values, which also can be treated as a hyper-parameter.

Once the full width of the input matrix has been parsed, the kernel moves down one or more rows, dependent on the stride, and repeats the process starting at the first available column of

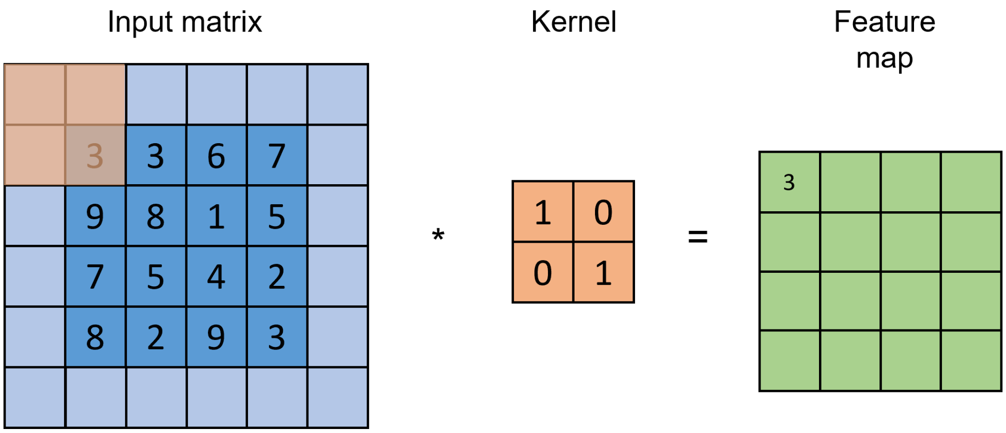
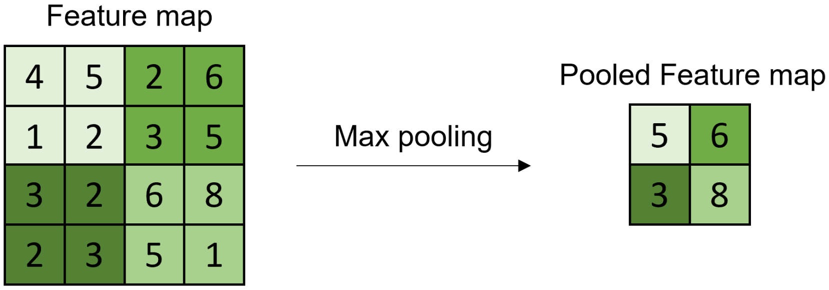


Figure 12. The light blue cells surrounding the input matrix are the data that is added in the padding procedure. As can be seen in the image, this allows the kernel to extract a feature from the outermost corner of the input matrix.

the input matrix, dependent on whether padding is used. This process repeats until the entire image has been traversed. The result of each convolution operation is stored in a new matrix, called the feature map. The feature map is essentially a representation of the original input matrix, where features in the original data consisting of multiple cells are aggregated into a single cell in the feature map. Since the kernel matrix consists of different values, the convolution operation emphasises certain parts of the receptive field, while it de-emphasises other parts. This allows one specific kernel to extract a specific type of feature and store this in the feature map. In a Deep CNN, the complexity of the type of features a convolutional layer can extract increases with the position of the convolutional layer. This means that later convolutional layers extract more complex and detailed features from the input matrix than earlier layers. The weights in the kernel matrix are often randomly initialized and then updated through back-propagation, similarly to normal feed forward neural networks. For each filter, a kernel and accompanying feature map is created for every channel. Right before being sent to the next layer, the multiple feature maps per filter are added up at the end and transformed by an activation function in order to introduce non-linearity in the output, similar to regular feed-forward neural networks. The activation functions used are generally ReLu and Tanh. Since each kernel detects a specific feature, using multiple filters and thus multiple kernels per channel allows a CNN to detect many different features.

Using convolution layers instead of regular fully-connected layers has two main benefits. Firstly, it greatly reduces the memory and computing power needed to train the model. It does so by performing dimensionality reduction on the input data by aggregating groups of cells into features. Additionally, CNNs use sparse interactions between the input and the output units, since each kernel uses the same weights for every part of the input matrix. This allows us to store fewer parameters and it improves the statistical efficiency of the model. The second benefit of convolution layers stems from the fact that the parameters are shared within a layer, which means the location of a feature within the input matrix does not matter for its detection, making CNNs equivariant to translation.

Convolutional layers are often followed by pooling layers. Pooling layers are primarily used to reduce the dimensionality of the feature map created in the preceding convolutional layer, but also serve a purpose for generalisation by removing small, potentially unimportant characteristics. It does so by sliding an area across the feature map, similar to the kernel convolution process, and calculating a summary statistic of the values in its receptive field, such as the maximum value or the mean value. These summary statistics are then stored in a new, smaller matrix, which serves as the input for the next layer. The two most popular pooling methods are 'Max Pooling' and 'Average Pooling'. 13 shows a Max Pooling procedure with a 2x2 filter size and a stride of 2. Whereas Max pooling calculates the maximum value present in the receptive field, average pooling calculates the mean value in the receptive field.



4	5	2	6
1	2	3	5
3	2	6	8
2	3	5	1

5	6
3	8

Figure 13. An example of the transformation the feature map undergoes after it is pooled using MaxPooling. The colors of the areas in the feature map correspond to the cells of the same color in the pooled feature map.

The final necessary component of a CNN is a fully-connected layer, such as the fully-connected hidden layers that can be found in regular feed-forward neural networks introduced earlier. Once the original input matrix has been through a number of convolutional and pooling layers, increasingly complex features are stored in the feature map. The fully-connected layer maps the final feature map to a classification or regression output. In order to provide the feature map, which is in matrix form, to the fully-connected layer, the feature map is often flattened into a vector.

Model architecture The model architecture of the Convolutional Neural Network used for wildfire occurrence prediction is as follows. The first layer is a convolutional layer with 32 filters, a kernel size of 2 by 2 and ReLu activation. Padding of 1 around the edges and a stride of 1 are used. Additionally, an L2 kernel regularizer with learning rate 0.01 was used to penalize complex models, since during the training fase, the model had a high propensity of overfitting. The first convolutional layer is followed by a MaxPooling layer of 2 by 2. The first two layers are followed by a second combination of convolutional and MaxPooling layer identical to the first one, except for the number of filters, which has been brought down to 16. The output of the second MaxPooling layer is flattened and used as input for a dense, fully-connected layer with 8 neurons, ReLu activation and 0.2 dropout. Lastly, the output layer is a dense, fully-connected layer with 1 neuron and sigmoid activation. The model is the compiled and fitted in the same way as the LSTM. The loss-function is binary cross-entropy. The model is trained over 50 epochs, using the Adam optimizer, a 10^{-4} learning rate and a batch size of 32.

4.3.2.3 ConvLSTM

Motivation The ConvLSTM model architecture is a variation on the LSTM model architecture, which replaces the matrix multiplication operation inside the LSTM cell with a convolution operation in both the input-to-state and state-to-state transitions. This difference in internal operation allows the ConvLSTM to use spatial sequence data as input, whereas a standard LSTM model can only use vector sequences. The combination of LSTM and CNN architectures results in a model that is sensitive to both spatial and temporal context, as the dependencies in the data in both two dimensions are retained and exploited during the training of the model.

Theory The theoretical underpinning of a ConvLSTM is extremely similar to that of a regular LSTM. The main difference with the standard LSTM architecture is that the set of learnable weights in the forget, input and output gates as well as in the cell state is derived through convolution instead of matrix multiplication. As a consequence, the input vector, forget vector, cell state and output vector are all changed from 2D tensors, vectors over time, to 3D tensors, matrices over time, whose first dimension is time and last two dimensions are the spatial dimensions. The ConvLSTM model determines the future cell state not only using its own input and past cell states, but also using those of its local surrounding cells. This allows the model to retain spatial structure in the data, while also learning from the temporal dependencies in the data using the regular LSTM steps of updating and using the cell state. The LSTM formulas introduced earlier are now as follows:

$$f^{(t)} = \sigma(b^f + W_x^f \star X^{(t)} + W_h^f \star H^{(t-1)})$$

$$g^{(t)} = \sigma(b^g + W_x^g \star X^{(t)} + W_h^g \star H^{(t-1)})$$

$$\tilde{C}^{(t)} = \tanh(b^c + W_x^c \star X^{(t)} + W_h^c \star H^{(t-1)})$$

$$C^{(t)} = f^{(t)} \bullet C^{(t-1)} + g^{(t)} \bullet \tilde{C}^{(t)}$$

$$o^{(t)} = \tanh(b^o + W_x^o \star X^{(t)} + W_h^o \star H^{(t-1)})$$

$$h^{(t)} = o^{(t)} \bullet \tanh(C^{(t)})$$

In these equations, the star denotes the convolution operator and the filled-in dot the Hadamard product. The Hadamard product of two matrices is the result of multiplying every element of the first matrix with the corresponding element (meaning in the same location) of the second matrix. Clearly, this requires the two elements to have the exact same shape, which is why padding is applied before the convolution operation in order to ensure that the cell state matrix and input matrix have the same shape. In all these equations, W is a set of learnable weights, which can be equated to a single kernel in the CNN architecture. The total number of filters, each containing the same number of kernels as there are data channels, is a hyper-parameter that can be adjusted, identically to a CNN.

Model architecture The ConvLSTM model architecture used to answer the main research question is similar to the architecture of the CNN. There are two combinations of ConvLSTM layer followed by a batch normalization layer with respectively 32 and 16 filters. The output is then flattened and fed into a fully-connected layer consisting 8 neurons with a 0.2 drop-out rate, which is followed by the output layer consisting of a single neuron. The learning rate is 0.001 and the model is trained for 50 epochs with a batch size of 64. The activation function in all layers are ReLu, except for in the output layer, where the Sigmoid activation function is used. The loss function is binary cross-entropy loss.

The ConvLSTM architecture used for the secondary research question is identical to the one described above, except for the output layer. The output layer consists of 5 neurons and uses the SoftMax activation function. Additionally, categorical cross-entropy loss is used for the loss function.

4.4 Implementation

In the following subsection, the practical implementation of the approach and models introduced earlier is described. Firstly, the metrics by which to judge and compare the relative performance of each model are presented and explained. Additionally, the cross-validation procedure is described, which is necessary to reduce variance in the final performance metrics. Secondly, the relevant hyperparameters and tuning procedure are described for each model. Thirdly, an overview is provided of the code written for this thesis and the libraries and software that were used.

4.4.1 Evaluation

Metrics The main research question is a binary classification problem and the secondary research question a multi-class classification problem. As such, all models trained for these two tasks generate a categorical output. For this reason, the metrics used to evaluate the models' performances in this thesis are as follows: accuracy, precision, recall, F1-score and Area Under the Receiver Operating Curve (AUROC). For the multi-class classification models, all these metrics are calculated by taking the macro-average over all classes. This means that every metric is calculated in a one-versus-all manner for each class and then averaged over the number of classes.

1. Accuracy is defined as the number of correct predictions divided by the total number of predictions. In a balanced data set, it gives a good indication of the performance of the model, as true positives and true negatives are given equal importance. In the following equations, TP and TN stand for respectively number of true positives and number of true negatives.

$$\text{Accuracy} = \frac{\text{TP + TN}}{\text{Total number of predictions}}$$

2. Precision represents the proportion of positive predictions that were correctly identified. It is calculated by dividing the number of true positives by the sum of true positives and false positives. In an unbalanced data set, precision gives a better indication of the performance of the model than accuracy. Additionally, in a situation where it is important to prevent false positives, such as automated fraud detection, precision can be an important metric to judge a model's performance. In the following equations, FP stands for the number of false positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

3. In contrast to precision, recall represents the proportion of actual positives that were correctly identified. It is calculated by dividing the number of true positives by the sum of true positives and false negatives. Similar to precision, recall is a good metric to use in case of an unbalanced data set. In situations where preventing false negatives is important, such as automated detection of tumors or other illnesses, recall also plays an important role.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

4. Fairly judging a model's performance means that precision and recall should both be taken into account. However, improving precision often reduces recall and vice versa. Comparing two models that each score highest in one of precision and recall is therefor difficult, as it might be unclear what to prioritise when selecting the best model. For this reason, F1-score is a useful metric, as it is the harmonic mean between precision and recall, which allows for a clear comparison between models. A high F1-score indicates that a model performs well in identifying a large proportion of true positives while not casting the net too wide and identifying many false positives.

$$\text{F1-score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. The last metric used in this thesis is AUROC. AUROC provides information on the ability of the model to discriminate classes. The AUROC value is derived by taking the area under the receiver operating characteristic, which is the ratio between true positives rate and false

positives rate for a range of decision thresholds. An AUROC score of 0.8 means that 80% of the time, the model will assign a higher probability of being positive to a randomly selected true positive than to a randomly selected true negative. As such, 0.5 is the worst AUROC score and 1 the best, since an AUROC score below 0.5 means the model can be inverted.

Cross-validation Cross-validation is a method used to reduce the variance in the final scores that each model achieves. It works by training and testing each model multiple times on slightly different sub-sets of the data set. In this thesis, two different cross-validation methods have been used due to the nature of the modelling tasks. The second research question is framed as a regular prediction task, which means k-fold cross-validation can be used to derive results, as illustrated in Figure 14.

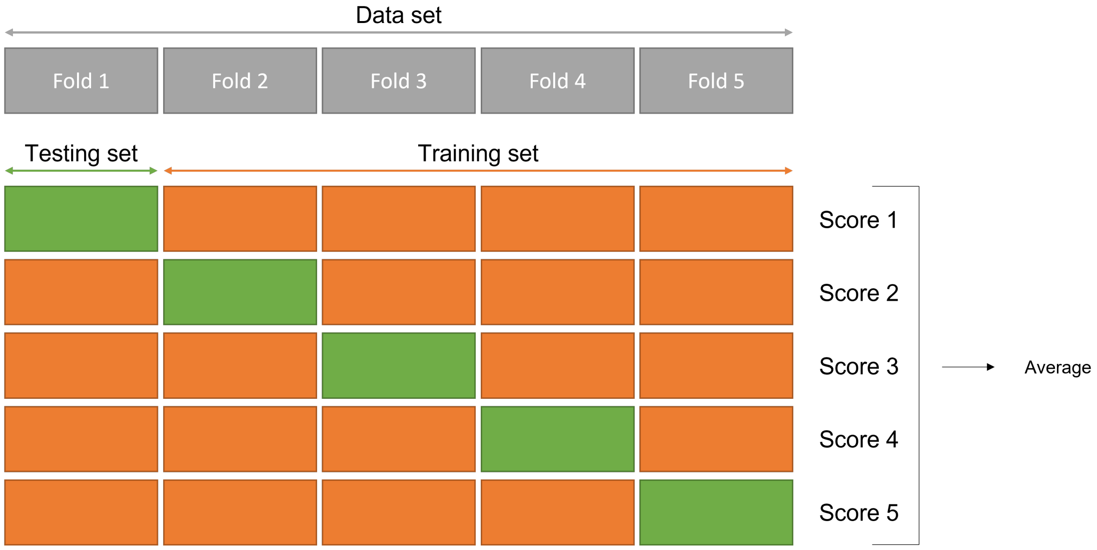


Figure 14. The split between training and testing for a regular 5-fold cross-validation procedure. The green boxes denote the folds used as the testing set and the orange boxes denote the folds used as the training set.

However, the first research question is framed as a forecasting problem, which means regular cross-validation can not be employed. As described earlier, a forecasting problem requires the training data to chronologically strictly precede the testing data. As can be seen in Figure 14, if the data set is divided into chronologically ordered folds this is not the case for regular k-fold cross-validation. Instead, time-series split cross-validation can be used, as illustrated in Figure 15. This method is similar to regular k-fold cross-validation, except for the fact that the part of the training set that does not precede the testing set is not used. As the testing set moves ahead chronologically after each cycle, the training set essentially grows. Consequently, the training and testing set in the last cycle are identical in both cross-validation method.

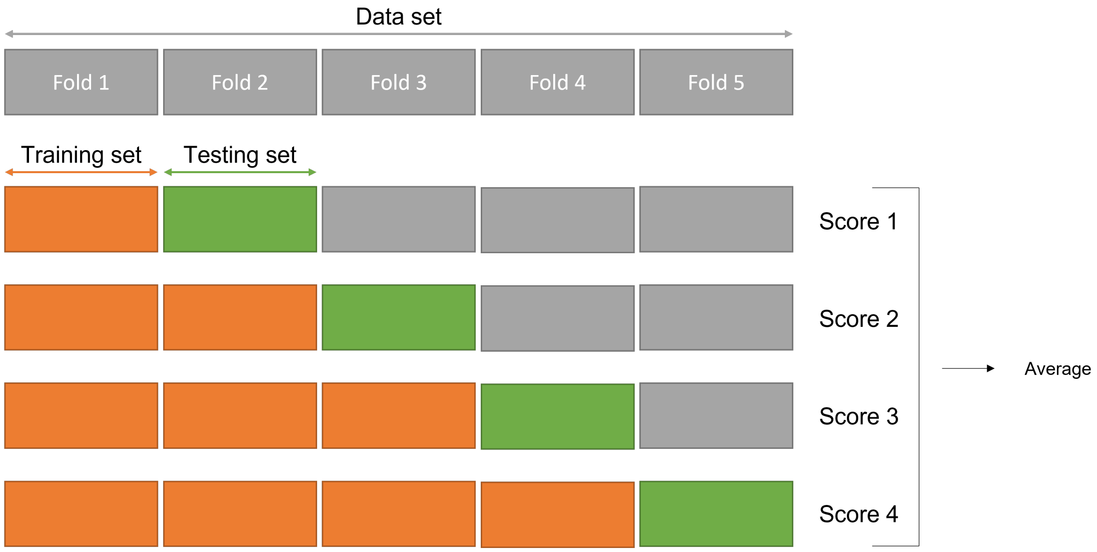


Figure 15. The split between training and testing for a time-series split cross-validation procedure. The green boxes denote the folds used as the testing set and the orange boxes denote the folds used as the training set. The grey boxes denote the folds not used per cycle as they do not chronologically precede the testing set.

4.4.2 Hyperparameter tuning

Due to the long training times of some model configurations in combination with computing power constraints, a grid search across many hyperparameters and respective options to find the optimal architecture for each model was not feasible. For this reason, manual experimentation was used to limit the number of times each model had to be retrained. The most important consideration was preventing overfitting while allowing the models to discover potentially highly complex, non-linear relations in the data. In order to guarantee a high level of generalizability, the validation loss was compared to the training loss after every training cycle. In the end, the manual experimentation resulted in the model architectures as described above, which balance practical training times, generalizibility and complexity of features.

4.4.3 Code

The programming language used to gather and process the data and create the predictive models is Python. For the data gathering and processing phase of the thesis, primary use was made of the Pandas, Geopandas, Numpy and Matplotlib libraries. The baseline Random Forest model was implemented using scikit-learn (Pedregosa et al., 2011), while the Deep Learning models were implemented using the Deep Learning framework Keras (Chollet et al., 2015).

5 Results

In the following section, the results of this study are presented. Firstly, the performance of the binary classification models are presented and interpreted. Sub

Secondly, the performance of the multi-label classification models are discussed.

5.1 Wildfire occurrence prediction

Model evaluation In order to predict next-day wildfire probability, four classification algorithms were trained and tested. These four models are: Random Forest, LSTM, CNN and ConvLSTM, as introduced in Section 4. In this study, the problem was approached as a forecasting problem, which means that the training data chronologically strictly preceded the testing data. Normally, k-fold cross-validation is used to reduce the variance in the model performance metrics, however regular k-fold cross-validation is not possible within the paradigm of a forecasting problem. For this reason, time-series split cross-validation was used, which was introduced in Section 4. The testing years were 2013, 2016, 2019 and 2021, which results in an approximately 90% train-test split. The metrics used to evaluate the models' performance are: accuracy, precision, recall, F1-score and AUROC, which were introduced and explained in Section 4. The results presented in table 2 shows the performance of each model when trained on all years prior to 2021 and tested on 2021, denoted by 'Split on 2021'. Additionally, table 2 shows the average values for each metric over the time-series split cross-validation procedure, denoted by 'Average over all splits'.

Classifier	*Accuracy*	*Precision*	*Recall*	*F1-score*	*AUROC*
Split on 2021					
Random Forest	0.582	0.677	0.300	0.416	0.709
Long Short-Term Memory Model	**0.735**	**0.716**	**0.770**	**0.743**	**0.805**
Convolutional Neural Network	0.680	0.726	0.660	0.690	0.745
ConvLSTM	0.666	0.670	0.661	0.670	0.754
Average over all splits					
Random Forest	0.613	**0.737**	0.325	0.443	**0.728**
Long Short-Term Memory Model	**0.661**	0.669	0.601	**0.632**	0.724
Convolutional Neural Network	0.647	0.597	**0.653**	0.616	0.718
ConvLSTM	0.623	0.557	0.629	0.586	0.700

Table 2: The models performances' in multiple metrics. The top shows the performance of the models trained on the years 2010 to 2020 and tested on 2021, while the bottom shows the average over all four splits. Per split and metric, the highest value is marked in bold-face.

Table 2 shows that for the 2021 split the LSTM model performed best in every metric. It scores a 71.3% score for total accuracy, which is significantly higher than the Random Forest model and slightly higher than the CNN and ConvLSTM models. Additionally, it scores a 76.2% score for AUROC, which indicates a good performance according to Hosmer & Lemeshow (2000), and 73.5% on precision, which indicates that the model has a low false positive rate. While achieving decent to good scores on accuracy, precision and AUROC, the LSTM model performed worse on recall with 66.3%. This score indicates that the model struggles to achieve a low false negative rate and mis-classifies a significant amount of fire samples. Scoring between 0.3% and 3.3% worse, the performance of the convolutional neural network model was slightly worse than the LSTM model in every metric. The CNN correctly classified 68% of all samples and achieved an AUROC of 74.5%, which, similarly to the LSTM model, indicates a good ability to distinguish fire from no-fire samples. For the 2021 split, the performance of the ConvLSTM model exceeded the baseline Random Forest model in all metrics, except for precision. It slightly exceeded the performance of the CNN model in recall and AUROC, with scores of respectively 67.0% and 75.4%, but was still outperformed by the LSTM model in every metric.

The models performances' averaged over all splits provide a less uniform indication of which model achieved the best performance. The best accuracy was again achieved by the LSTM model with a score of 66.1%, which is slightly higher than the CNN and ConvLSTM models with respectively 64.7% and 63.2% and significantly higher than the Random Forest model with 61.3%. While the highest precision was achieved by the Random Forest model with 73.7%, it achieved an extremely poor recall and subsequently F1-score with 32.5% and 44.3% respectively. This combination of high precision but low recall could indicate that the Random Forest model mis-classified a high number of fire samples and only correctly classified fire samples with extreme values, making them easy to classify. The highest F1-score was achieved by the LSTM model with 63.2%. In terms of AUROC, the Random Forest model achieved the highest score with 72.8%, closely followed by the LSTM model with 72.4%. The average performance of the ConvLSTM was generally poor, finishing either last or second-last in all metrics.

Based on all these results, a few observations can be made. Firstly, for each neural network model, the performance when trained on the 2021 split was higher than the average performance over all four splits. However, this is not the case for the Random Forest model, which could indicate that the neural network models have a higher capacity for generalisation. Secondly, while the general performance of all four models seems relatively low, due to the highly stochastic nature of wildfires, an accuracy of 71.3% and 66.1% can be seen as a decent to good performance.

Figure 16 shows the ROC curve for all four models. An ROC-curve shows the ability of a model to distinguish the two classes by showing the True Positive Rate and False Positive Rate at different decision thresholds. A decision threshold is defined as the minimum probability that must be reached for a data point to be classified to class 1. Traditionally, this threshold is set to 0.5, however other options are possible. A threshold of 0 means the True Positive Rate and False Positive Rate are 0, while a threshold of 1 means the True Positive Rate and False Positive Rate are 1. A random classifier has no ability to distinguish between classes, meaning that its ROC-curve follows the line $x = y$. On the contrary, the ROC-curve of a perfect classifier follows the equation $y = 1$, as its prediction contains no false positives and all true positives are predicted correctly, regardless of the chosen threshold. With this information, Figure 16 clearly shows that the best ROC curve belongs to the LSTM model followed by respectively the CNN model, the ConvLSTM model and finally the RF model. This corroborates the general findings shown in Table 2 as the LSTM model performs quite well and all models outperform the RF baseline model.

Figure 16 clearly shows that for all three Deep Learning models, a relatively high true positive rate can be achieved while keeping the false positive rate. This can be seen from the fact that the curve climbs steeply until an inflection point, after which the slope of the curve decreases quite significantly. This reflects that different decision thresholds will result in a different true positive-false positive ratio. While the standard is naturally 0.5, in the case of natural hazard mitigation it may be more important to minimize false negatives than it is to minimize false positives, which could lead to a different decision threshold being optimal. The optimal value is irrevocably tied to real-world policy and its practicalities, however the the high distinguishing ability of the LSTM model, as shown in Figure 16, allows for different decision thresholds to still produce good results.

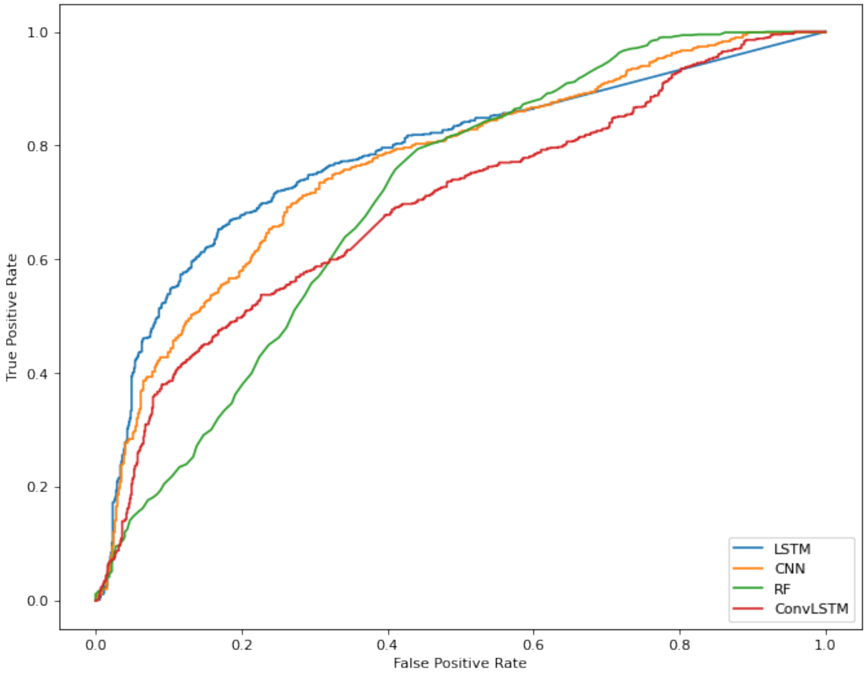
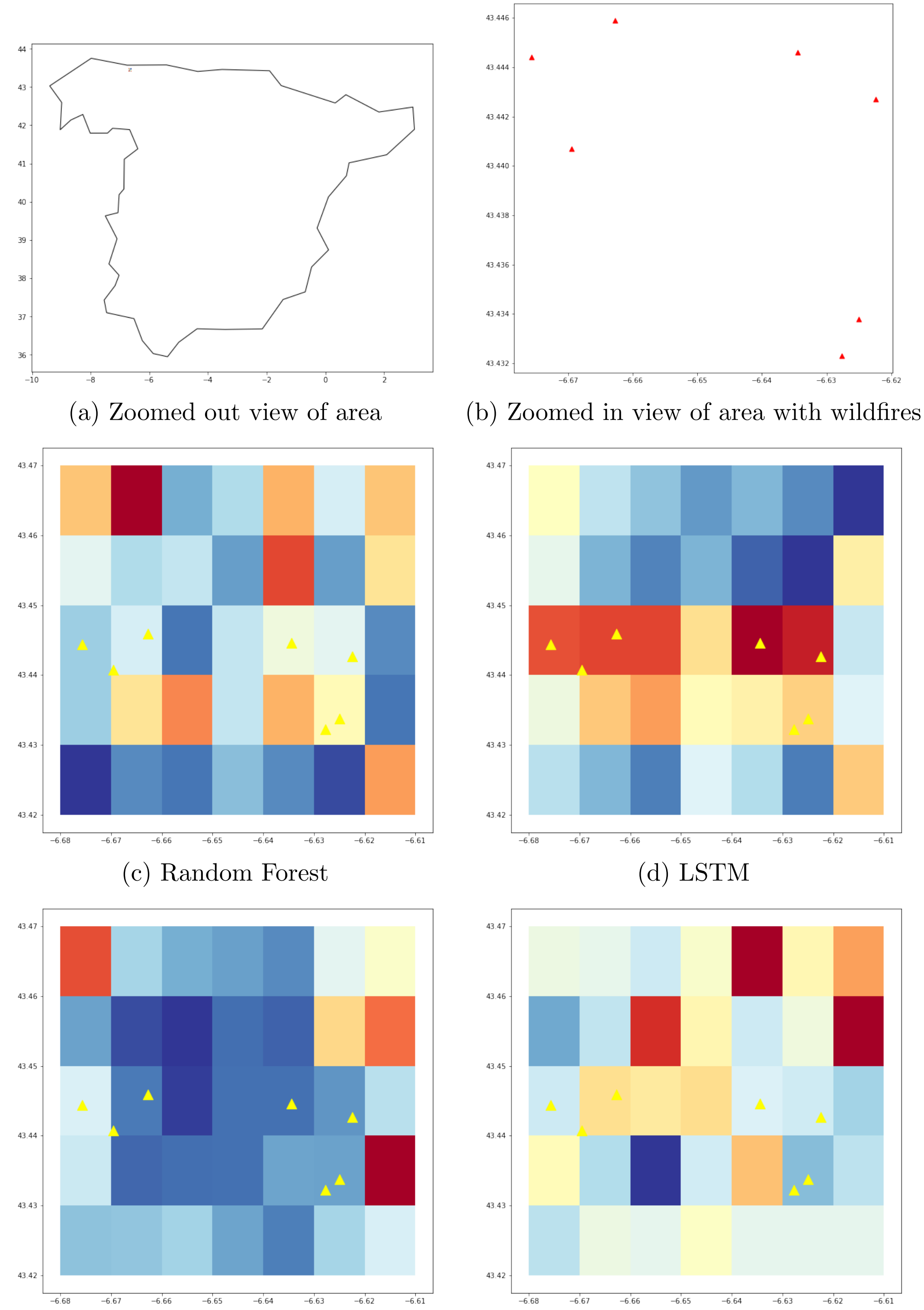


Figure 16. The Receiver-operating-characteristic curve for all four predictive models. The x-axis shows false positive rate and the y-axis shows true positive rate for different classification thresholds. A line that lies closely to the left and top edge indicates good performance.

Performance in practical application In order to illustrate the practical applicability of the models, a small area of Spain on the date 2017-04-24 was chosen and all necessary data in this area gathered. This date and region were chosen due to the occurrence of a high number of wildfires in close proximity, which is useful for illustrating the models' ability to locate high-risk areas.

The trained models were used to predict the fire danger for all cells contained in this area. Figure 17 shows, respectively, the location of this area of interest within Spain, the location of wildfires within this area and the predicted fire danger maps by all four models. In the top right plot, wildfires are indicated by the red pyramids. In the bottom four plots, wildfires are indicated by the yellow pyramids and low to high fire danger is indicated by the blue to red color gradient. As can be seen, in this particular example the CNN does not accurately locate the risk areas. The Random Forest and ConvLSTM perform slightly better, predicting medium fire danger in the approximate area of the wildfires. The LSTM model, however, almost perfectly predicts the areas that were at high risk of wildfires on that particular date. As can be seen in the LSTM plot, five of the seven wildfires are located in areas that have a high predicted wildfire risk relative to their surroundings, while the remaining two have a medium predicted wildfire risk relative to their surroundings.



(a) Zoomed out view of area

(b) Zoomed in view of area with wildfires

(c) Random Forest

(d) LSTM

(a) CNN

(b) ConvLSTM

Figure 17. An illustration of how such spatial susceptibility maps could be used in practice. The top two plots show the location of the area of interest and the distribution of wildfires on the date of interest denoted by red pyramids. The bottom four plots show the fire danger per tile as predicted by all four models. The fire danger ranging from low to high danger is denoted by the blue to red gradient and the yellow pyramids denote wildfires.

Figure 18 shows a more zoomed-out view of the fire danger as predicted by the LSTM model. In addition to the cells that contain the wildfires, the model predicts high fire danger in the top left corner as well as in the bottom left and right corner. This visualisation shows both the accurate performance of the LSTM model, which is in agreement with the other insights in the results, as well as the manner in which such models could be used in practice. The zoomed-out view of the other three models can be found in appendix B.

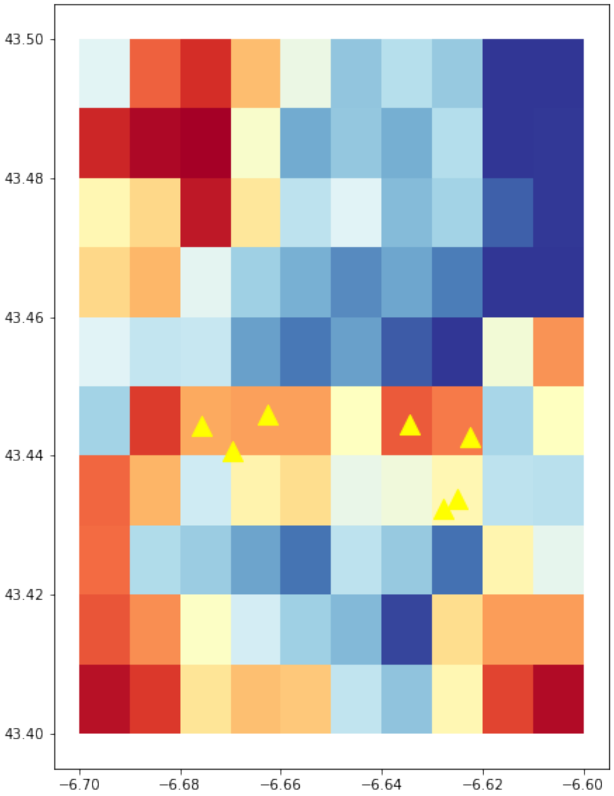


Figure 18. Zoomed-out view of wildfire risk predicted by the LSTM model.

Currently, such fire danger maps are created using the Fire Weather Index. However, the good performance as shown in Figures 17 and 18 show the two main benefits of the algorithms developed in this thesis, taking the LSTM model as the main example as it achieved the best performance. Firstly, the LSTM model provides predictions on a more detailed scale of approximately 1 km by 1 km, whereas the FWI is calculated in cells of 9 km by 9 km. Secondly, FWI is purely weather-based, which means it does not take into account the other conditions that increase an areas susceptibility to wildfires, such as the moisture level of the vegetation, type of land or level of human activity. This allows for both a more accurate and more fine-scaled prediction of fire danger within an area. Ideally, this visualisation would be performed over a larger area and for more dates to further substantiate the qualitative comparison, however due to data storage and computing constraints this was infeasible.

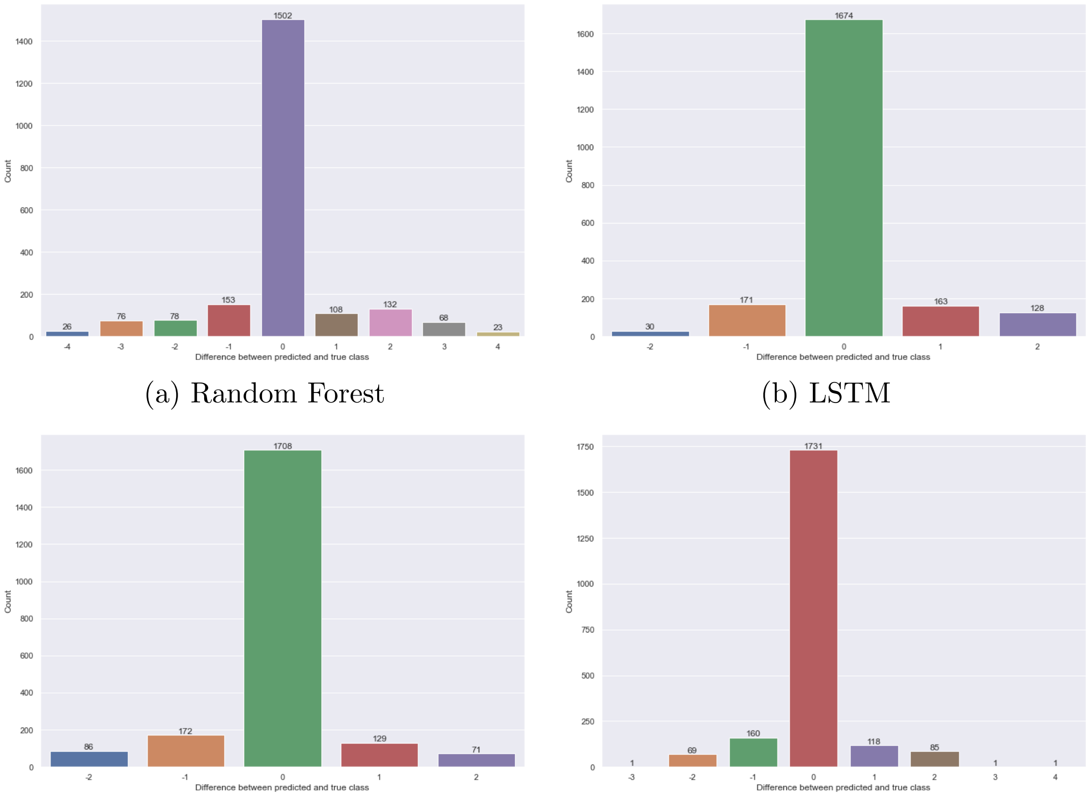
5.2 Wildfire severity prediction

The secondary research goal in this thesis is attempting to predict the severity of wildfires by a 5-class classification procedure. The data set with which to answer this question contains only wildfires that burned a non-zero number of hectares, which means the prediction is being conditioned on the fact that the wildfire event occurs. For this reason, this problem is not approached as a forecasting problem, but instead a regular prediction problem. As a consequence, the time-series split cross-validation procedure that was used to answer Question 1 is no longer necessary and a regular cross-validation procedure can be employed. The baseline model and three Deep Learning model are trained in a 5-fold cross-validation procedure, using 80% of the data for training and 20% for testing. The metrics used to evaluate the models are identical, however, due to the increased number of output classes, the relative importance of each metric and benchmark scores are different. Table 3 shows the performance in each metric for the four models averaged over all five folds.

Classifier	*Accuracy*	*Precision*	*Recall*	*F1-score*
Random Forest	0.717	0.708	0.706	0.705
Long Short-Term Memory Model	0.793	0.770	0.777	0.771
Convolutional Neural Network	0.803	0.782	0.787	0.782
ConvLSTM	**0.812**	**0.794**	**0.796**	**0.793**

Table 3

Table 3 shows that the performance of all three Deep Learning models is remarkably strong. While the Random Forest baseline performs quite well with an accuracy of 71.7%, it is outperformed by all three context-sensitive models and achieves the lowest scores in all metrics. Across all four metrics, the ConvLSTM scores the best, followed by the CNN, which in turn is followed by the LSTM. The ConvLSTM model reaches an exceptionally high score in all four metrics. It places 81.2% of samples in the right output class on average, whereas a random model would score 20% accuracy due to there being five output classes. It additionally reaches similarly high scores for precision, recall and F1-score, making it the best model out of the four for wildfire severity prediction. None of the three Deep Learning models achieved a specifically poor performance, which suggests that a strong benefit can be derived from using context. Specifically, the use of spatio-temporal context seems to be the most beneficial.



(a) Random Forest

(b) LSTM

(c) CNN

(d) ConvLSTM

Figure 19. For each of the four models, predicted values are generated for a test set of 20% of the total data set. The difference between the predicted and true output class is then calculated for each sample. How often each discrepancy occurred is displayed for each of the four models.

A more in-depth look at the results reveals that all three Deep Learning models not only achieve a high accuracy, precision, recall and F1-score, but also that the difference between the predicted and the true class is often very small. Figure 19 shows how often a discrepancy between the true and predicted class occurs for each of the four models. As can be seen in the figure, no test sample is predicted to be more than 2 classes higher or lower than its true value by the LSTM and CNN models. The ConvLSTM makes three predictions that are more than 2 classes away from the true value, but scores a higher accuracy overall. On the other hand, while the Random Forest achieves a decent performance, it generates large discrepancies between the predicted and true value far more often than the context-sensitive models, with, for example, 49 samples being predicted the maximum of four classes apart from their true value. This shows that the use of context-sensitive Deep Learning models not only improves overall accuracy, but also results in smaller mistakes. Lastly, it is important to note that the fact that the ConvLSTM is the only context-sensitive model that makes a prediction mistake of 2 classes or higher is counter-evidence to the claim that it is the undisputed best model for this prediction task. However, the extremely rare occurrence of prediction discrepancies of this size (3 out of 2166) coupled with its dominant performance across all evaluation metrics, still points to the result that the ConvLSTM model and with that spatio-temporal context is most suitable for wildfire severity prediction.

6 Conclusion

In this thesis, an answer was sought to two questions:

1. How can context-sensitive Deep Learning models be used to predict wildfire occurrence in Spain?
2. To what extent are context-sensitive Deep Learning models suitable for wildfire severity prediction?

In order to answer these two questions, two different classification tasks were defined. The first was a binary classification of next-day wildfire occurrence, defined as a forecasting problem. The second task was a multi-class classification of wildfire severity, not defined as a forecasting problem. Three Deep Learning classification models, each sensitive to either spatial, temporal or spatio-temporal context, were trained and compared to a baseline Random Forest classification model. Subsequently, the same three Deep Learning architectures and Random Forest baseline were extended and trained to predict the severity of a wildfire, measured in total hectares burned, given the same predictive variables as for the first question.

With regards to Question 1, the results showed a significantly better performance by the context-sensitive Deep Learning models than by the context-insensitive baseline. While the average scores over all splits did not indicate a clear winner, the LSTM model achieved the best average accuracy with 66.1%. Since the data set that was used was a balanced data set, accuracy can be regarded as the most important metric of comparison. Additionally, the results of the 2021 split, which uses the most data during training, provided further evidence that LSTM is best suitable for this task. It achieved the highest score in every metric, which included a 73.5% accuracy score and 80.5% AUROC score. The CNN and ConvLSTM model also outperformed the baseline albeit with significantly lower scores than the LSTM. This suggests that context-sensitive Deep Learning models can be highly useful for wildfire occurrence prediction in Spain, especially models sensitive to temporal context.

The answer to the second question is less clear due to the strong performance achieved by all three Deep Learning models. The baseline Random Forest model was outperformed by all three Deep Learning models, however it still achieved a decent score. The ConvLSTM model achieved the highest accuracy of 81.3%, which can be seen as strong evidence that Deep Learning models sensitive to spatio-temporal context are suitable to wildfire severity prediction. All in all, while the novel approach tested in this thesis has shown spatio-temporal context to be most suitable to burn area prediction, the spatial model outperformed the temporal model and performed relatively better than in the occurrence prediction task. This result intuitively stands to reason as the spread of a wildfire seems to be influenced more than the ignition by the characteristics of the entire tile and its neighbours.

7 Discussion

Reflection on method and results

Wildfire occurrence prediction The relative performance in the wildfire occurrence prediction task achieved by the baseline models and three Deep Learning models are similar to those achieved in the literature. Prapas et al. (2021) found that their LSTM, CNN and ConvLSTM architectures performed better than their baseline model in every metric but precision. Additionally, their LSTM model scored the highest in precision and recall, but they saw their ConvLSTM model score the highest AUROC score. Similar relative performances were seen in this paper, with the LSTM model scoring the highest across all metrics, the ConvLSTM model achieving the second-highest AUROC score and the baseline model achieving the lowest score in every metric but precision. The absolute performances, however, differ as the models in this thesis achieve generally similar or higher scores for recall and F1-score and lower scores for precision and AUROC.

A few points of critique can be made on the method used in this thesis that might have decreased the final models' performances. Firstly, compared to for example Prapas et al. (2021), fewer samples are used. In their study, a 1 to 2 ratio is used for sampling no-fire samples, which increases the total number of training samples and with that the potential for the models to find patterns in the data. This by itself makes a straight-forward comparison between the model performances difficult. Secondly, the choice of predictive variables was based on the literature. Theoretically, neural networks do not require a large amount of feature selection and engineering, since the optimization method should be capable of suppressing the variables with low predictive power. Whether this is practically feasible depends on the noise in the data and an efficient convergence of the optimization procedure, which is not guaranteed. Due to the black-box nature of neural networks, it is not an easy task to find the variables it uses the most for its predictions, which means the variables with low predictive power cannot simply be removed by hand. As such, a feature selection procedure, whether this be simple or complex, could potentially increase the performance of the Deep Learning models. Thirdly, there are a few data quality concerns that might reduce the ability of the Deep Learning models to generalize to unseen data sets. The first data quality issue has to do with the target variable. The original data set the location and date of wildfires was extracted from was first filtered to only contain fires that were detected with 80% confidence or above. Thus, actual fire-samples that had a lower confidence score, might have been sampled as no-fire samples, which introduces noise in the data set and increases the difficulty for the models to observe non-trivial mappings. On top of this, the target data set contains no information on the cause of the fire, which means some fires might have ignited spontaneously while others may have been ignited by a fire in a neighbouring cell. Again, this introduces noise in the data, since the conditions needed for spontaneous ignition may be less extreme than those needed for ignition due to an already existing wildfire. However, this does not threaten the general validity of the research, but merely changes the interpretation of the model output from the probability of a spontaneous or human-caused ignition within the cell to the general probability of a wildfire occurring within that cell.

Wildfire severity prediction There are no prior studies similar enough to directly compare results, however the high scores achieved seem promising, especially in a 5-class classification problem. The data quality issues that are relevant to the main question are not shared by the data set used for the second question, since no negative samples were needed. A different data quality issue, however, has to do with the fact that the total hectares burned by a wildfire is highly dependent on humans interfering during the burning phase. Attempts to extinguish the fire, albeit often once the fire has reached a high severity level already, could strongly influence the total hectares burned and the human extinguishing effort is not taken account in the model. However, this is likely not the source of major noise in the data, because the size a fire has to reach in order to warrant active fire-suppressing efforts is likely higher than the threshold for the fifth class, which is approximately 11 hectares. As such, the fire would already have reached the threshold and the accompanying size before it is suppressed.

Interpretation

For both research questions, an important question regarding the interpretation of the results is whether the Deep Learning models performed better due to the added value of the context or simply due to an increase in data points per sample that may or may not be beneficial. Seeing as how the baseline model performed better than the CNN and ConvLSTM in various metrics in the first research questions, it does not seem to be the case that providing the model with more data will automatically increase the performance of the model. In fact, it seems to introduce enough noise in the data for the performance of these two models to fall below that of the baseline. In turn, this implies that the temporal-context available to the LSTM model does in fact bring added value. Regarding the second research question, there seems to be strong evidence for the spatio-temporal context to bring added value as it outperforms all other models.

Future research

There is a clear policy application for these DL models as they could potentially be used to assist forestry management groups in preparing for and extinguishing wildfires. However, similar to many other DL models, there is no indication on the confidence that can be put in their prediction, even if their accuracy and other performance scores are acceptable. In order to practically implement ML models, especially when it concerns mitigating the risks and damages of natural hazards, it is important to know how much trust can be put in the model and whether it contains any unforeseen biases. Herein lies the main potential for future work. Firstly, establishing to what extent such models can be confided in through methods such as Bayesian Neural Networks, which can tell us how uncertain our predictions are by learning a probability distribution over possible neural networks. Secondly, using explainable AI methods, such as Shap-values, partial dependency plots or other methods to examine how the final predictions are derived and whether the models contain any biases.

A Extended data exploration

Table 4 shows five random samples from the context-less data set that is used by the baseline Random Forest model for the first research question. Table 5 shows the mean, standard deviation, minimum, median and maximum value for every continuous variable in the context-less data set used to answer the first research question. Figure 6 show the distributions for the remaining continuous variables split by positive or negative population. Table 7 shows summary statistics for all variables in the data set used by the context-less model for the second research question.

y	u10	v10	t2m	str_	e	tp	t2m_max
1.0	-0.930945	0.926695	293.762634	-4580202.0	-0.000670	5.699694e-07	298.346558
0.0	0.899566	-2.140720	292.530365	-5355336.0	-0.001115	1.925975e-06	297.422394
1.0	0.968847	-0.108770	289.750946	-4941545.0	-0.000392	1.711771e-06	294.600433
0.0	-2.528738	2.638594	298.951019	-5465721.0	-0.000546	5.699694e-07	304.561401
1.0	-0.637765	2.548837	294.836029	-3884036.5	-0.000745	7.679686e-06	301.829041

dc	bui	ffmc	ic	population	u_rate	landcover	ndvi
711.434387	107.567711	90.447899	2.260417	0.000000	21.39	1.0	31560000.0
576.968750	192.164062	91.278809	17.468750	2.466026	15.53	4.0	63840000.0
764.554199	153.614578	92.326149	11.166667	4.576644	21.39	4.0	51810000.0
628.161438	177.638016	92.940346	15.843750	3.559698	26.09	4.0	31410000.0
715.901978	100.533981	92.219536	20.006771	192.547276	17.22	4.0	65760000.0

evi	elevation	road_density	wildland_urban
10190000.0	122.777778	2	True
35210000.0	137.500000	8	True
27515000.0	154.000000	3	True
20720000.0	58.000000	4	False
26150000.0	92.666667	8	True

Table 4: Five randomly selected samples from the context-less data set used for the first research question.

	y	u10	v10	t2m	str_	e	tp
Mean	0.497448	0.059799	-0.047871	291.824527	-4.795791e+06	-0.001015	0.000301
Std.	0.500004	1.660536	1.467330	6.857073	1.198333e+06	0.000635	0.001170
Min.	0.000000	-12.256624	-7.092863	265.722198	-7.917503e+06	-0.003980	0.000000
Median	0.000000	-0.018703	0.046847	292.246277	-4.850969e+06	-0.000904	0.000002
Max.	1.000000	8.205373	8.794882	308.710693	-1.754898e+05	0.000335	0.029777

	t2m_max	dc	bui	ffmc	ic	population	u_rate
Mean	296.932518	520.839900	128.854861	88.176431	12.170408	179.140550	20.261050
Std.	7.407585	330.870421	107.823551	10.528316	9.874488	1331.860030	3.917255
Min.	268.530945	0.000000	0.000000	4.327148	0.000000	0.000000	14.100000
Median	297.415161	537.830353	111.454422	90.919312	10.400000	8.996182	21.390000
Max.	315.772766	1816.226562	791.872375	99.353416	52.125000	49120.628906	26.090000

	ndvi	evi	elevation	road_density
Mean	5.177322e+07	2.868978e+07	124.942122	10.643813
Std.	1.683811e+07	1.035523e+07	42.361164	31.984750
Min.	-2.000000e+07	-3.310000e+06	0.000000	0.000000
Median	5.304500e+07	2.720000e+07	135.000000	4.000000
Max.	9.123000e+07	7.424000e+07	187.000000	977.000000

Table 5: The mean, standard deviation, minimum, median and maximum of the continuous variables used by the context-less Random Forest model for the first research question.

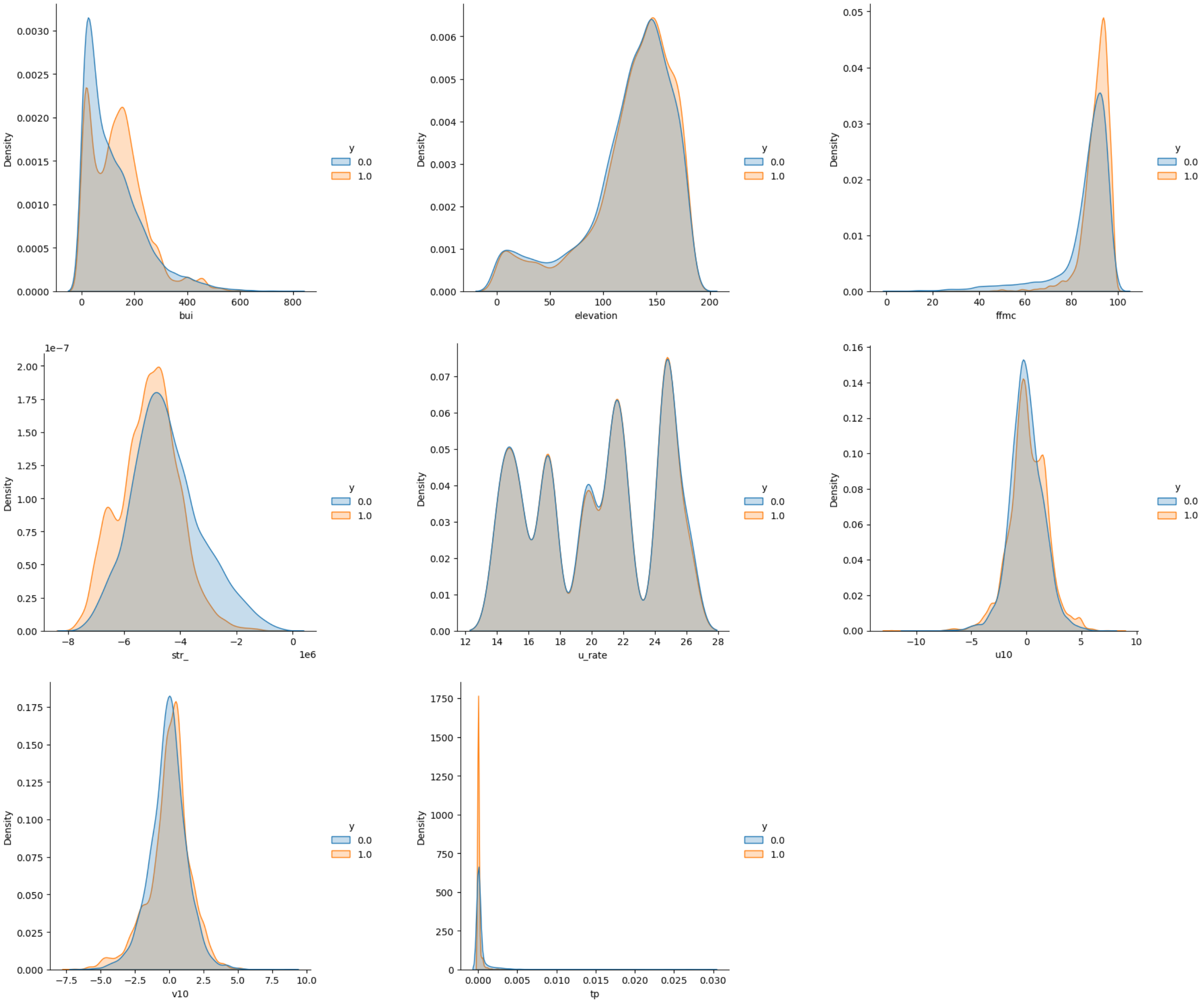


Table 6: The distribution for fire and no-fire data points for respectively Build-Up Index, elevation, Fine Fuel Moisture Code, surface temperature, unemployment rate, Eastward wind speed, Northward wind speed and precipitation in the baseline data set for the first research question.

	u10	v10	t2m	str_	e	tp
Mean	-0.096638	0.125971	287.306093	-4.378123e+06	-0.001086	0.000309
Std.	1.471177	1.707697	7.020161	1.038295e+06	0.000594	0.000893
Min.	-8.657542	-8.445972	261.373779	-7.704670e+06	-0.004686	0.000000
Median	-0.140905	0.114127	287.138794	-4.440670e+06	-0.001006	0.000005
Max.	8.025438	7.981335	307.838135	-4.842312e+05	-0.000030	0.020451

	t2m_max	dc	bui	ffmc	ic	population
Mean	292.318412	333.515333	75.558269	84.952313	8.015384	123.997523
Std.	7.605487	328.147187	89.126367	10.705277	7.543763	773.432471
Min.	262.499237	0.000000	0.000000	10.491862	0.000000	0.000000
Median	291.983887	278.964569	36.312500	87.401306	6.250000	17.817086
Max.	313.680695	1664.187500	657.648438	98.893753	48.281250	44060.373047

	u_rate	ndvi	evi	elevation	road_density
Mean	21.747642	5.440941e+07	3.086873e+07	122.130756	15.125473
Std.	1.760177	1.567974e+07	1.035560e+07	34.407475	31.730616
Min.	19.860000	-2.000000e+07	-1.900000e+05	0.000000	0.000000
Median	21.390000	5.747000e+07	3.030000e+07	128.000000	7.000000
Max.	24.790000	8.949000e+07	6.935000e+07	183.000000	1346.000000

Table 7: The mean, standard deviation, minimum, median and maximum of the continuous variables used by the context-less Random Forest model for the second research question.

B Wildfire risk visualisations

The following three plots show the wildfire risk per tile predicted by, respectively, the RF, CNN and ConvLSTM models, as also shown in the results for the LSTM model.

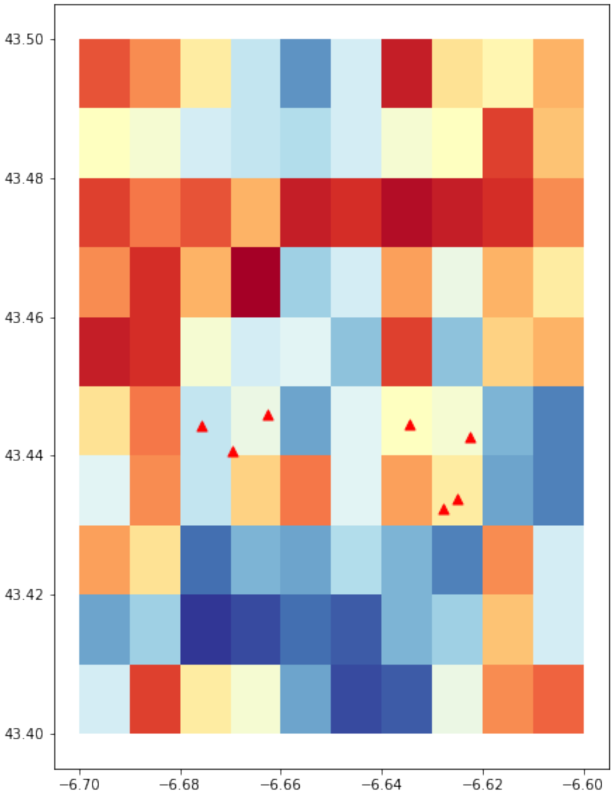


Figure 20. Zoomed-out view of wildfire risk predicted by the Random Forest model

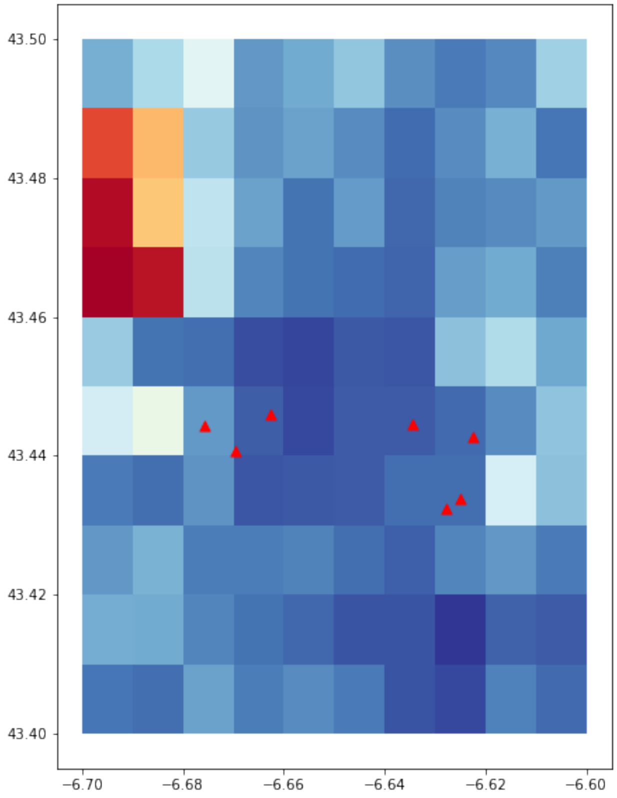


Figure 21. Zoomed-out view of wildfire risk predicted by the CNN model

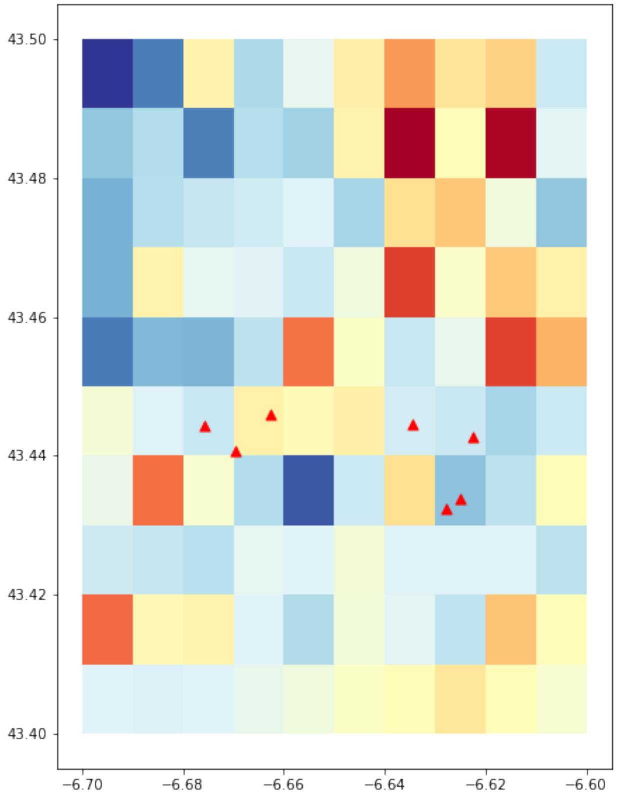


Figure 22. Zoomed-out view of wildfire risk predicted by the ConvLSTM model

References

Bjånes, A., De La Fuente, R., & Mena, P. (2021). A deep learning ensemble model for wildfire susceptibility mapping. *Ecological Informatics*, *65*, 101397. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1574954121001886` doi: https://doi.org/10.1016/j.ecoinf.2021.101397

Carr, R. (2023, Jan). *Spain.* Encyclopædia Britannica, inc. Retrieved from `https://www.britannica.com/place/Spain/Castilian`

Castelli, M., Vanneschi, L., & Popovič, A. (2015, 04). Predicting burned areas of forest fires: An artificial intelligence approach. *Fire Ecology*, *11*, 106-118. doi: 10.4996/fireecology.1101106

Center, C. I. F. F. (2003). Glossary of forest fire management terms. *Winnipeg, Manitoba, Canada*, *61*.

Cheng, T., & Wang, J. (2008, 10). Integrated spatio-temporal data mining for forest fire prediction. *T. GIS*, *12*, 591-611. doi: 10.1111/j.1467-9671.2008.01117.x

Chollet, F., et al. (2015). *Keras.* GitHub. Retrieved from `https://github.com/fchollet/keras`

Doerr, S., Santín, C., Maynard, T., Smith, N., & Gonzalez, S. (2013, 01). Wildfire: A burning issue for insurers?
doi: 10.13140/2.1.2551.9681

Flannigan, M., Amiro, B., Logan, K., Stocks, B., & Wotton, M. (2006, 07). Forest fires and climate change in the 21st century. *Mitigation and Adaptation Strategies for Global Change*, *11*, 847-859. doi: 10.1007/s11027-005-9020-7

Ghorbanzadeh, O., Valizadeh Kamran, K., Blaschke, T., Aryal, J., Naboureh, A., Einali, J., & Bian, J. (2019). Spatial prediction of wildfire susceptibility using field survey gps data and machine learning approaches. *Fire*, *2*(3). Retrieved from `https://www.mdpi.com/2571-6255/2/3/43` doi: 10.3390/fire2030043

Giglio, L., Schroeder, W., & Justice, C. O. (2016). The collection 6 modis active fire detection algorithm and fire products. *Remote Sensing of Environment*, *178*, 31-41. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0034425716300827` doi: https://doi.org/10.1016/j.rse.2016.02.054

Guo, F., Zhang, L., Jin, S., Tigabu, M., Su, Z., & Wang, W. (2016). Modeling anthropogenic fire occurrence in the boreal forest of china using logistic regression and random forests. *Forests*, *7*(11). Retrieved from `https://www.mdpi.com/1999-4907/7/11/250` doi: 10.3390/f7110250

He, Q., Jiang, Z., Wang, M., & Liu, K. (2021). Landslide and wildfire susceptibility assessment in southeast asia using ensemble machine learning methods. *Remote Sensing*, *13*(8). Retrieved from `https://www.mdpi.com/2072-4292/13/8/1572` doi: 10.3390/rs13081572

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression.* John Wiley and Sons.

Huot, F., Hu, R. L., Ihme, M., Wang, Q., Burge, J., Lu, T., ... Anderson, J. R. (2020). Deep learning models for predicting wildfires from historical remote-sensing data. *ArXiv*, *abs/2010.07445*.

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

Juang, C. S., Williams, A. P., Abatzoglou, J. T., Balch, J. K., Hurteau, M. D., & Moritz, M. A. (2022). Rapid growth of large forest fires drives the exponential response of annual forest-fire area to aridity in the western united states. *Geophysical Research Letters*, *49*(5), e2021GL097131. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021GL097131` (e2021GL097131 2021GL097131) doi: https://doi.org/10.1029/2021GL097131

Keetch, J. J., & Byram, G. M. (1968). A drought index for forest fire control. *Res. Pap. SE-38. Asheville, NC: U.S. Department of Agriculture, Forest Service, Southeastern Forest Experiment Station. 35 p.*, *038*.

Lai, C., Zeng, S., Guo, W., Liu, X., Li, Y., & Liao, B. (2022). Forest fire prediction with imbalanced data using a deep neural network method. *Forests*, *13*(7). Retrieved from `https://www.mdpi.com/1999-4907/13/7/1129` doi: 10.3390/f13071129

Le, H., Hoang, A.-D., Trung Tran, C., Phi, Q., Tran, V.-H.-T., Hoang, N.-D., ... Bui, D. (2021, 04). A new approach of deep neural computing for spatial prediction of wildfire danger at tropical climate areas. *Ecological Informatics*, *63*, 101300. doi: 10.1016/j.ecoinf.2021.101300

Liang, H., Zhang, M., & Wang, H. (2019). A neural network model for wildfire scale prediction using meteorological factors. *IEEE Access*, *7*, 176746-176755. doi: 10.1109/ACCESS.2019.2957837

Mann, M., Batllori, E., Moritz, M., Waller, E., Berck, P., Flint, A., ... Dolfi, E. (2016, 04). Incorporating anthropogenic influences into fire probability models: Effects of human activity and climate change on fire activity in california. *PLoS ONE*, *11*. doi: 10.1371/journal.pone.0153589

Martell, D. L., Bevilacqua, E., & Stocks, B. J. (1989). Modelling seasonal variation in daily people-caused forest fire occurrence. *Canadian Journal of Forest Research*, *19*(12), 1555–1563.

Martell, D. L., Otukol, S., & Stocks, B. J. (1987). A logistic model for predicting daily people-caused forest fire occurrence in ontario. *Canadian Journal of Forest Research*, *17*(5), 394-401. doi: 10.1139/x87-068

Mayr, M., Vanselow, K., & Samimi, C. (2018). Fire regimes at the arid fringe: A 16-year remote sensing perspective (2000–2016) on the controls of fire activity in namibia from spatial predictive models. *Ecological Indicators*, *91*, 324-337. Retrieved from https://www.sciencedirect.com/science/article/pii/S1470160X18302759 doi: https://doi.org/10.1016/j.ecolind.2018.04.022

McArthur, A. G. A. G., Forestry, A., Bureau, T., & McArthur, A. (1967). *Fire behaviour in eucalypt forests* [Book; Book/Illustrated]. Canberra : Forestry and Timber Bureau. ("Paper submitted to the Ninth Commonwealth Forestry Conference, India, 1968.")

Mhawej, M., Faour, G., & Adjizian-Gerard, J. (2015). Wildfire likelihood's elements: A literature review. *Challenges*, *6*(2), 282–293. Retrieved from https://www.mdpi.com/2078-1547/6/2/282 doi: 10.3390/challe6020282

Milanovic, S., Marković, N., Pamucar, D., Gigović, L., Kostic, P., & Milanovic, S. (2020, 12). Forest fire probability mapping in eastern serbia: Logistic regression versus random forest method. *Forests*, *12*. doi: 10.3390/f12010005

Moreno, M., Malamud, B., & Chuvieco, E. (2011). Wildfire frequency-area statistics in spain. *Procedia Environmental Sciences*, *7*, 182-187. Retrieved from https://www.sciencedirect.com/science/article/pii/S1878029611001599 (Spatial Statistics 2011: Mapping Global Change) doi: https://doi.org/10.1016/j.proenv.2011.07.032

Muñoz Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., … Thépaut, J.-N. (2021). Era5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, *13*(9), 4349–4383. Retrieved from https://essd.copernicus.org/articles/13/4349/2021/ doi: 10.5194/essd-13-4349-2021

Nunes, M., Vasconcelos, M., Pereira, J., Dasgupta, N., Alldredge, R., & Rego, F. (2005, 09). Land cover type and fire in portugal: Do fires burn land cover selectively? *Landscape Ecology*, *20*, 661-673. doi: 10.1007/s10980-005-0070-8

Oliveira, S., Moreira, F., Boca, R., San-Miguel-Ayanz, J., & Pereira, J. (2013, 01). Assessment of fire selectivity in relation to land cover and topography. a comparison between southern european countries. *International Journal of Wildland Fire*, *23*, 620-630. doi: 10.1071/WF12053

O'Shea, K., & Nash, R. (2015). *An introduction to convolutional neural networks.* arXiv. Retrieved from https://arxiv.org/abs/1511.08458 doi: 10.48550/ARXIV.1511.08458

Pawar, K., & Rothkar, R. (2015, 12). Forest conservation environmental awareness. *Procedia Earth and Planetary Science*, *11*. doi: 10.1016/j.proeps.2015.06.027

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Prapas, I., Kondylatos, S., Papoutsis, I., Camps-Valls, G., Ronco, M., Fernández-Torres, M.-, … Carvalhais, N. (2021, 11). Deep learning methods for daily wildfire danger forecasting.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*, 195-204.

Reyes-Bueno, F., & Loján-Córdova, J. (2022). Assessment of three machine learning techniques with open-access geographic data for forest fire susceptibility monitoringmdash;evidence from southern ecuador. *Forests*, *13*(3). Retrieved from `https://www.mdpi.com/1999-4907/13/3/474`

Su, Z., Zheng, L., Sisheng, L., Tigabu, M., & Guo, F.-T. (2021, 08). Modeling wildfire drivers in chinese tropical forest ecosystems using global logistic regression and geographically weighted logistic regression. *Natural Hazards*, *108*. doi: 10.1007/s11069-021-04733-6

Van Wagner, C. (1987). Development and structure of the canadian forest fire weather index system. canadian forestry service. *Forestry technical report*, *35*, 37.

Vega-Garcia, C., Woodardl, P., Titus, S., Adamowicz, L., & Let, S. (1995, 01). A logit model for predicting the daily occurrence of human caused forest fires. *International Journal of Wildland Fire*, *5*, 101-111.

Vilar, L., Woolford, D., Martell, D., & Martín, M. (2010, 01). A model for predicting human-caused wildfire occurrence in the region of madrid, spain. *International Journal of Wildland Fire - INT J WILDLAND FIRE*, *19*. doi: 10.1071/WF09030

Wang, D., Guan, D., Zhu, S., Mac Kinnon, M., Geng, G., Zhang, Q., … Davis, S. (2021, 03). Economic footprint of california wildfires in 2018. *Nature Sustainability*, *4*. doi: 10.1038/s41893-020-00646-7

Zhang, G., Wang, M., & Liu, K. (2019). Forest fire susceptibility modeling using a convolutional neural network for yunnan province of china. *International Journal of Disaster Risk Science*, *10*, 386 - 403.

Zhang, G., Wang, M., & Liu, K. (2021). Deep neural networks for global wildfire susceptibility modelling. *Ecological Indicators*, *127*, 107735. Retrieved from `https://www.sciencedirect.com/science/article/pii/S1470160X21004003` doi: https://doi.org/10.1016/j.ecolind.2021.107735