Fire Safety Journal 104 (2019) 130–146

Contents lists available at ScienceDirect

Fire Safety Journal

journal homepage: www.elsevier.com/locate/firesaf

Predictive modeling of wildfires: A new dataset and machine learning approach

Younes Oulad Sayad[a,*], Hajar Mousannif[b], Hassan Al Moatassime[a]

[a] Cadi Ayyad University, Faculty of Sciences and Technologies, LAMAI Laboratory, Marrakesh, 40000, Morocco
[b] Cadi Ayyad University, Faculty of Sciences Semlalia, LISI Laboratory, Marrakesh, 40000, Morocco

ARTICLE INFO

Keywords:
Big data
Remote sensing
Machine learning
Wildfire prediction
Data mining
Artificial intelligence

ABSTRACT

Wildfires, whether natural or caused by humans, are considered among the most dangerous and devastating disasters around the world. Their complexity comes from the fact that they are hard to predict, hard to extinguish and cause enormous financial losses. To address this issue, many research efforts have been conducted in order to monitor, predict and prevent wildfires using several Artificial Intelligence techniques and strategies such as Big Data, Machine Learning, and Remote Sensing. The latter offers a rich source of satellite images, from which we can retrieve a huge amount of data that can be used to monitor wildfires. The method used in this paper combines Big Data, Remote Sensing and Data Mining algorithms (**Artificial Neural Network and SVM**) to process data collected from satellite images over large areas and extract insights from them to predict the occurrence of wildfires and avoid such disasters. For this reason, we implemented a methodology that serves this purpose by building a dataset based on Remote Sensing data related to the state of the crops (NDVI), meteorological conditions (LST), as well as the fire indicator "Thermal Anomalies", these data, were acquired from "MODIS" (Moderate Resolution Imaging Spectroradiometer), a key instrument aboard the Terra and Aqua satellites. This dataset is available on GitHub via this link (https://github.com/ouladsayadyounes/Wildfires). Experiments were made using the big data platform "Databricks". Experimental results gave high prediction accuracy **(98.32%)**. These results were assessed using several validation strategies (e.g., classification metrics, cross-validation, and regularization) as well as a comparison with some wildfire early warning systems.

1. Introduction

Wildfires are a serious problem that threatens to destruct thousands of square kilometers of forest every year. It is considered a worldwide disaster affecting various aspects of life, such as natural environments, economy, and health. A large number of fires are caused by humans, although other factors like drought, wind, lightning strikes, and topography have an important influence on fire occurrence and spread.

Wildfires present serious monitoring challenges. Their behavior is ambiguous and hard to predict, especially large, intense wildfires because they may combine complex meteorological scenarios, complicated topography, and complex fuel structures. Modeling and predicting wildfire behavior is a multi-disciplinary challenge addressed by many researchers from various backgrounds in engineering, ecology, physics, computer science, chemistry, mathematics, forestry, and other fields. Besides, predicting the occurrence of large wildfires is a computational challenge because the processes composing wildfires span a wide range of spatial and temporal scales that contribute substantially to these nonlinear phenomena. Another challenge consists of predicting the spread of a wildfire that has already started, which requires the collection of several weather parameters in real time (e.g., air temperature, wind speed, soil moisture), representing in itself a real challenge especially when collecting weather data in inaccessible and dangerous environments. Addressing all these issues could help land and fire managers in making decisions that could save the lives of firefighters, area residents, lands as well as reduce the costs of fire extinction [1].

The improvement of wildfire management lays on a better understanding of the scientific factors behind fire occurrence, behavior and spreading. Some of the primary keys are assessing fuel distribution, combustion factors, weather changes, and intensive energy production. Effective new tools and techniques could emerge from technological progress in wildfire prediction. Several improvements may be anticipated from these new technologies: firefighter assistance, advanced logistical services planning and forest management assessment of potential threats. One of the most interesting future improvements in wildfire prediction is the integration of spatial data into prediction

* Corresponding author.
E-mail address: oulad.sayad.younes@gmail.com (Y.O. Sayad).

https://doi.org/10.1016/j.firesaf.2019.01.006
Received 5 September 2018; Received in revised form 14 January 2019; Accepted 20 January 2019
Available online 24 January 2019
0379-7112/

models. Remote Sensing is becoming both more frequent and of increasingly higher resolution. These data have multiple applications for fires (e.g., the distribution and amount of fuel, water content and land cover, crop's health, and density). Another promising future direction is the combination of wildfire models with atmospheric weather models. This combination is helping researchers to understand the complex processes between the fire and the atmosphere. Furthermore, Internet of Things technologies such as wireless sensors and cameras can help improve wildfire monitoring systems in the future, by reaching remote, inaccessible and dangerous areas [2].

Wildfires monitoring can be more effective when combined with Artificial Intelligence technologies (AI), which will help researchers in building solid models for monitoring wildfires and detecting anomalies in real time. Many recent AI technologies have been widely applied for developing new wildfires prediction systems. These technologies include Big Data, Machine Learning, and Data Mining among others [3].

- Big Data refers to a collection of extremely huge, unstructured and continuously growing datasets that it is beyond the capacity of standard data management tools to capture, store, manage and analyze. The term not only refers to data, but also to the various frameworks, tools, and techniques involved.
- Data Mining [9] is a process for extracting useful information and knowledge from huge, unstructured and random data (Big Data). It combines statistical analysis, Machine Learning, and database technology to retrieve hidden patterns and relationships from large databases.
- Machine Learning (ML) is a computational study of algorithms based on automated learning approaches. It uses algorithms to acquire knowledge from data. The basic of ML is to build algorithms that can receive input data and use statistical analysis to predict new entries.

There are several machine learning algorithms to discover patterns in big data that lead to actionable insights. These different algorithms can be classified into two categories based on how they "learn" from data to make predictions: supervised and unsupervised learning.

Supervised learning (SL) is the most commonly used category. In this category, algorithms learn from the data what conclusions they should draw. SL requires that the algorithm's possible outputs are already known and that the data used to train the algorithm are already labeled with correct answers. SL uses classification and regression techniques to develop predictive models. Classification techniques predict categorical responses, while regression techniques predict continuous responses. SL includes algorithms such as linear and logistic regression, Artificial Neural Networks, and Support Vector Machines [10].

On the other hand, Unsupervised Learning (UL) algorithms are mostly used in exploratory data analysis because it is used to determine hidden patterns in data from unlabeled input data. Clustering is the most common unsupervised learning technique. It consists of dividing the dataset into multiple clusters with the same size. Some examples of unsupervised learning algorithms include k-means clustering, and association rules [10].

The primary difference between supervised and unsupervised learning is the data used for training. The input data used in supervised learning is well known and labeled, while in unsupervised learning the data is not known nor labeled.

The increased use of Artificial Intelligence requires the use of powerful software tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes more and more difficult. Many big companies have integrated AI solutions into their Database products, such as: Oracle (Tool: Oracle Data Mining), SAS (Tool: SAS Enterprise Miner), SPSS (Tool: IBM SPSS Modeler), Microsoft (Tool: SQL Server Analysis Services), Teradata (Tool: Teradata Database, former name TeraMiner) and TIBCO (Tool: TIBCO Spotfire) [15,16]. Artificial Intelligence tools are making use of big data by collecting, pre-processing and analyzing data to extract useful information. These tools have the potential to address many problems including disasters monitoring, crops, and water management [4], among others. Wildfires can be characterized by big data's three features: volume, variety, and velocity, defined as three "V" dimensions [5].

1) **Volume:** Huge amount of data is generated every day. This big volume of data has increased from terabytes to petabytes, and even to Exabytes. Wildfires monitoring requires the manipulation of large volumes of data coming from different sources: Remote sensing generates tons of valuable data for wildfires (e.g., NDVI, LAI, and LST) obtained from different satellites (e.g., Terra, Landsat, Sentinel), at multiple time spans (e.g., hourly, daily, weekly). In-situ weather stations also collect valuable data for wildfire monitoring (e.g., temperature, wind, moisture, precipitation). Another rich source of data that can be used in fighting wildfires is sensors which can be embedded in forests and collect data in real-time. All the mentioned data sources generate loads of potentially valuable data for wildfire monitoring [6].
2) **Variety:** There is a wide variety of data that can be used to monitor wildfires, satellite images generate various data with different aspects: multi-sources (e.g., laser, radar, optical), multi-temporal (collected on different dates), and multi-resolution (different spatial resolution) [7]. Weather stations also generate different kind of data (e.g., Air Temperature, Humidity, Precipitation, Solar radiation, Wind speed, and Direction), with different formats (e.g., fluctuation, numerical, wave). These data are essential in predicting the occurrence and spreading of wildfires.
3) **Velocity:** Big Data are undergoing fast growth, estimated to 4 TB daily. The velocity of big data goes beyond data generation at a rapidly growing rate. Big Data velocity not only involves high generation rate, but also the efficiency of data processing and analysis. In other words, the data should be analyzed on a real and reasonable time to achieve a given task. In the context of wildfires, seconds can save hundreds of thousands of lives, by processing the collected data in real time.

Remote Sensing is one of the main sources of big data. This rapidly developing technology offers the advantages of consistent, repeatable, large-area coverage, and can easily provide a massive amount of data from remote regions. In addition, Remote Sensing data represent the integrated response of the vegetation to the different factors influencing its status as a fuel base for wildfires. Several parameters related to wildfires can be extracted using Remote Sensing, including The Normalized Difference Vegetation Index (NDVI), which is a vegetation index that indicates the state of crop health; it can be used to assess spatiotemporal changes in green vegetation. Leaf Area Index (LAI) represents the projected green leaf area over a large ground area; it is used to characterize canopy light conditions. Thermal Anomalies (TA) are a fire detection strategy based on complete detection of a fire (only when the fire is strong enough to be detected), and Land Surface Temperature (LST) represents the radiative skin temperature of the land surface derived from solar radiation, it depends on the vegetation cover and the soil moisture.

Wildfires can be managed using Artificial Intelligence technologies (e.g., Remote Sensing, Big Data, and Machine Learning). However, some technological and natural limitations can make it difficult to detect fire occurrence and can lead to inaccurate data. This includes dense cloud covers, technical problems on the satellite, on the ground system, on the image distribution service, dense forest, fires that occur in the period between images and the fires that occur outside the satellite's field of view. Besides, wildfires management still suffers from a lack of data. Scientific researchers tend to rely on Artificial Intelligence and having enough data quality to build models that provide meaningful learning and results. However, AI cannot be effective with small amounts of data. Machine Learning algorithms require big and

meaningful data in order to establish a knowledge base and be able to discover new entries. Researchers looking to adopt AI often find difficulties in aggregating data across multiple sources; they are also limited by missing and incomplete data. Luckily, new methods and techniques are helping wildfires researchers overcome these challenges.

Several Artificial Intelligence solutions have been developed to deal with wildfire problems. For example, authors in Ref. [11]] adopted Neural Networks (NN) to predict human-caused wildfires. Infrared scanners and NN were combined in Ref. [12] to reduce wildfires false alarms with a 90% precision. In Ref. [13], a spatial clustering (FASTCiD) was adopted to detect wildfire spots in satellite images. In 2005, satellite images from North America wildfires were fed into a Support Vector Machine (SVM), which obtained 75% accuracy at detecting fire at the 1.1-km pixel level [14].

Authors in [17] have applied multiple data mining techniques such as Logistic Regression, Random Forest (RF) and Decision Trees (DT) to predict wildfires in different regions of the Slovenian forests, using both satellite-based and meteorological data. These data are divided into three groups: geographic data, multi-temporal MODIS data and meteorological data (temperature, humidity, evaporation, transpiration wind speed and direction). Along with the collected data, the authors needed positive and negative samples of fire occurrence to build predictive models of wildfires. Those samples were located in the past, where wildfires occurrences were noticed along with the date and hour. Negative samples are represented by an equal number of occurrences with random timestamps and location. The best model was obtained by Bagging of decision trees that gave the best results in terms of predictive accuracy, precision and kappa statistics compared to the other algorithms, with an overall 80% accuracy.

In [18] the authors proposed a data mining approach that uses meteorological data, collected in real-time by sensors in local weather stations. The cost of the collected data is low compared with the high-resolution satellite images. The dataset also included spatial and temporal components from the Canadian Fire Weather Index (FWI) and four weather conditions. The aim of this work was the prediction of the burned area in the northeast region of Portugal using five different data mining algorithms, including Support Vector Machines (SVM) and four feature selections. The proposed solution required four weather parameters capable of predicting small fires (temperature, precipitation, relative humidity, and wind speed), which constitute the majority of the fire occurrences. However, this solution gives low predictive accuracy for large fires. Nevertheless, the proposed model is used to improve firefighting resources management in case of small fires, or the early stages of large fires. Such management would be very helpful in dramatic fire seasons when simultaneous fires occur at different locations.

Another approach in [19] reviewed the research strategy behind the development of the Canadian Forest Fire Behavior System (FBP) (Forestry Canada Fire Danger Group 1992). This system is used across Canada in fire management operations, and it is unique in two ways. First, the FBP strategy makes separate estimates of spread rate and fuel consumption and then combines them to obtain fire intensity. Second, the FBP is also known as one of the few fire behavior systems to be based on a large set of field data that are analyzed and presented in a framework. The results are generally realistic and of good value in fire management. However, the current version does contain one drawback in its physical logic. The problem lies in the research philosophy behind the FBP System, careful collection of field data, first analyzed by graphical methods, set the stage, followed by the increasing use of the physical principle as a framework for statistical correlation analysis, but never to supplant the empirical evidence of the field data themselves.

With respect to all the models used in the literature, the existing efforts provided general models to predict the occurrence and spreading of wildfires based on Remote Sensing, weather data, and climate parameters. However, it is also necessary to take into account the different causes of wildfires, the aspects of data processing and geographical location. The present work puts forward a strategy that makes use of big data and remote sensing to build a dataset, later processed by Data Mining algorithms to predict the occurrence of wildfires. The proposed strategy is composed of seven steps ranging from data collection to data extraction. The data were acquired from MODIS (Moderate Resolution Imaging Spectroradiometer) [23], a sensor embedded in both Terra and Aqua satellites. These data were made available by NASA's Land Processes Distributed Active Archive Center (LP DAAC) 20]. We have chosen to use this sensor because it covers the entire earth and provides a multitude of data products (e.g., NDVI: MOD13Q1, LST: MOD11A1, TA: MOD14A1) in regular and continuous time spans. Thus, the model can be applied anywhere in the world (A guide of the applicability of the strategy is presented in section 5.3). The originality of this paper lies within the fact of it seeking the development of a multi-disciplinary model for predictive analysis based on big remote sensing data and data mining. While doing so, this paper attempts to provide satisfying answers to the following questions:

- What are the necessary data for wildfire occurrence prediction?
- How can we build a dataset based on remote sensing data?
- How can we process our dataset using data mining algorithms?
- How can we improve the predictability of our model?

In this paper, an experiment has been set up to analyze the created dataset in order to predict the occurrence of wildfires in a specific region of Canada's forests between 2013 and 2014. The fire zones were acquired from The Canadian Wild-land Fire Information System (CWFIS) [20]. The constructed dataset is composed of 804 instances (386 fire instances and 418 no_fire instances). The experiment is considered as a case study to illustrate what can be done at larger scales, for this reason, the dataset used in the simulation will not contain many instances. In this experiment, we used two of the most known data mining algorithms (Artificial Neural Networks and Support Vector Machine), these algorithms were implemented in “Databricks”, a big data platform. 70% of the data were used for training and the remaining 30% for testing. The results showed high fire occurrence prediction accuracy (ANNs: 98.32%, SVM: 97.48%). The model was also validated using classification metrics, cross-validation, regularization as well as a comparison with some existing wildfire models.

The remainder of this paper is organized as follows: Section 2 describes our research methodology. Section 3 highlights the data processing and validation strategies. In section 4 we present the experiment. Section 5 discusses the results. Finally, conclusions and future works are given in section 6.

2. Research methodology

This section is dedicated to present the proposed methodology for building the dataset that will be used for wildfire occurrence prediction. The process of building our dataset is composed of multiple steps, starting from collecting numerous data, applying multiple preprocessing techniques to remove noises and correct inconsistencies, and finally extracting useful information. Big Data and remote sensing coupled with data mining algorithms are now allowing researchers to predict the risk of wildfires accurately and thus allow their prevention.

In order to create our dataset, we applied the methodology proposed in [21], which is a generic methodology for building any big data project, and adapted it to the context of wildfire prediction. Our approach consists of multiple steps ranging from data collection to data extraction as shown in Fig. 1.

2.1. Data Collection

The first step in the process of building the dataset is to collect pertinent data that can represent the appearance of wildfires; many parameters can be used in this matter. These parameters are related to the state of the crops, the state of the soil and numerous meteorological

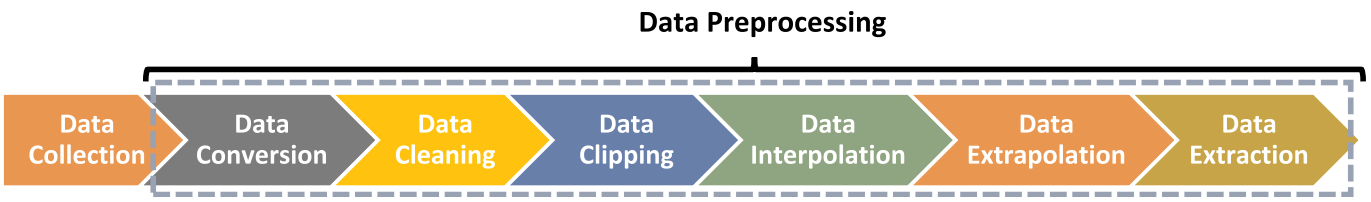

Fig. 1. The proposed methodology for building our dataset.

data influencing wildfires. In this study, we selected three parameters related to three aspects (crop's health, soil temperature, and a fire indicator).

The first parameter is the Normalized Difference Vegetation Index (NDVI) which is considered to be the most popular and most used vegetation index for crops, the second parameter is LST (Land Surface Temperature) which is the soil temperature, and finally, the fire indicator (Thermal Anomalies).

The choice of these three parameters was conducted by the nature of the issue addressed by the paper, which is predicting the occurrence of wildfires caused naturally by heat or lightning, in both cases the state of the crops, as well as the soil, play a primary and important role in catching fire; a dry crop in a dry and heated soil is more likely to catch fire than a wet crop in a moisturized land. We added the third parameter (i.e., Thermal Anomalies) to strengthen the first two because it gives direct information about the fire with a detection confidence when the fire is strong enough to be detected. After validating the model, additional parameters will be added in future works to improve the final wildfire monitoring model.

All these data can be extracted from different satellites such as Terra, Aqua, Landsat, and Aster. Each satellite contains a different product, and these products differ in their sizes, their spatial resolution, and spectral range. Below an overview of the three parameters used.

2.1.1. The Normalized Difference Vegetation Index (NDVI)

The normalized difference vegetation index (NDVI) is an index of a crop's photosynthetic activity "greenness" and is considered as the most widely used vegetation index. It is the fraction of the red and near-infrared reflectance, and it is useful for evaluating the health, and density of vegetation. This index has been used in many applications, including estimation of crop yields, water resources management, drought and wildfires prediction among others. The two channels used in an NDVI sense through different depths of vegetation canopies. The near Infra-Red channel can see about eight layers, while the red channel sees only one layer due to the strong absorption of chlorophyll. NDVI values are between 0 and 1, values near 0 indicate very sparse vegetation and values near 1 indicate dense vegetation.

Multiple other vegetation indices with varying complexity have been developed using the same set of near-IR and red channels during the past decade. These indices do not have the same popularity as that of NDVI. The recently developed Atmospherically Resistant Vegetation Index (ARVI) uses a blue channel near 0.47 pm alongside the near-IR and the red channels. This index has self-correction properties for the atmospheric effect [22], NDVI can be very useful regarding predicting wildfire occurrence because it indicates whether the crop is dry or wet: a dry crop is more flammable than a wet one. The NDVI is available using Remote Sensing and can be extracted from different Satellites. In this work, we used the satellite Terra. This satellite contains many sensors (e.g., ASTER, CERES, MISR, MODIS, MOPITT). Each sensor collects multiple data known as products. The NDVI product reference is "MOD13Q1". It provides regular spatial and temporal comparisons of vegetation conditions. "MOD13Q1" data are provided every 16 days at a 250-m spatial resolution as a gridded level-3 product in the Sinusoidal projection. The EVI algorithm uses the 500 m blue band to correct for residual atmospheric effects, with negligible spatial artifacts [20].

2.1.2. The Land Surface Temperature (LST)

Land Surface Temperature (LST) represents the radiative temperature of the ground. It is the combination of vegetation and bare soil temperatures and is used to detect water-stressed crops; the higher it gets, the more stressed crops are [20]. It is also considered as one of the most influencing parameters in the physics of land-surface processes on global scales, combining the results of all surface-atmosphere interactions and energy fluxes between the atmosphere and the ground. LST can be retrieved from satellite data and can be used in multiple applications. One of the most important applications of the LST is to improve the global meteorological prediction model after parameterization, in land-cover change analysis, in the production of the Moderate Resolution Imaging Spectroradiometer (MODIS) land cover product and in estimation and parameterization of surface fluxes. LST can also be used to monitor drought, predict wildfires, to estimate surface soil moisture, to evaluate water requirements of wheat and to determine frosts in different crops. “MOD11A1” is the MODIS product name of the Land Surface Temperature (LST); this product provides daily data extracted at 1 km pixels by the split-window algorithm and 6 km grids by the day/night algorithm [24].

The MODIS LST products are archived in Hierarchical Data Format-Earth Observing System (HDF-EOS) format files. The LST product files contain global attributes and scientific datasets (SDSs) with local attributes. SDSs are local attributes including the coefficients of calibration which converts the SDS value to real LST value in Kelvin or Celsius. HDF predefined local attributes (Table 1) to describe the characteristics of the data.

2.1.3. Thermal Anomalies (TA)

This parameter gives direct information on fire when it occurs (only when the fire is strong enough to be detected). This parameter is available at 1-km resolution, and composed of fire pixels detected in each grid cell daily and can be extracted from the MODIS product “MOD14A1”. The fire detection strategy is based on complete detection of fire, and on detection relative to its background (to take into account the variability of the surface temperature and sunlight reflection). Multiple tests are employed to reject typical false alarm sources like sunlight. The MODIS daily and 8-day Level 3 fire products are tile-based products, with each product file spanning one of the 460 MODIS tiles, of which 326 contain land pixels. The 8-day composite represents the maximum value of the individual Level 2 pixels that fell into each 1-km grid cell over the 8-day compositing period [25].

2.2. Data preprocessing

After data are collected, it is important to lay the ground for data analysis by applying various preprocessing operations to address potential imperfections in the raw collected data. Data Collection methods and conditions are not flawless, especially in the context of Remote Sensing data where faulty equipment (sensors' constrained resources and intermittent loss of connection.) or human's inattention can lead to a noisy dataset containing errors, redundancies, and outliers. There may also be missing data due to multiple reasons, such as cloud covers, technical problems on the satellite or the image distribution service. Finally, data may need to fit the requirements of analysis algorithms. As a consequence, collected data must be inspected, fused and all the

Table 1
Land Surface Temperature data characteristics.

Attribute Name	Reserved Label(s)	Definition	Sample Value
SDS Name	Short Name	Short Name of the SDS	LST
Label	Long Name	Long Name of the SDS	Land Surface Temperature
Format	Number Type	How the data are stored	uint16 (16-bit unsigned integer)
Unit	Units	SI units of the data	K (Kelvin)
Range	Valid Range	Max and min values within a selected data range	7500–65535
Fill Value	Fill Value	Data used to fill gaps in the swath	0
Calibration	Scale Factor	Scaling factor	0.02
	Add Offset	Add offset	0

The data range of SDS is between 7500 and 65535; these values have to be calibrated and converted into Kelvin or Celsius. The calibration formula for the "LST" SDS is.
Kelvin: LST = (SDS * scale factor) K.
Celsius: LST = (SDS * scale factor) - 273.15 ˚C.
Example: for SDS = 14584, LST = (14584 * 0.02) – 273.15 = 18.53 ˚C.

above problems treated during a preprocessing phase. It is clear that insights extracted from "dirty data" (i.e., unreliable data) are probably erroneous. Thus, the decisions to be made are likely unsound. The preprocessing step aims to investigate many data-enhancement approaches, such as Data Conversion, Data Cleaning, Data Clipping, Data Interpolation and Data Extrapolation [26].

2.2.1. Data Conversion

The First step in data preprocessing is data conversion. This step consists in converting the collected data into a readable format. MODIS products have two types of metadata: the embedded HDF metadata and the external ECS metadata. The HDF metadata contains valuable information including global attributes and dataset. The dataset attributes contain specific information such as the data range and applicable scaling factors for the data. A hierarchical data format is essentially a format that groups various aspects of a dataset (e.g., metadata, raster data) into a binary file. There are many tools for manipulating, converting and reading HDF:

- **GDAL**: Geospatial Data Abstraction Library (GDAL) is a translator library for raster geospatial data formats. GDAL can be used to convert HDF file to a raster data using two commands: 1) gdalinfo: to report information about a file, 2) gdal_translate: to convert a raster file, with control of output format.
- **HEG**: The HDF-EOS to GeoTIFF Conversion Tool (HEG) is a tool that allows a user to reformat, re-project and perform subsetting operations on HDF-EOS objects. The output GeoTIFF file can be used in most GIS applications. HEG works with MODIS (AQUA and TERRA), ASTER, MISR, AIRS, and AMSR-E HDF-EOS datasets. It also handles OMI HDF-EOS5 grid and SMAP L3/L4 HDF5 datasets.

In this step, we used the GDAL library to convert the HDF metadata collected into raster data (GeoTIFF) using the “gdal_translate” utility. We have chosen to use GDAL because it has a wide range of image processing utilities, and it can be used in many programming languages. Hence, converting multiple satellite images sequentially.

2.2.2. Data Cleaning

After converting the collected metadata into raster data, the second step of data preprocessing consists of applying various data cleaning techniques to correct imperfections in raw collected data. Imperfections on satellite images can be caused by cloud covers, technical problems on the satellite, on the ground system or the image distribution service. In this case, the satellite image may be damaged and must undergo several cleaning techniques. Cleaning multispectral satellite data presents a significant challenge due to the following reasons:

- **Clouds and atmospheric conditions:** Cloud cover and shadows can make the data unreadable, given reflectance values are either too bright (due to sunlight reflectance on the sensor) or too dark (shadows which represent clouds blocking the sensor or absorbed light).
- **Geometric Distortions:** satellite images may contain some geometric distortions due to changes in sensor position, Earth rotation on its axis while recording images and finally due to ground effects. Some distortions are predictable and can easily be corrected, while others are complicated and difficult to handle.

For this reason, data cleaning is a crucial step before data analysis, in this step, we can apply many data cleaning techniques such as radiometric calibration, geometric correction, and atmospheric correction to remove noises and correct inconsistencies. Below a set of techniques and methods that can be used for satellites data cleaning [27].

a) **Geo-referencing:** is a Geometric correction method that aims to convert the satellite images' geographic coordinates to a normalized coordinate system so it can be displayed and analyzed with other geographic data. Geo-referencing includes multiple operations such as shifting, rotating, scaling, and warping (Ortho-rectifying). Many raster data must be geo-referenced before they can be processed with other geographic data. For example, the Earth Explorer historical satellite images are generally not Geo-referenced
b) **Ortho-rectification**: is also a Geometric correction process for correcting raster data distortions from sensor tilt and the Earth's ground. Ortho-rectification requires rational polynomial coefficients (RPCs) and an accurate digital elevation model (DEM). RCPs are usually provided by the satellite images suppliers and require knowledge of the sensor. If the RCPs are not provided, it is possible to use some specialized programs designed to generate and complete the ortho-rectification process.
c) **Radiometric Corrections**: Radiometric correction aims to correct errors in raster data to improve their quality and ease their interpretability. The atmospheric conditions, as well as the elevation, can influence the observed energy by sensors which might differ from the real energy emitted or reflected from a surface on the ground. Thus, it is necessary to use radiometric correction to obtain the real ground radiance or reflectance.
d) **Atmospheric Correction**: Atmospheric correction is the process of producing surface reflectance values by removing the effects of the atmosphere. This operation improves the satellite images interpretability and usage. Preferably, this process requires information about the atmospheric conditions as well as the aerosol properties when acquiring the image.

In this step, we did not use Geo-referencing and Ortho-rectification techniques, because the collected data from “MODIS” were already Geo-referenced and not distorted. However, some satellite images

Table 2
The most popular methods for data interpolation.

Categories	Methods	PROS	CONS
Deterministic methods	**Nearest Neighborhood (NN)**: Assign the value from the nearest observation to a particular grid cell.	Fast and simple	The interpolated fields do not look realistic in all cases.
	Inverse Distance Weighting (IDW): IDW is an advanced nearest neighbor approach. The value at a specific date is obtained from a linear combination of the surrounding locations.	Fast, universal and easy to implement	Does not Support Ancillary data
	Linear regression: identify the relationship between a predicted variable and one or more explanatory variables.	Ancillary data included	Can be stochastic in some cases.
Probabilistic methods	**Optimum interpolation**: Based on a spatial correlation function, requires a first guess field like the model output from numerical weather prediction models.	Minimize the expected interpolation error	Difficult to define the error covariance accurately
	Kriging: Kriging is a geostatistical method for spatial interpolation. Kriging can evaluate the quality of prediction with estimated prediction errors. The three fundamental Kriging methods are simple Kriging, ordinary Kriging, and universal Kriging.	Provide a reliable interpolation	Complex procedure, many steps, many version
Other methods	**MISH**: Incorporates information from time series in the interpolation procedure. It consists of two modules: MISH for interpolation and MASH to obtain homogenized data series.	Include ancillary Data	Require many meteorological information
	PRISM: uses point measurements of temperature, precipitation, and other climate values to produce continuous, digital coverage. It is used with Geographic Information Systems (GIS) to build maps and do many types of analysis.	Very powerful in dense networks	Include many parameters, requires great skill about local climatic conditions

contained incorrect values due to wrong energy reading from the sensor, which required radiometric corrections. The weather conditions also affected the satellite images, cloud covers hindered sensors from capturing correct values, for this reason, we used atmospheric correction to remove the effects of the atmosphere. In some cases, none of the data cleaning techniques helped, because the images were either too damaged or did not contain any data, in these cases the images were deleted.

2.2.3. Data Clipping

After applying various data cleaning techniques and removing multiple imperfections in raw collected data, the next step is to clip or subset the satellite images in order to include only the burned areas. This operation is crucial for the crops data abstraction from the satellite images. However, many difficulties can be faced by performing this operation. The main problem is the image volume; the higher the resolution of the image is, the more it is difficult to be processed as few tools and platforms support processing large images. The most known library for satellite image processing is the GDAL library.

In this step, we used the GDAL utility “gdalwarp”. It can be used for image mosaicing, reprojection, and warping. To clip satellite images into the shape of the burned areas, “gdalwarp” requires two parameters as inputs (the shapefile of the burned areas and the satellite image to be clipped). The “gdalwarp” utility was also used in the “python” programming language to clip all the satellite images sequentially.

2.2.4. Spatio-temporal data interpolation

After generating the raster data matching the shape of the Burned areas, we move to data interpolation, which consists in normalizing the data by keeping only one timestamp, preferably "daily". The collected satellite images have different timestamps, (NDVI: 16 days, LST: daily and Thermal Anomalies: 8 days), so it is necessary to interpolate the images for both NDVI and Thermal Anomalies to have the same timestamp as LST (daily). This operation is called, "spatiotemporal interpolation", which consists in creating satellite images on missing time locations based on existing satellite images. Several libraries and data interpolation methods can be used in this step. Data interpolation methods can be classified into three categories: **deterministic**, **probabilistic** and **other methods**.

- The **deterministic methods** create a continuous surface by using the geometric characteristics of point observations.
- The **probabilistic methods** use a probabilistic theory; they allow including the variance in the interpolation process and computing the statistical significance of the predicted values.
- The **other methods** are specially developed for meteorological purposes using both deterministic and probabilistic methods.

Table 2 summarizes some of the most popular methods in each category [29]:

In this step, we interpolated the clipped images using the deterministic method “Inverse Distance Weighting (IDW)” because it gives the most realistic interpolated values. In addition to being universal, easy to implement, fast and can also be implemented in a wide range of programming languages.

2.2.5. Spatio-temporal data extrapolation

Once spatiotemporal data interpolation is complete, we move to a very important step in the process of building the dataset which is Data Extrapolation. This operation consists of data estimation based on the collected data in order to forecast wildfires. Extrapolating satellite images requires the use of spatiotemporal data extrapolation methods, which consists in creating new satellite images on future time locations based on existing ones in order to forecast wildfires. The extrapolated satellite images are considered approximate estimates of the future based on previous satellite image time series using spatiotemporal extrapolation methods.

Extrapolation methods are quantitative because they use past data to forecast future values. Many extrapolation methods are available, including trend-based regression and auto-regression. All of the existing extrapolation methods search for existing values and then extrapolate these values into the future within a short or long range. Table 3 below summarizes some of the most popular extrapolation methods [29]:

To obtain approximate estimates of future conditions, we extrapolated our satellite images using the “Temporal (trend) extrapolation” method. It is suitable for our model thanks to the presence of historical data (already collected and interpolated). The method is also easy to implement and has lower data requirements.

2.2.6. Data extraction

Once spatiotemporal data interpolation and Extrapolation are finished, we move to the final step of the methodology which is data extraction; this stage produces various data products including NDVI, LST, and Thermal Anomalies. For example, to calculate the NDVI values, crops' pixels are extracted "one by one" and their average is calculated.

Table 3
The most popular methods for data extrapolation.

Methods	Principle	PROS	CONS
Rule-Based Forecasting	- Uses judgment to develop and apply rules for combining extrapolations. - Uses domain knowledge to combine forecasts from various extrapolation methods.	More accurate than traditional extrapolation methods for long-range forecasts	Weak accuracy in the absence of domain knowledge More expensive to develop and use
Temporal (trend) extrapolation methods	Applies a set of predefined principles or expectations based on a prior understanding of the system, together with recent data, to interpret future developments.	- Easy to obtain approximate estimates of the future - Has lower data requirements	Unreliable predictions in the absence of historical data
Auto-regression	- Predict time series and perform prediction error filtering. - Replace missing or corrupted samples with estimates of their true samples.	Good performance even with random noise	Captures only linear relationships

For the MODIS products "MOD13Q1" the NDVI is already calculated in each pixel. For other products such as LANDSAT and ASTER, the NDVI is more difficult to calculate. For example, LANDSAT's Thematic Mapper (TM) instruments include seven spectral bands, including a thermal band. In order to calculate the value of NDVI, we use the third band (Visible Red) and the fourth band (Near-Infrared) and calculate the NDVI according to this formula: NDVI = (NIR-RED)/(NIR + RED) [28].

In this step, we extracted the values of the parameters from the satellite images using the Python Imaging Library (PIL). The first step of data extraction consists of loading the satellite images, then, reading the images, and finally extracting their values pixel by pixel and computing their average. This operation is done for all the satellite images.

3. Data processing and validation strategies

3.1. Data mining algorithms

In order to predict the occurrence of wildfires, we need to use data mining algorithms. Multiple algorithms can be used for wildfire occurrence prediction using data mining techniques. Some classification algorithms such as neural networks, naïve bayes, SVM, and random forests, have been already applied to develop predictive models of fire occurrence based on satellite data. Table 4 below describes some of the most known data mining algorithms.

In this study, we used some of the best data mining algorithms to process the created dataset. Those algorithms are ***Neural Networks and SVM***. Each algorithm uses some specific functions to load the data, train the model and display the results. Below the classifier functions used for each algorithm.

3.1.1. Neural Networks

"MLPClassifier" [36] is a Neural Networks classifier that implements a multi-layer perceptron (MLP) algorithm that trains using **Backpropagation**. The classifier has 21 parameters (e.g., hidden_layer_sizes, activation, solver). It can also have a regularization parameter (alpha) added to the loss function that penalizes some model parameters to prevent overfitting.

3.1.2. SVM

The SVM algorithm uses the C-Support Vector Classifier **"SVC"** [38] for classification which is implemented based on the Library for Support Vector Machines (libsvm). The SVC function has 14 parameters (e.g., C, kernel, degree).

3.2. Simulation steps

The simulation of the dataset is composed of three steps (see Fig. 2) :

1. **Model Training:** Each algorithm uses a specific function to train the model **"Neural Networks: MLPClassifier(parameters) and SVM: SVC(parameters)"**, and then call the function **fit(X, y)** to fit the model to data matrix X and targets y.
2. **Prediction:** predict(X) is a function that predicts target values of the test data (X) given in parameter.
3. **Evaluation:** prediction evaluation is done using the score function **"score(X, y)"** which returns the mean accuracy of the given test data (X) and labels (y). **classification_report(y_true, y_pred)** is also used for evaluation. It builds a text report showing the main classification metrics (precision, recall, and f1-score).

3.3. Model validation

In order to validate the stability of our model, it is required to estimate the error rate after training; this operation is called evaluation of residuals. However, this only gives an idea about how well our model learns from the data used for training. Low performance in Machine Learning may have different reasons (e.g., over-fitting and under-fitting [39]). To overcome these problems, multiple operations and techniques can be used to increase the performance of the classifiers such as classification metrics, cross-validation, and regularization.

3.3.1. Classification metrics

Classification metrics are a set of metrics used to validate the model. These metrics are calculated based on a confusion matrix. The confusion matrix is a representation of the prediction results on a classification problem. It summarizes the number of correct and incorrect predictions for each class.

Table 5 above describes a confusion matrix with two classes (class 1: positive and class 2: negative), and four cells explained as follows:

- **True Positives (TP):** cases where class 1 is correctly predicted.
- **False Negatives (FN):** cases where class 1 is incorrectly predicted.
- **False Positives (FP):** cases where class 2 is incorrectly predicted.
- **True Negatives (TN):** cases where class 2 is correctly predicted.

Based on the confusion matrix, classification metrics can be calculated. The performance of the model is assessed by these metrics to describe the accuracy of the classification:

- **The True Positive Rate (TPR)** known also as (sensitivity or recall) is the proportion of positive instances (fire) correctly predicted. This metric is calculated using the following formula: TPR = TP/(TP + FN).
- **The True Negative Rate (TNR) or (specificity)** is the proportion of negative instances correctly predicted, TNR is computed using the following formula: TNR = TN/(FP + TN).
- **The False Positive Rate (FPR) or (1-specificity)** is the fraction of negative samples that are predicted as positive.
- **The False Negative Rate (FNR**) is the fraction of positive samples that are predicted as negative.
- **Precision or (Positive predictive value**) (TP/(TP + FP)) is the proportion of positive prediction which is actually positive,

Table 4
Data mining algorithms.

Description	Performance	Limitations
- **Bagging of decision trees** creates several subsets of data from the training sample chosen randomly with replacement. - It is used to decrease the variance of a decision tree. - Each collection of the subset data is used to train decision trees. - After training, we obtain different models. The final prediction is the average of all the model predictions, which is more robust than a single decision tree.	- Handles higher dimensionality data very well. - Handles missing values and maintains accuracy for missing data.	- Since the final prediction is based on the mean predictions from subset trees, it will not give precise values for the regression model.
- **Boosting of decision trees** iteratively constructs a series of decision trees. - Each decision tree is trained and pruned on examples that have been filtered by previously trained trees. - It fits randomly consecutive trees at every step; the goal is to solve for net error from the prior tree.	- Supports different loss function. - Works well with interactions.	- Prone to over-fitting. - Requires careful tuning of different hyper-parameters.
- **K-nearest neighbor** classification is one of the simplest algorithms in data mining. - It memorizes the entire training set and performs classification only if the attributes of the test match exactly one of the training examples [34].	- KNN is easy to implement and can perform well in many situations. - KNN is particularly appropriate for multi-modal classes.	The drawback of this method is that many test records will not be used in classification because they do not exactly match any of the training set.
- **The Naive Bayes** also called simple Bayes, belongs to the supervised classification family. - It allows constructing rules which will enable predicting classes for future objects: based on a given set of objects; each one belongs to a known class.	- Naïve Bayes is easy to construct, interpret and may be applied to large datasets. - Does not require complicated iterative parameters.	- Require computing several conditional probabilities. - Impossible to compute the probabilities by traditional methods when an attribute is continuous.
- **Support vector machines** (SVM) [37] are a set of supervised learning algorithms used for classification, regression and outliers' detection. - SVM aims to determine the best classification method to distinguish between members of two classes in the training data. - The determination of the "best" classification method can be realized geometrically.	- It requires only a couple of sets for training. - It offers one of the most accurate methods among all well-known algorithms.	- SVMs are slow in both the training and testing phases. - SVMs are so complex and require extensive memory in large-scale tasks.
- **K-means** is an unsupervised algorithm that uses an iterative method to divide a given dataset into several clusters noted as "k". - The clusters are then positioned as points, and all observations or data points are associated with the nearest cluster, computed, adjusted, and then the process starts over using the new adjustments until the desired result is reached.	- Easy to implement. - Fast and efficient in terms of computational cost. - Easy to interpret the clustering results.	- The algorithm is quite sensitive to initialization. - The algorithm is also sensitive to the presence of outliers, since "mean" is not a robust statistic.
- **Artificial Neural Networks** (ANNs) [35] are a set of supervised learning algorithms that learn iteratively by training on a dataset. - ANNs are widely used for solving various classifications and forecasting problems. - ANNs consist of three layers: an input layer, an intermediate layer, and an output layer.	- Storing information on the entire network - Ability to work with incomplete knowledge - Perform more than one job at the same time.	- Require processors with parallel processing power. - The difficulty of showing the problem to the network.

- **Accuracy** (TP + TN/TP + FP + FN + TN) is the ratio of correctly predicted instances to the total instances.
- **F-score or F-measure** can be interpreted as the average of the precision and recall. F-score reaches its best score at 1 when all the predicted instances are correct, and thus the precision and recall are equals. The formula for the F1 score is: F-score = 2 * (precision * recall)/(precision + recall),

Table 5
Confusion matrix.

		Predicted Values	
		Class 1	Class 2
Actual Values	Class 1	True Positives	False Negatives
	Class 2	False Positives	True Negatives

3.3.2. Cross-Validation

Cross-Validation is a statistical method for evaluating and validating data mining algorithms by dividing the dataset into two parts: one used for training and the other used for testing. In cross-validation, the training and validation sets must cross-over in successive rounds in order to run and validate the model in each cluster. There are many cross-validation methods such as K-fold, leave out one, adversarial validation and shuffle split.

3.3.2.1. K-fold. In order to minimize the low performance associated with the random dataset splitting of the training and test data, researchers tend to use K-fold cross-validation. In k-fold, the entire dataset (D) is randomly divided into k subsets (D1, D2, …, Dk) of the same size. The classification model is trained and tested k times. Each time (t1, t2, …, tk), it is trained on all except one subset (Dt) and tested on the remaining single subset (Dt). The overall accuracy is calculated as the average of the k accuracy measures.

3.3.2.2. Shuffle split cross-validation. Shuffle Split is an iterative cross-validation method that randomly samples the entire dataset during each iteration to generate a training set and a test set. The dataset is first shuffled and then split into a pair of train and test sets. The test_size and train_size parameters control how large the test set and training set should be in each iteration.

3.3.3. Regularization

One of the major aspects of Machine Learning is avoiding

Fig. 2. Dataset simulation steps.

overfitting. In this case, the model will have low accuracy. This happens when the model struggles to learn noisy data that does not represent the true properties of the dataset. Learning such data makes the model at the risk of overfitting. Regularization penalizes some of the model parameters to zero. Hence, some variables will not play any role in the model.

There are many types of regularization techniques, such as L1 regularization, L2 regularization, Elastic Net and in the context of Deep Learning, we also have a dropout. Classification models use "C" as regularization parameter; other estimators use "alpha". The relation between them is given as **C = 1/Alpha**.

- In SVM, the regularization parameter used is C, which implements the L1-Regularization method. The default value of "C" is 1. However, if the dataset contains noisy observations, the parameter should be decreased to regularize more the estimation. A high value of "C" will result in low prediction rate "under-fitting", and a low value of C will result in high prediction rate "over-fitting".
- Neural Networks regularization parameter is alpha; this parameter combats overfitting by constraining the size of the weights. Increasing alpha may fix high accuracy (a sign of over-fitting) by encouraging smaller weights. Similarly, decreasing alpha may fix low accuracy (a sign of under-fitting) by encouraging larger weights.

4. Experiment

In this experiment, we followed the paper methodology to create our dataset. Before proceeding with data analysis, we will start by identifying the study area, the experiment environment and finally the fire zones used as tests.

4.1. Study area

The study area is composed of multiple zones located in the center of Canada (mostly in British Columbia and Quebec) (Fig. 3). The surface of this area is approximately 2 million hectares. Plots in yellow represent burned areas. Those plots differ in their size, burn period, date of burn and extent. We have chosen to apply the experiment in a big region of Canada's forests because it is known for its high rate of wildfires and also for the availability of fire information (start and end fire date, cause of fire and the surface of the burned area in hectares), these information were acquired from The Canadian Wild-land Fire Information System (CWFIS) which creates daily fire weather and fire behavior maps year-round and hot spot maps throughout the forest fire season [30].

4.2. Experiment environment

In order to analyze our dataset, we have chosen to use a big data platform called "Databricks".

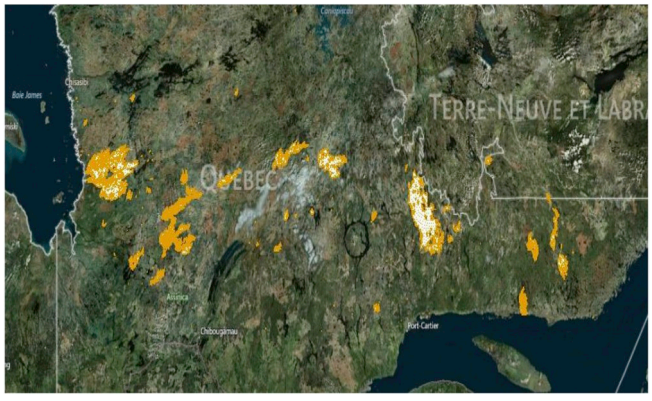

Fig. 3. Study area for wildfire occurrence prediction.

Table 6
Burned areas attribute table.

Fire No	Section	Cause	Start Date	End Date	Area (ha)
271	restricted	lightning	2014/08/26	2014/09/03	2,4
265	intensive	human	2014/08/26	2014/09/15	178
254	restricted	lightning	2014/08/07	2014/08/20	521,2
253	restricted	lightning	2014/08/08	2014/08/18	69,5

Table 7
Attributes of Canada's wildfire dataset.

	Attribute	Domain	Description
1	NDVI	-1–1	Normalized Difference Vegetation Index
2	LST	7500–65535	Land Surface Temperature
3	Thermal Anomalies	1–10	Detection of fire

4.2.1. What is “databricks”?

Databricks is a unified analytical Platform founded by Apache® Spark™; it can read from Amazon S3, MySQL, HDFS and Cassandra. Databricks is an open source project in the big data ecosystem. It allows individuals and organizations to build and deploy advanced analytic solutions through its virtual analytics platform [31].

4.3. Fire zones identification

In order to apply the proposed methodology to build our dataset using Remote Sensing data, we selected some fire zones that occurred in Canada's forest between 2013 and 2014, The burned areas were downloaded from the official website of the Natural Resources Canada [32], and it contains 386 fire zones; each one differs in its size, the nature of its vegetation, its position and its start and end date. Table 6 below presents an example of the burned areas attribute table.

After extracting Remote Sensing data (NDVI, LST, and Thermal Anomalies) for each fire zone, we added to our dataset “no_fire” instances where the fire did not happen. This way our dataset will be composed of 2 classes (**fire and no_fire**), and three integer-valued attributes. Table 7 shows a description of attributes used in the dataset and Table 8 Illustrates class distributions while Table 9 shows an example of the dataset used in the simulation.

In order to analyze the dataset in databricks, the first thing we need to do is to create a cluster which will serve as a database. Next, we will create a table and attach it to the created cluster. This table will be filled with our dataset. This operation is done by importing a CSV file containing all the data. The CSV file is composed of four columns, the three-first columns contain the parameters, and the last one represents the corresponding class.

Table 8
Class distribution.

	Class	Distribution
1	Fire	386 (48%)
2	No Fire	418 (52%)

Table 9
Wildfire prediction dataset.

NDVI	LST	Burned_Area	Class
0.51	14584.27	4.69	No_fire
0.68	14780.11	5	Fire
0.52	14655.83	5	No_fire
0.69	14658.42	3	No_fire
0.51	14333.37	4.97	Fire
0.45	14929.57	4.90	Fire
0.52	14738.22	5	Fire

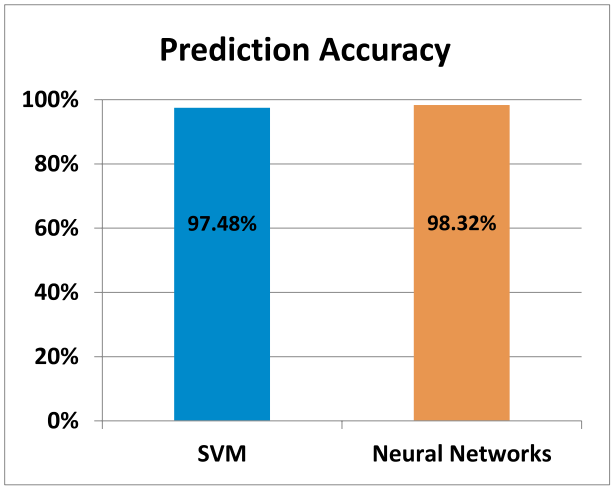

Fig. 4. Prediction accuracy of the algorithms.

5. Results and discussion

In this section, we present the results obtained from the experiment. In Figs. 4 and 5, we present the prediction rate of the class "fire" for the algorithms used, the performance of the algorithms are expressed using classification metrics, cross-validation, and regularization.

In this study, we used supervised learning [33] because we have a labeled dataset, composed of a set of existing data including the target values. The targets can have two possible outcomes (fire or no_fire). Supervised learning has proven very useful as a method for the discovery of new entries after sufficient training; they also have been successfully applied to spatial data in many areas including wildfires. In our simulation, we used two of the most popular data mining algorithms: Neural Networks and SVM [34]

5.1. Evaluation

As mentioned earlier, the data analysis is done using the "Databricks" platform. The training dataset consists of 804 instances with two distinct classes "Fire and No fire", In addition to instances of the class "no_fire", the dataset contains instances of the class "fire" which occurred in different areas in Canada. The performance of the classifiers is evaluated, and their results are analyzed using multiple validation techniques. According to the attributes, the dataset is divided into two parts: 70% (566) of the data are used for training, and 30% (238) are used for testing.

Table 10
Performance of the classifiers.

	Classifiers	
	SVM	Neural Networks
Correctly classified instances	116	117
Incorrectly classified instances	3	2
Accuracy (%)	**97.48**	**98.32**

5.1.1. Simulation accuracy

After running all three algorithms in "Databricks" platform using the steps described above, we obtained a high prediction rate of wildfire occurrence (**98.32% for Neural Networks and 97.48% for SVM**), An average of 117 instances out of total 119 instances of the class "fire" is found to be correctly classified using "NN", while SVM managed to predict 116 fire instances out of 119 accurately. Table 10 and Fig. 4 below show the performance of the classifiers.

As shown in Fig. 5, the dots in blue represent the incorrect prediction, while the red ones represent the correct prediction.

This result proves that the chosen parameters (**NDVI, LST, and Thermal anomalies**) can be used for wildfire occurrence prediction. However, prediction accuracy is not enough for evaluating the performance of the model. For this reason, we computed multiple performance classification metrics to validate the model. The metrics used in this simulation are calculated for the class "fire" (fire occurrence = yes). We used the **classification metrics** which include: "True Positive Rate", "True Negative Rate", "False Positive Rate", "False Negative Rate", F-Score, Recall, and precision.

5.1.2. Model validation

In this section, we are going to use three validation techniques (classification metrics, cross-validation, and regularization).

5.1.2.1. Classification metrics. In order to compare the actual classes with the predicted results, we generated a confusion matrix for each classifier using the python function **"confusion_matrix(y_true, y_pred)"**, with: (**y_true:** the actual values of the test set and **y_pred:** the predicted values from the test set). Table 11 below shows the confusion matrix for both classifiers (**NN and SVM**) using the test dataset composed of 238 instances divided equally between both classes (**fire:** 119 and **no_fire:** 119).

As observed in Table 11, the number of "no_fire" instances incorrectly predicted as "fire" instances noted as "false positives" (NN: 22, SVM: 35) are considerably higher than the number of "fire" instances incorrectly predicted as "no_fire" instances noted as "False Negatives" (NN: 2, SVM: 3). This can be explained by having some "no_fire" instances in the test set with values near some "fire" instances values in

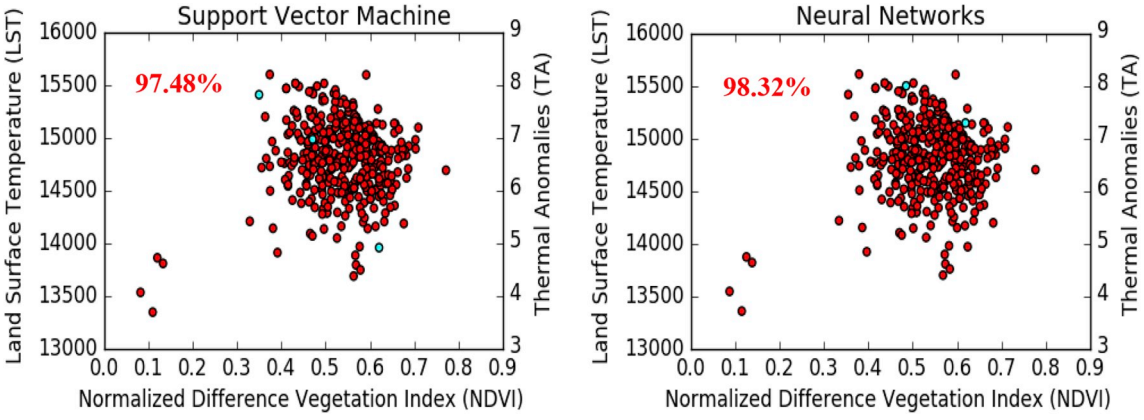

Fig. 5. Wildfire occurrence prediction accuracy for both classifiers for the class "fire.

Table 11
Confusion matrix for both classifiers (NN and SVM).

- **Neural Networks** correctly predicted 214 instances out of 238 instances (117 fire instances that are effectively fire and 97 no_fire instances that are actually no_fire), and 24 instances incorrectly predicted (2 instances of fire class predicted as no_fire and 22 instances of no_fire class predicted as fire)
- **SVM** correctly predicted 200 instances out of 238 instances (116 fire instances that are effectively fire and 84 no_fire instances that are actually no_fire), and 38 instances incorrectly predicted (3 instances of fire class predicted as no_fire and 35 instances of no_fire class predicted as fire)

		Predicted Values			
		Neural Networks		SVM	
		Fire	No_Fire	Fire	No_Fire
Actual Values	Fire	117	2	116	3
	No_Fire	22	97	35	84

Table 12
Accuracy measures for NN and SVM.

	TPR	FPR	Precision	Recall	F-Measure
Neural Networks	0.98	0.02	0.97	0.98	0.97
SVM	0.97	0.03	0.96	0.97	0.97

the training set, and thus the classifiers predict them as "fire" instances. This similarity can happen in some cases (e.g., Arid regions, Hot weather), in these cases the crops get dried and heated, and the soil temperature gets higher than usual (usually happens in the summer). This can affect NDVI, LST as well as TA values, and thus becoming similar to those of "fire" instances. To overcome this problem, additional parameters will be added in the dataset to enhance the prediction accuracy of the model for both classes “no_fire” and “fire.

Based on the confusion matrix, we computed classification metrics for both classifiers. As shown in Table 12 and Fig. 6 below, both algorithms gave good results. However, Neural Networks outperformed the SVM in terms of sensitivity, specificity, precision, and F-score. Performance is typically assessed by these metrics to describe the accuracy of the classification:

Both algorithms used (NN, SVM) gave good results in term of accuracy, but we can notice that Neural Networks outperformed SVM. The obtained results of the classification metrics prove the efficiency of the model used for predicting the occurrence of wildfires.

5.1.2.2. Cross-Validation. This section presents the results obtained by the cross-validation techniques: K-Fold and Shuffle split.

5.1.2.2.1. K-Fold. We divided our dataset into 5 clusters using K-Fold, each cluster gave different accuracy. Table 13 and Fig. 7 show the accuracy for each cluster as well as their average using k-fold cross-validation.

5.1.2.2.2. Shuffle split cross-validation. We shuffled our dataset and split it into five pairs of training, and test sets, Table 14 and Fig. 8 show the accuracy for each cluster as well as their average using shuffle split cross-validation.

The results obtained by both methods are almost identical to the ones obtained in the simulation.

5.1.2.3. Regularization. In SVM, increasing the regularization parameter “C” resulted in a fixed low prediction rate (91.60%), a sign of "under-fitting". Conversely, decreasing the value of “C” resulted in a fixed high prediction rate (100%), a sign of "over-fitting". Variable values of prediction rate [92.60%–97.48%] are considered as “Appropriate-Fitting”. Table 15 and Fig. 9 illustrate the fire prediction accuracy variation with regularization in SVM.

In Neural Networks, the regularization parameter is “alpha”. Unlike SVM, Increasing “alpha” fixed high prediction accuracy “Over-fitting”. While decreasing “alpha” fixed low prediction accuracy “Under-fitting”. Similarly to SVM, “Appropriate-Fitting” accuracies are between 91.43% and 98.32%. Table 16 and Fig. 10 present the accuracy variation with regularization in Neural Networks (see Fig. 11) (see Table 17).

5.2. Model comparison

After validating our model with classification metrics, regularization, and cross-validation. Now we move to another aspect of validation, which consists in comparing the results of our model with other wildfires prediction models. In this comparison, we have chosen three frameworks: 1) STIFF (spatiotemporal forecasting framework) [40], 2) ISTFF (integrated spatiotemporal forecasting framework) [41] and 3) ARIMA (auto-regressive integrated moving average) [42], all three frameworks were used in Canada's forest to predict fires.

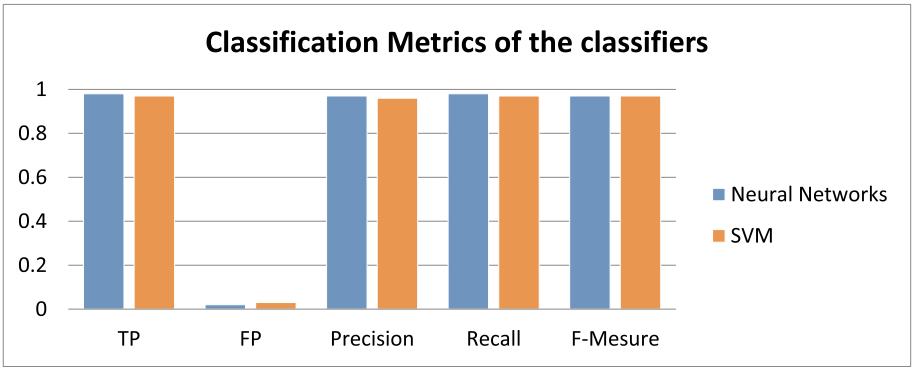

Fig. 6. Classification metrics of NN and SVM

Table 13
K-Fold Cross-Validation Accuracy Prediction for both classifiers.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Average
Neural Network (%)	97.52	99.54	99.27	99.32	98.81	**98.89**
SVM (%)	97.64	98.77	97.54	96.46	95.84	**97.22**

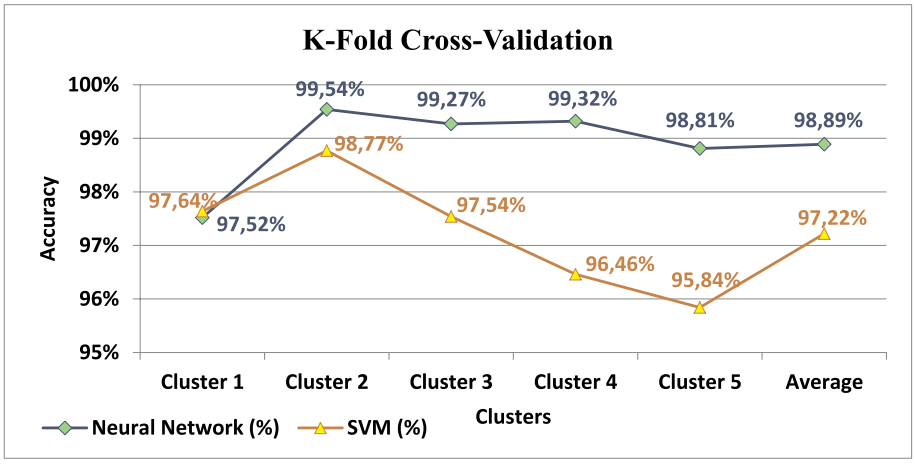

Fig. 7. K-fold cross-validation for neural network and SVM

Table 14
Shuffle Split Cross-Validation Accuracy Prediction for both classifiers.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Average
Neural Network (%)	98.53	99.61	98.24	99.4	98.22	**98.8**
SVM (%)	97.21	98.22	97.52	96.27	96.52	**97.15**

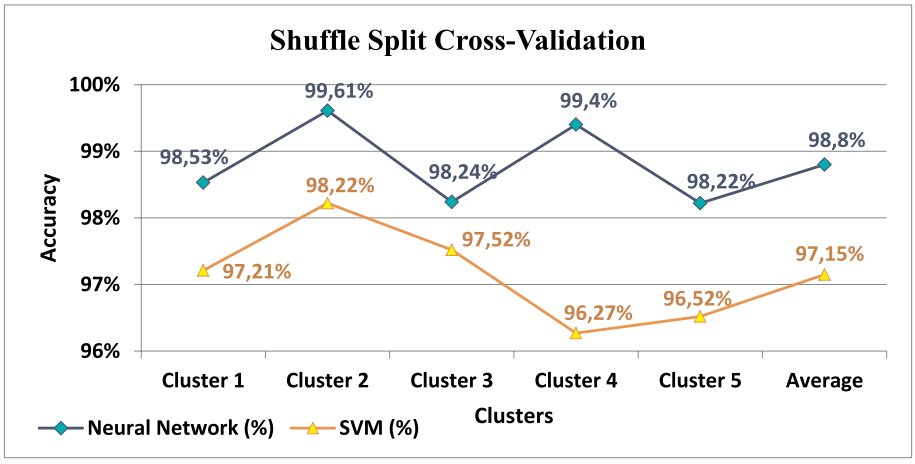

Fig. 8. Shuffle split cross-validation for neural network and SVM

Table 15
Regularization table for SVM.

(C)	0.01	0.1	0.8	1	2	3	4	5	6
SVM (%)	Over-Fitting		Appropriate-Fitting				Under-Fitting		
	100	100	97.48	95.79	94.12	92.60	91.60	91.60	91.60

5.2.1. TIFF: Spatio-temporal forecasting framework

TIFF is a spatiotemporal forecasting framework that combines both statistical analysis and data mining. The framework consists of several different functional components; the algorithm is composed of four steps as follows:

1) Determining the target location and its spatially separated siblings.
2) Building a time series model for each location in order to provide the necessary temporal forecasting capability at the location. Specifically, forecasting from the initial model recorded as f_T.
3) Constructing an artificial neural network based on the spatial correlation of all locations in order to capture the spatial influence over the target location. Once the neural network has been built, the network is fed by the forecasting from each time series model. The output of the network is considered as the spatially-influenced forecasting noted as f_S.
4) Combining both forecasting f_T and f_S with a statistical regression mechanism which is the final overall spatiotemporal forecasting [43].

5.2.2. ISTFF: integrated Spatio-temporal forecasting framework

ISTFF (integrated spatiotemporal forecasting framework) is an improved version of STIFF that aims to provide better forecasting accuracy. The new framework has three objectives:

✓ Constructing random time series models to catch the temporal characteristics of each spatially independent subcomponent,

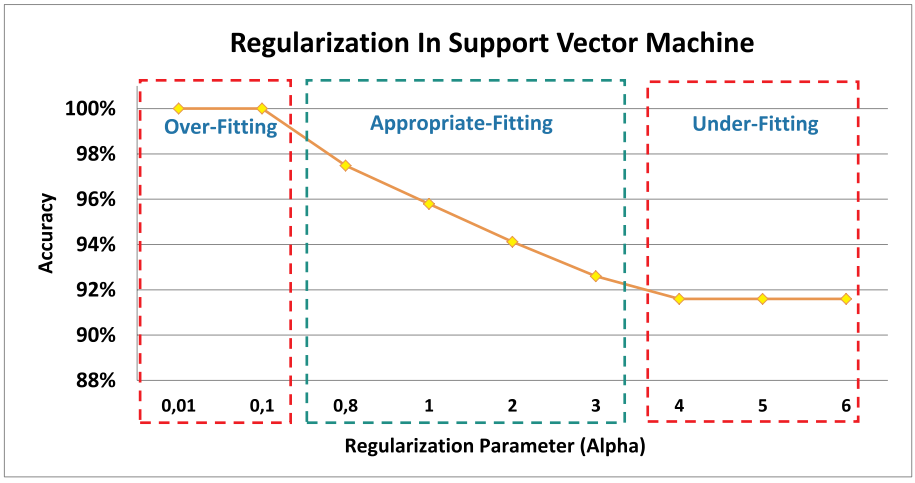

Fig. 9. Varying regularization in SVM

Table 16
Regularization table for Neural Networks.

(Alpha)	0.01	0.1	0.8	1	2	3	4	5	6
Neural Networks (%)	**Under-Fitting**		**Appropriate-Fitting**					**Over-Fitting**	
	89.07	89.07	91.43	92.59	93.28	95.79	98.32	100	100

Table 17
Accuracy evaluation of wildfire occurrence prediction in Canada with different models.

	Accuracy	RANK
ARIMA	83.5%	4
STIFF	65%	5
ISTFF	95%	3
NN	98.32%	1
SVM	97.48%	2

✓ Building a dynamic recurrent neural network (DRNN) to determine the hidden spatial correlation between the target subcomponent and other subcomponents,
✓ Combining the previous individual temporal and spatial forecasts, based on statistical regression, to generate the final forecasting result of the target subcomponent.

The algorithm of ISTFF is composed of 5 steps: 1) Defining the forecasting problem in terms of specification, 2) Time series analysis and temporal forecasting, 3) Spatial forecasting, 4) Overall spatio-temporal forecasting using linear regression, 5) Model validation and accuracy evaluation [43].

5.2.3. *ARIMA: auto-regressive integrated moving average*

The ARIMA modeling is an approach oriented by data that has the flexibility to fit an appropriate model adapted from the structure of the data. The stochastic nature of the time series can be approximately modeled using autocorrelation function, from which information can be discovered. And thus, forecasting future values of the series with some degree of accuracy.

The auto-regressive moving average (ARMA), or auto-regressive integrated moving average (ARIMA) models are often applied for time series forecasting. However, the application of the ARIMA model requires the time series to be stable; the algorithm of ARIMA assumes that the process remains stable at a constant mean level. If the series are unstable or have clear variability, the ARIMA model can be used.

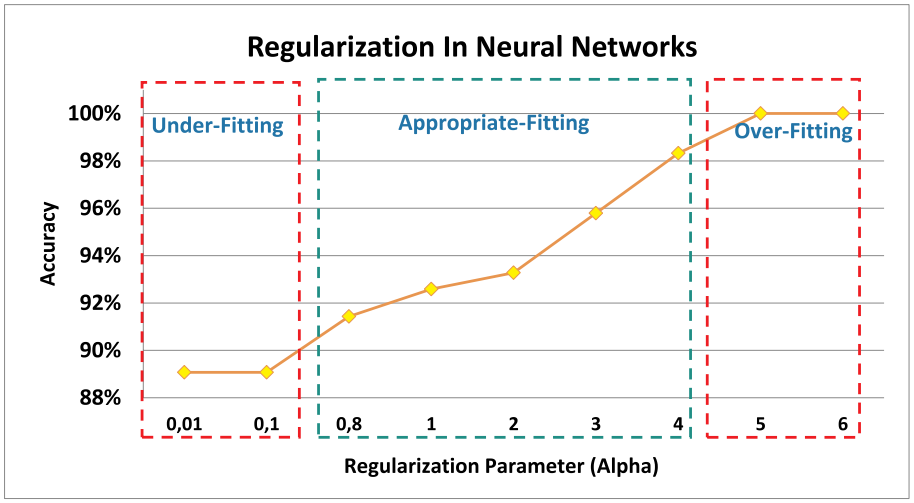

Fig. 10. Varying regularization in neural networks.

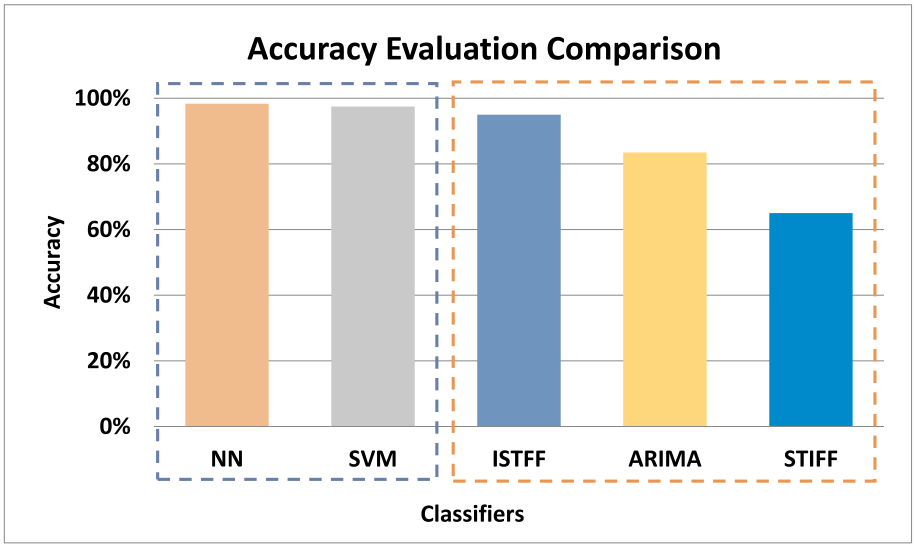

Fig. 11. Accuracy evaluation comparison between our model and other models.

Table 18
CFFDRS' wildfire occurrence prediction reports in the study area between 2013 and 2014Wildfires/Region.

Wildfires/Regions Date	British Colombia		Quebec	
	Actual	Predicted	Actual	Predicted
10/06/2013			15	10–15
11/06/2013			3	3–5
13/06/2013	1	2	6	3–5
15/06/2013			9	5–10
06/07/2013			21	12–20
26/08/2013	8	8		
03/09/2013	6	7		
01/07/2014			9	5–10
19/07/2014	5	5		
23/07/2014	4	3		
07/08/2014	21	18		
08/08/2014	11	8		
10/08/2014	8	8		
14/08/2014			3	1–3
Accuracy	**92,19%**		**98,48%**	

The ARIMA model consists of three major steps: model identification, parameter estimation, and application. Among these three steps, the identification step is the most important one and includes two phases: 1) performing appropriates differencing of the series to achieve stationary and normality, 2) identifying the order of the AR and MA parts of ARMA model [43].

5.2.4. Accuracy evaluation

The above models were applied in different regions of Canada to predict the occurrence of wildfires. The forecasting accuracy of each model was assessed and compared with the results obtained in this work.

According to the accuracy evaluation comparison between the algorithm used in this work (NN and SVM) and the other models (STIFF, ISTFF, and ARIMA), we can see that NN have the highest accuracy value **(98.32%)**. This comparison proves that the methodology used in this paper outperformed the other models.

5.2.5. CFFDRS: the Canadian Forest Fire Danger Rating System

Forest fire danger rating researches in Canada have been under development since 1925. The current system is **the Canadian Forest Fire Danger Rating System (CFFDRS)**, which was developed in 1968. It is considered the main source of fire intelligence for all forest fire management agencies in Canada. CFFDRS is composed of two modules [32]:

1. **The Canadian Forest Fire Weather Index (FWI) System**, which depends only on weather parameters (e.g., Fuel moisture codes, Fine Fuel Moisture Code (FFMC), Duff Moisture Code (DMC), and Drought Code (DC)). FWI provides a general measure of fire danger throughout forested and rural areas.
2. **The Canadian Forest Fire Behavior Prediction (FBP) System** helps forest managers evaluate the spread of fire in a particular forest type, the amount of fuel it might consume and, finally, the possible intensity of the fire. The FBP system relies on 14 primary data inputs in five general categories: fuels, weather, topography, foliar moisture content, and type and duration of prediction.

In order to compare the wildfire occurrence prediction accuracy of our model with the **Canadian Forest Fire Danger Rating System (CFFDRS)**, we collected CFFDRS' archived reports from **the Canadian Interagency Forest Fire Center** [44,45] relevant to our study area and period (2013–2014). The archived reports provide daily Hotspot reports of wildfires in multiple regions in Canada, these reports include diverse information (e.g., number of wildfires occurred, predicted, preparedness level). The predicted wildfires are estimated based on multiple satellites, (e.g., Advanced Very High-Resolution Radiometer (**AVHRR**), Moderate Resolution Imaging Spectroradiometer (**MODIS**), and Visible Infrared Imaging Radiometer Suite (**VIIRS**)). Table 18 below presents an example of the CFFDRS' wildfire occurrence prediction reports in the study area (**British Columbia and Quebec**). Each cell contains the number of wildfires occurred/predicted in both regions between 2013 and 2014.

According to the CFFDRS' archived reports, the accuracies of the wildfire occurrence prediction in the studied area regions are:

- "British Colombia": (92.19%), "Quebec": (98.48%).

The CFFDRS' overall accuracy of the wildfire occurrence prediction in both regions between 2013 and 2014 is approximately (**95.34%**), which is lower than the accuracy obtained by our model (**98.32%**). However, the performance of the CFFDRS varies according to multiple criteria, including the studied area, time period, and meteorological conditions among others. For example in other Canadian regions, such as "**Manitoba**" and "**Ontario**", the CFFDRS' accuracy of wildfire occurrence prediction in the same period is **98.72%** and **99.14%** respectively.

In our study, we took into account only the wildfires located exactly inside our area of study. In the CFFDRS some wildfires might be located

Table 19
General Applicability of the Model: steps, operations, challenges and solutions.

Steps	Principle	Challenges	Solutions
Fire zones identification	This operation aims to identify some fire zones in the study area, which consists in defining delimited areas where wildfires happened in the past, date of the fire and the exact area where the fire occurred, the format of this zones is generally "SHP: Shapefile ".	Information about real historical wildfires in specific regions is not always available.	Real fires zone is made available by some Wild-land Fire Information Systems, such as (CWFIS) "*the Canadian Wild-land Fire Information System*" [30].
Data Collection	To collect satellite data, we need to specify the data sources, which are generally chosen according to the availability of data for the study area. After selecting the data source, we then download the satellite images that contain the study area. The obtained format is either in HDF or GeoTIFF.	- The volume of satellite images. - Unavailability of satellite images.	- The use of powerful servers and high internet connection. - Checking many data sources (**e.g., USGS, ESA, NOAA, Digital Globe**), if not found use data interpolation.
Data Conversion	This operation is done when the collected satellite images are not in GeoTIFF format, so it is necessary to use specific libraries (**GDAL, HEG**) to convert the satellite images into a readable format (raster).	The satellite image might contain many bands, and when converting it, we might neglect some important bands.	Before converting a satellite image, we should know all its bands and its corresponding meaning.
Data Cleaning	As already mentioned, satellite images may contain many errors and should be cleaned before being processed using data cleaning techniques (e.g., radiometric calibration, geometric correction, and atmospheric correction).	The satellite images might be extremely distorted, and thus geometric corrections are complex and need more processing time.	To correct images distortion, divide the geometric correction into three phases: first-order, second-order and third-order transformation.
Data Clipping	Data Clipping or sub-setting is the operation of cutting satellite images to match the shape of the study area using the GDAL library.	The main problem is the image volume; it is difficult to process large images.	This problem is resolved by using machines with high Memory capacity.
Data Interpolation	As satellite images may have different time spans (e.g., hourly, daily, weekly), data interpolation is a crucial step to normalize the satellite images in one timespan (daily). Data interpolation methods are divided into two major categories (deterministic and probabilistic)	The interpolated data are not always realistic and can be stochastic in some cases.	Choose the right data interpolation method; this choice is made based on the type of Satellite data.
Data Extrapolation	Predicting upcoming wildfires requires processing satellite images of a future date, so we need to extrapolate the raster data we have to get raster data of future date using data Extrapolation techniques (e.g., Rule-Based Forecasting, trend extrapolation methods, auto-regression)	The extrapolated data might not be accurate and thus causes unreliable predictions.	In order to obtain high accuracy, it is recommended to use domain knowledge as well as reliable historical data before any Data Extrapolation.
Data Extraction	The final step consists in extracting data from the satellite images and prepares them to be processed. Retrieving data from the image's pixels is done using the GDAL library.	Reading data contained in the satellite images pixels is a challenging operation due to: raster data format (e.g., multiple bands, high volume, complex format)	It is recommended to specify only the band(s) that contains data and use medium size parcels to avoid memory problems or use servers with high memory.

in areas nearby the studied region while still belonging to the same forest. Thus, the number of fire events in the CFFDRS (130) is slightly higher than the ones used in our study (119).

Despite the high prediction accuracy of the CFFDRS, the proposed model brings a significant added value in predicting the occurrence of wildfires as described in the points below:

- The accuracy of the wildfire occurrence prediction has increased from 95% to 98%, which is significant because 3% might represent thousands of hectares and cost millions of dollars and save hundreds of lives.
- The CFFDRS is considered as a black box used to predict the occurrence of wildfires without further detail about the techniques nor the methodologies used, while in our model we describe in detail all the steps of the methodology including the parameters, tools, methods, as well as the techniques used to pre-process and analyze the satellite data to predict the occurrence of wildfires.
- The CFFDRS depends on both remote sensing data and weather parameters, which restrict its use in zones equipped with weather stations, while our model uses only remote sensing data, which can be acquired from multiples sources. Thus, expanding the range of its use.
- The CFFDRS efficiency is demonstrated in limited zones in the world, while our model might be used in any region in the world, as long as the satellite images are available.
- In this methodology, we provided a guide for the general applicability of the model, which covers all its steps and highlights the challenges and the proposed solutions.
- The CFFDRS predicts the occurrence of wildfires in wide regions, while our system predicts their occurrence in accurate restricted zones.

All the above-mentioned arguments are proofs of the added value of the proposed model in predicting the occurrence of wildfires compared to the CFFDRS.

5.3. General applicability of the model

In order to apply the proposed model to predict wildfires in any region of the world, we need to follow the methodology described in this paper, which consists in building a dataset based on Remote Sensing data. Table 19 below describes the methodology's steps alongside with the challenges and possible solutions in each step.

6. Conclusion & future works

In this paper, we dealt with a very serious problem that threatens our lives as well as our ecosystem. Each year, thousands of hectares of forest around the world are destroyed by fire. The same amount is lost to logging and agriculture combined. These fires not only damage the

structure and composition of forests, but they also open up forests to invasive species, threaten biological diversity, alter water cycles and soil fertility, and destroy the livelihoods of the people who live in and around the forests. Hence, in order to reduce the damages caused by this disaster, we implemented a solution that predicts wildfires by making use of a very rich source of big data which is “Remote Sensing”, this latter offers different data through multiple satellite images. Those data need to be extracted and used for the right purpose.

In this paper, we implemented a methodology to build a dataset based on three parameters related to the state of the crops: NDVI, LST, and Thermal Anomalies. After identifying the studied area, we collected the corresponding data from Terra's instrument MODIS, preprocessed them and finally saved them in a dataset; which was analyzed using two of the most known supervised data mining algorithms: Neural Networks and SVM. The simulation was run using the big data platform “Databricks” which implements the spark framework as an engine for big data processing. The model gave good results for both algorithms (**SVM 97.48%, NN 98.32%);** the accuracy was assessed using classification metrics, cross-validation, and regularization. A comparison with some wildfire prediction models also proved the performance of our model. All these results confirm the efficiency of the model in predicting the occurrence of wildfires.

Future works will mainly consist of strengthening the model by including weather data. Weather plays a major role in the occurrence, growth, spread and the Extinction of wildfires. It can impact on the strength and movement of fire, and thus burn more land, which makes its extinction even more difficult. There are three weather parameters that can affect wildfires: Air Temperature, Wind and Soil Moisture.

- **Air Temperature** influences the occurrence of wildfires, by heating trees and crops on the ground, which makes them sensitive toward catching fire.
- **The Wind** has the most prominent and strongest impact on wildfires behaviors. Wind speed and direction are unpredictable. Besides, winds supply the fire with additional oxygen, which pushes the fire to move faster across the land.
- **Soil Moisture** is directly affected by precipitation and air humidity. When the soil moisture is low, the risk of wildfires is high. Conversely, high soil moisture lowers the chances of a wildfire igniting,

Acknowledgment

The MODIS data products were retrieved from the online Data Pool, courtesy of the NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, https://lpdaac.usgs.gov.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.firesaf.2019.01.006.

References

[1] J. L. Coen and C. C. Douglas, “Computational Modeling of Large Wildfires : Status and Challenges”.
[2] F. Tedim, V. Leone, M. Amraoui, C. Bouillon, M. Coughlan, G. Delogu, P. Fernandes, C. Ferreira, S. McCaffrey, T. McGee, J. Parente, D. Paton, M. Pereira, L. Ribeiro, D. Viegas, G. Xanthopoulos, Defining extreme wildfire events: difficulties, challenges, and impacts, Fire 1 (1) (2018) 9.
[3] R. Yaakob, N. Mustapha, A. Ainudin, I. Sukaesih Sitanggang, Modeling wildfires risk using spatial decision tree, Data Min. Optim. (2011) 28–29 no. June.
[4] Y.O. Sayad, H. Mousannif, M. Le Page, Crop management using big data, 2015 Int. Conf. Cloud Technol. Appl, 2015, pp. 1–6.
[5] M. Chi, A. Plaza, J.A. Benediktsson, Z. Sun, J. Shen, Y. Zhu, Big data for remote sensing: challenges and opportunities, Proc. IEEE 104 (11) (2016) 2207–2219.
[6] H. Ramapriyan, J. Brennan, J. Walter, J. Behnke, Managing big Data: NASA tackles complex Data challenges, Earth Imaging J. (2013) [Online]. Available: http://eijournal.com/print/articles/managing-bigData.
[7] M. Chi, J. Shen, Z. Sun, F. Chen, J.A. Benediktsson, Oil spill detection based on web crawling images, Proc. IEEE Int. Geosci. Remote Sensing Symp. Que'bec, QC, Canada, Jul. 2014, pp. 1–4 [Online]. Available: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3DData-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf.
[9] N.A. Sundar, P.P. Latha, M.R. Chandra, Performance analysis of classification data mining techniques over heart disease data base, Int. J. Eng. Sci. Adv. Technol. 2 (3) (2012) 470–478.
[10] X. Wu, et al., Top 10 Algorithms in Data Mining 14 (1) (2008).
[11] C. Vega-Garcia, B. Lee, P. Woodard, S. Titus, Applying neural network technology to human-caused wildfire occurrence prediction, AI Appl. 10 (3) (1996) 9–18.
[12] B. Arrue, A. Ollero, J. Matinez de Dios, An intelligent system for false alarm reduction in infrared forest-fire detection, IEEE Intell. Syst. 15 (3) (2000) 64–73.
[13] W. Hsu, M. Lee, J. Zhang, Image mining: trends and developments, J. Intell. Inf. Syst. 19 (1) (2002) 7–23.
[14] P. Cortez, A. Morais, A data mining approach to predict wildfires using meteorological data, Proc. 13th Port. Conf. Artif. Intell, 2007, pp. 512–523.
[15] R. Mikut, M. Reischl, Data mining tools, Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 1 (5) (2011) 431–443.
[16] X. Wu, X. Zhu, G.-Q. Wu, W. Ding, Data mining with big data, Knowl. Data Eng. IEEE Trans. 26 (1) (2014) 97–107.
[17] D. Stojanova, P. Panov, A. Kobler, S. Džeroski, K. Taškova, Learning to predict wildfires, Knowl. Creat. Diffus. Util. 9 (14) (2006) 255–258.
[18] P. Cortez, A. Morais, A data mining approach to predict wildfires using meteorological data, Proc. 13th Port. Conf. Artif. Intell, 2007, pp. 512–523.
[19] C.E. Van Wagner, Modelling logic and the Canadian forest fire prediction system, For. Chron. 74 (1) (1998) 50–52.
[20] MODIS data products, Courtesy of the NASA Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, https://lpdaac.usgs.gov.
[21] H. Mousannif, H. Sabah, Y. Douiji, Y. Oulad Sayad, Big Data projects: just jump right in!, Int. J. Pervasive Comput. Commun. 12 (2) (2016) 260–288.
[22] B.C. Gao, NDWI - a normalized difference water index for Remote Sensing of vegetation liquid water from space, Remote Sens. Environ. 58 (3) (1996) 257–266.
[23] Moderate resolution imaging spectroradiometer (MODIS), http://modis.gsfc.nasa.gov/.
[24] Z. Wan, Y. Zhang, Q. Zhang, Z. liang Li, Validation of the land-surface temperature products retrieved from terra moderate resolution imaging spectroradiometer Data, Remote Sens. Environ. 83 (1–2) (2002) 163–180.
[25] C.O. Justice, et al., The MODIS fire products, Remote Sens. Environ. 83 (1–2) (2002) 244–262.
[26] Y.O. Sayad, H. Mousannif, M. Le Page, “Crop management using big data,” 2015, Int. Conf. Cloud Technol. Appl. (2015) 1–6.
[27] Jonathan Beaudoin, J.E. Clarke, Hughes, Edward J. Van den Ameele, James V. Gardner, Geometric and radiometric correction of multibeam backscatter derived from reson 8101 systems, Canadian Hydrographic Conference, vol. 242, 2002 http://scholars.unh.edu/ccom/242.
[28] R. Allen, L.S. Pereira, D. Raes, M. Smith, Crop Evapotranspiration: Guidelines for Computing Crop Requirements, Irrig. Drain. Pap. No. 56, FAO, no. 56 (1998), p. 300.
[29] R. Sluiter, Interpolation methods for climate Data: literature review, KNMI, R&D Inf. Obs. Technology, 2009, pp. 1–28.
[30] The Canadian Wildland Fire Information System (CWFIS) http://cwfis.cfs.nrcan.gc.ca/.
[31] Databricks official website, https://docs.Databricks.com/.
[32] Natural Resources Canada, http://www.nrcan.gc.ca/home.
[33] P. Ozer, Data Mining Algorithms for Classification, (2008), p. 26 no. January.
[34] X. Wu, V. Kumar, Q.J. Ross, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.H. Zhou, M. Steinbach, D.J. Hand, D. Steinberg, Top 10 Algorithms in Data Mining, 14 (2008), pp. 1–37.
[35] Neural Networks, Scikit Learn, http://scikit-learn.org/sTable/modules/neural_networks_supervised.html.
[36] MLPClassifier, http://scikit-learn.org/sTable/modules/generated/sklearn.neural_network.MLPClassifier.html.
[37] SVM, http://scikit-learn.org/sTable/modules/svm.html#svm.
[38] SVC Classifier, http://scikit-learn.org/sTable/modules/generated/sklearn.svm.SVC.html.
[39] W.M.P. Van Der Aalst, V. Rubin, H.M.W. Verbeek, B.F. Van Dongen, E. Kindler, C.W. Günther, Process mining: a two-step approach to balance between underfitting and overfitting, Software Syst. Model 9 (1) (2010) 87–111.
[40] Z. Li, M.H. Dunham, Y. Xiao, STIFF: a forecasting framework for SpatioTemporal data, in: O.R. Zaïane, S.J. Simoff, C. Djeraba (Eds.), Mining Multimedia and Complex Data. PAKDD 2002. Lecture Notes in Computer Science, vol. 2797, Springer, Berlin, Heidelberg, 2003.
[41] T. Cheng, J. Wang, Integrated spatiotemporal Data Mining for forest fire prediction, Trans. GIS 12 (5) (2008) 591–611.
[42] T.H.E. Use, O.F. Arima, M. For, A the Centre for quality, ngee ann polytechnic, Singapore. 535, clementi road, Singapore 599489. b department of industrial and systems engineering, national university of Singapore. 10, kent ridge crescent, Singapore 119260, Science 35 (98) (1998) 213–216 (80-.).
[43] W. Wang, W. Wang, K. Chau, D. Xu, Improving Forecasting Accuracy of Annual Runoff Time Series Using ARIMA Based on EEMD Decomposition, (June 2015).
[44] Archived reports of Canadian Wildland Fire Information System: http://cwfis.cfs.nrcan.gc.ca/report/archives.
[45] Archived Situation Reports: http://www.ciffc.ca/.

Oulad Sayad Younes is a Ph.D. candidate in Computer sciences at the University of CADI AYYAD, Marrakesh Morocco with research interests primarily in Big Data, Remote Sensing and Machine Learning. His work focuses on developing a Remote Sensing system using Big Data and Machine Learning. Younes graduated from the same University at the faculty of sciences Semlalia with a master degree in engineering information systems in 2013.

Hajar Mousannif is an associate professor within the department of computer science at the Faculty of Sciences Semlalia (Cadi Ayyad University, Morocco). She holds a PhD degree in Computer Sciences on her work on Wireless Sensor Networks and Vehicular Networks. She received an engineering degree in Telecommunications in 2005. Her primary research interests include Big Data, IoT, Human-Computer Interaction, and next-generation internet technologies. In addition to her academic experience, she chaired the Program Committee of many international conferences. Hajar Mousannif holds a patent on her work on Affective Computing and was selected among 5 best female researchers in North Africa. She received many international awards such as L'Oréal-UNESCO Award and the Emerald Litterati Prize for Excellence.

A native of Rabat city (Morocco), **Hassan Al Moatassime** graduated from the Mohamed V University, Rabat (Morocco) in 1988 with a B.S. in applied mathematics. In 1989, he received the M.S. degree in numerical analysis from the Paris Sud University, France. He got a Ph.D. degree in numerical analysis from the Paris Sud University in 2003. After a year, he joined the Faculty of Sciences and Techniques Marrakech (Morocco) as an Assistant Professor. Currently, he is a Professor within the Department of Mathematics and Computer Sciences. His primary research interests are numerical analysis, computational fluid dynamic, network systems, and numerical methods for wave equations. He was an author and co-author of many peer-reviewed scientific papers in the field of numerical analysis.