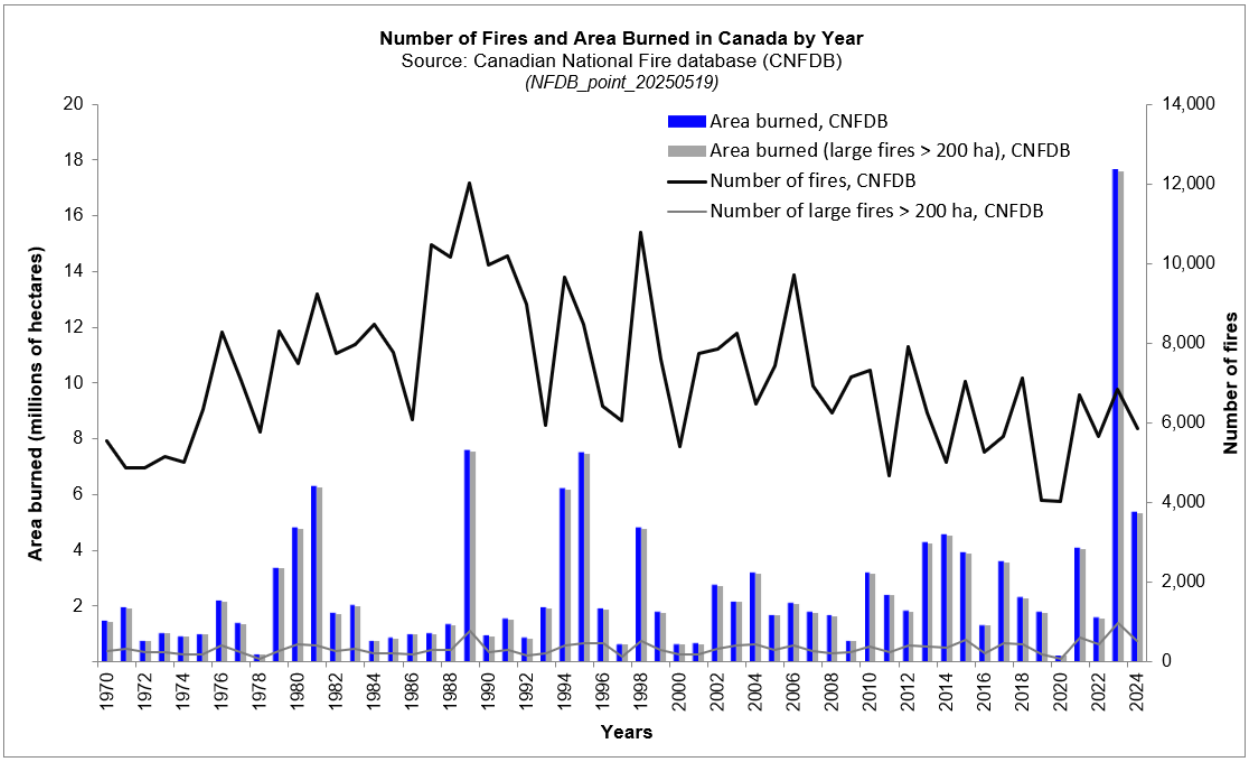
Question:

Can AI and Machine Learning (ML) accurately predict wildfires?

Introduction

Did you know that the majority of fires in Alberta are human caused?[1] Many are accidental, but due to Global Warming, with increasing dry seasons and droughts, fires grow more common.[1] Each year, fires threaten forests, wildlife, communities and air quality, causing economic and ecological damage. WIth quickly changing weather conditions, predicting where and how severe a wildfire might be is challenging. The National Forestry Database shows that over 8000 fires occur each year, and burn an average of over 2.1 million hectares.[2]



How we prevent and reduce the burden of forest fires has evolved with scientific discovery and better data collection by the federal government. More recent scientific findings have shown that wildfire suppression and performing controlled burns in fire-prone habitats are important measures to reduce overall wildfire intensity and frequency.[3] Wildfire patterns have become more unpredictable year to year, making it more challenging to predict where to allocate resources for fire prevention, management, suppression, and control.

[1] https://www.canadawildfire.org/wildfirefacts
[2] https://cwfis.cfs.nrcan.gc.ca/ha/nfdb
[3] https://natural-resources.canada.ca/forest-forestry/wildland-fires/fire-management

Could Artificial Intelligence (AI) — computer systems designed to perform tasks that normally require human intelligence, such as pattern recognition, learning, and decision-making — be used to help us better predict where to focus our efforts to help control wildfires?

This project explores how artificial intelligence (AI) could be used to help predict wildfires and identify higher risk areas in Alberta. By analyzing past fire records, an AI model may be able to recognize patterns that humans may miss. These patterns can then be used to estimate where fires are more likely to occur.

The goal of this project is to demonstrate how AI can support early warning systems and improve wildfire preparedness. If accurate predictions are made, emergency responders could use this information to plan evacuations, allocate resources more effectively, and reduce the overall impact of wildfires on people and the environment.

Real-World Applications

This project has important real-world applications for wildfire prevention and emergency management in Alberta and world wide. An AI-based fire prediction system could help firefighters and emergency responders prepare for fires before they start or spread. By identifying high-risk areas, authorities could pre-position firefighting equipment, issue early warnings, and plan evacuations more efficiently.

In addition, government agencies could use this technology to improve land-management strategies, such as controlled burns or vegetation clearing in high-risk zones. Accurate fire-risk predictions could also reduce economic losses, protect wildlife habitats, and improve public safety. Overall, this project demonstrates how artificial intelligence can be used as a powerful tool to support environmental protection and disaster preparedness.

Background Information:

Machine Learning (ML) is the study of computer algorithms that can improve automatically through experience.[4] This approach is heavily data-centric, dependent on the quality and quantity of available data relevant to the task.[5] ML can be defined as a set of methods that detect patterns in data, use the uncovered patterns to predict further data or other outcomes of interest.[6]

ML can generally be identified as either belonging to supervised learning, unsupervised learning or agent-based learning.

In Supervised ML, all problems can be seen as a function passing through specific parameters, often called a “model”, which maps known inputs (i.e, predictor variables) to known outputs (i.e., target variables). The goal of supervised learning is to use an algorithm to reach a specific outcome given a specific set of parameters.[5]

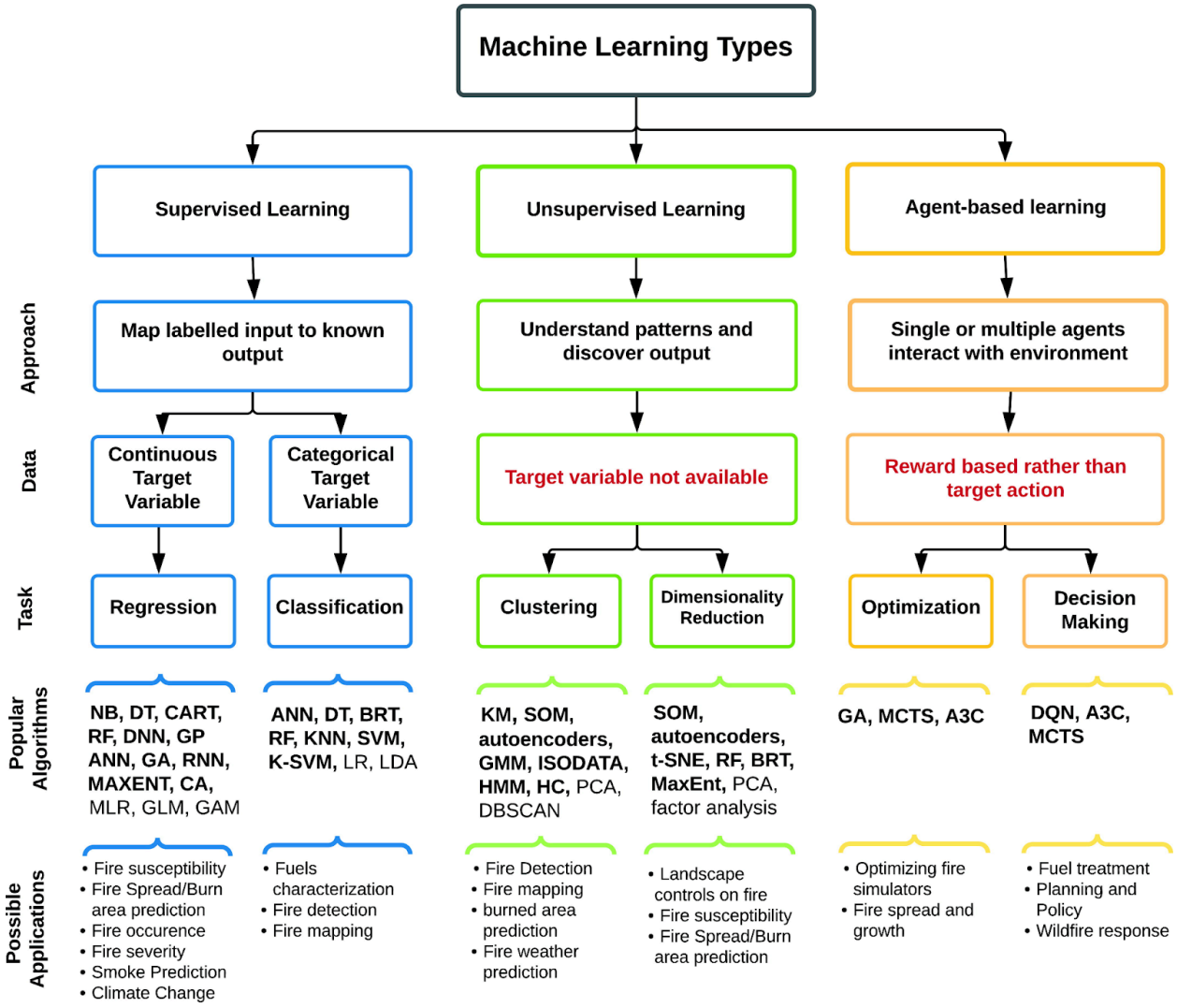
---

[4] Mitchell, T.M. 1997. Machine learning. McGraw-Hill.
[5] https://cdnsciencepub.com/doi/full/10.1139/er-2020-0019#core-refg246-1
[6] Murphy, K. 2012. Machine learning: a probabilistic perspective.

Unsupervised ML is when the target variables are not available. These problems are much more difficult to solve. Algorithms are used to discover hidden patterns or groupings without the need for human intervention.[7]

Agent-based ML refers to a model between supervised and unsupervised learning by which learning happens by simulating behaviours and interactions of autonomous agents. Given a set of incomplete base parameters, the model will autonomously go through trial and error to reach an outcome.[5]

Below is a diagram showing how different ML types can be used for possible application in wildfire science and management.[5]



Method:

I performed a Google Scholar search using keywords, “machine learning” and “wildfire prediction”.

[7] https://www.ibm.com/think/topics/unsupervised-learning

I found a review of machine learning applications in wildfire science and management which identified 300 relevant publications from 1996 to 2019. I found many articles on wildfire spread prediction, but few on predictive modeling. I focused on published data after 2019 to present.

Research:

Fire occurrence prediction (FOP) models dates back 100 years.[8] FOP models usually use regression methods to relate response variables, such as fire reports or hotspots, to weather, lightning, and other covariates for a geographic unit. The ML method most commonly used in studies predicting fire occurrence were Artificial Neural Networks (ANN).[5] The basic unit of an ANN is a neuron, also called a perception or logistic unit. A neuron involves a set of inputs which are combined linearly through multiplication with weights associated with the input resulting in an output signal.[5]

In 1996, an ANN for human-caused wildfire prediction in Whitecourt Provincial Forest of Alberta, Canada, accurately predicted 85% of no-fire observations and 78% of fire observations.[9] A 314 fire and no-fire data set for the period of 1986 to 1990 was used for training. The model was then tested using data from the 1991 to 1992 fire seasons, which had 58 observed fires. The input variables were the Canadian Fire Weather Index for the day, area in km^2 of the geographic zone, and district.[9] This study showed that human-caused fires are predictable when enough environmental and human-activity data are included. It is often cited because it was one of the first applications of AI in wildfire prediction. It laid the groundwork for modern wildlife ML research.

A 2017 study in Yunnan Province, China found that a cost-sensitive Random Forest (RF) analysis outperformed ANN models for predicting wildfire ignition susceptibility. This study performed wildfire susceptibility assessment using multiple methods, including logistic regression, probit regression, an ANN and a RF algorithm. Sample ratio of ignition and nonignition data was investigated from 2002 to 2010. Nonignition data included meteorological, landform, and vegetation variables. The models used nonignition data because ignition events were comparatively rare. The results showed a cost-sensitive RF had the highest overall accuracy (88.47%) for all samples, and 94.23% accuracy for ignition prediction. The ANN had an accuracy of 83.78% for predicting wildfire ignitions, but only 78.44% for predicting nonignitions. The predication accuracy, variable importance, and spatial pattern of each model result were evaluated and compared for class percentage of true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

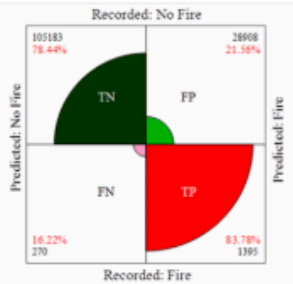
True positive: RF-cost sensitive (94.23%) > RF (84.26%) > ANN (83.78%)
True negative: RF-cost sensitive (88.4%) > RF (88.35%) > ANN (78.44%)
False positive: RF-cost sensitive (11.6%) < RF (11.65%) < ANN (21.56%)

---

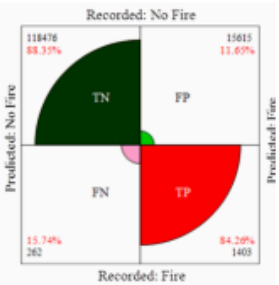
[8] Nadeem K., Taylor S.W., Woolford D.G., and Dean C.B. 2020. Mesoscale spatiotemporal predictive models of daily human- and lightning-caused wildland fire occurrence in British Columbia. Int. J. Wildl. Fire, 29(1): 11–27.
[9] Vega-Garcia C., Lee B.S., Woodard P.M., and Titus S.J. 1996. Applying neural network technology to human-caused wildfire occurrence prediction.

False negative: RF-cost sensitive (5.77%) < RF (15.74%) < ANN (16.22%)

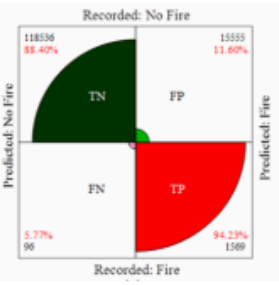
ANN



RF-original



RF-cost sensitive



Class Prediction Accuracy of Each Model:[15]

	Ignition (%)	Non-ignition (%)	Total accuracy (%)
ANN	83.78	78.44	78.51
RF-original	84.26	88.35	88.30
RF-cost sensitive	94.23	88.40	88.47

The factors that influence Yunnan wildfire occurrence the most were forest coverage ratio, month / season (temporal patterns), surface roughness (obtained from the Climate Forest System Reanalysis - CFSR), humidity measures (minimum of the 6 h humidity over 10 days) , temperature measures (maximum of the 6 h average temperatures over 10 days).[10]

A RF is a model composed of many individually trained decision tress (DT). A DT is a set of if-then-else rules with many branches joined by decision nodes and terminated by leaf nodes. The decision node is where the tree splits into different branches, with each branch corresponding to the decision being taken by the algorithm, and the leaf nodes represents the output.[5] Cost-sensitive RF is an RF adjusted to handle data imbalance (few ignitions vs many non-ignitions).[10]

Maximum Entropy Models (MAXENT) have also been used for fire occurrence prediction. MAXENT is a ML method used to predict where events are likely to occur based on known occurrences and environmental conditions. MAXENT was originally developed for species distribution modeling.[11] A case study of Canton Ticino, Switzerland used this ecological niche modeling to define what atmospheric conditions are most common when fires occur (“fire days”).

They compared this approach to traditional logistic regression models and also tested different sets of input variables (e.g., temperature / humidity / wind, fire weather indices or a combination). The daily meteorological variables used are: Air Temperature in Celsius degree (T), Air Humidity in percentage value (H), Wind velocity in m/s (U), Precipitation in mm (P), coverage of sky in ratio between 0 and 1

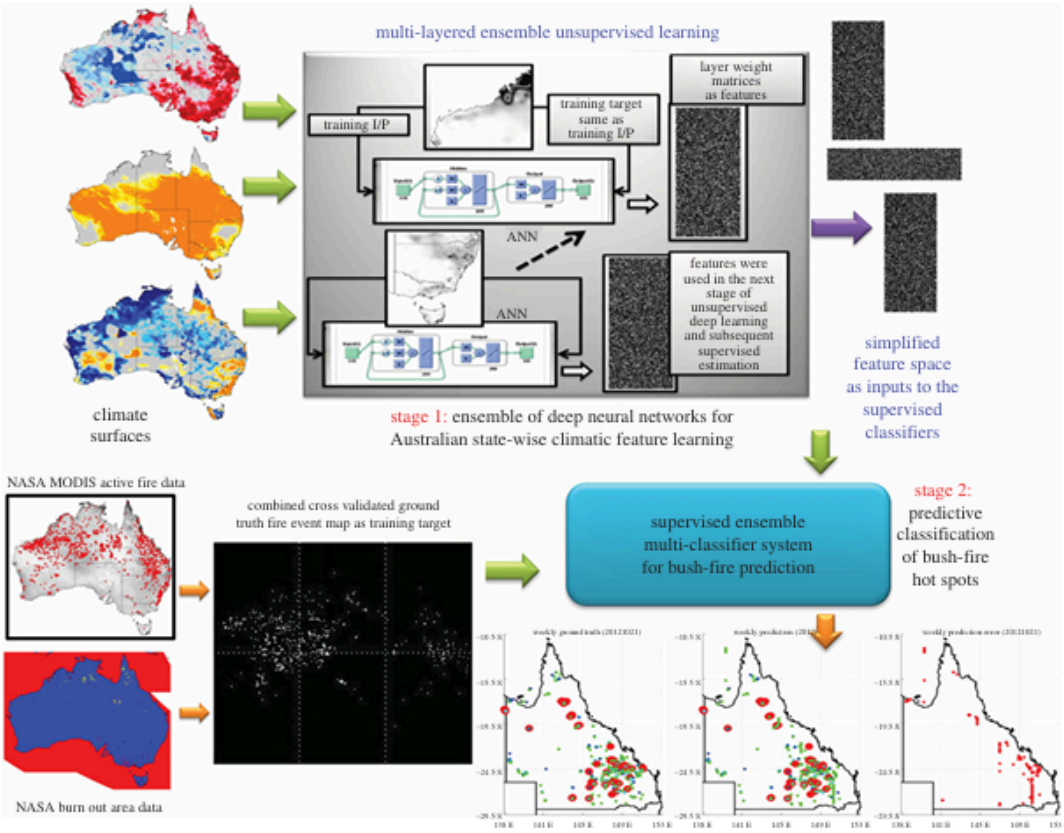
[10] https://link.springer.com/content/pdf/10.1007/s13753-017-0129-6.pdf

[11] De Angelis A., Ricotta C., Conedera M., and Pezzatti G.B. 2015. Modelling the meteorological forest fire niche in heterogeneous pyrologic conditions. PLoS ONE, 10(2): e0116875.

(CloudCover) and coverage of snow expressed as snow presence or absence (SnowCover). Model input variables: total rainfall over the last seven days (WeekRain), the days since the last rainfall (DaysSinceRain), the sum of the last rainfall (conse cutive days with rain) (LastRainSum), the dew point temperature (Tdew), that is the temperature to which air needs to be cooled to make air water vapour saturated [40] and the vapour pressure deficit (VPD) that is the difference between the saturation vapour pressure and the actual vapour pressure at a particular temperature.

Using combinations of variables generally improved model performance. MAXENT niche modeling showed slightly better predictive power, but results varied across different fire types (e.g., winter, summer natural, summer anthropogenic - influenced by human activity). Their results support the idea of flexible statistical and ML methods can enhance fire danger forecasting.[11]

An Australian study investigated trends in bush-fire frequency and set out to develop a predictive model linking climatic conditions to bush-fire occurrence. They developed a two-stage ensemble ML model using unsupervised deep learning (a multi-layered neural network to learn climatic patterns without prior labels) and supervised ensemble classification. In the first unsupervised deep learning (DL) phase, multi-layered deep neural networks [Deep Belief Neural Networks with Conventional Supervised Ensemble ML (DBNet)] were used to learn about the given data and generate simple features (from environmental and climatic surfaces) without any prior information or training targets. In the second supervised ensemble classification stage, features extracted from the first stage were fed as training inputs to 10 ML classifiers to establish the best classifier for bush fire hotspot estimation. Multiple supervised classifiers were used to learn the extracted features against the ground truth bush-fire hot spot map.[5]

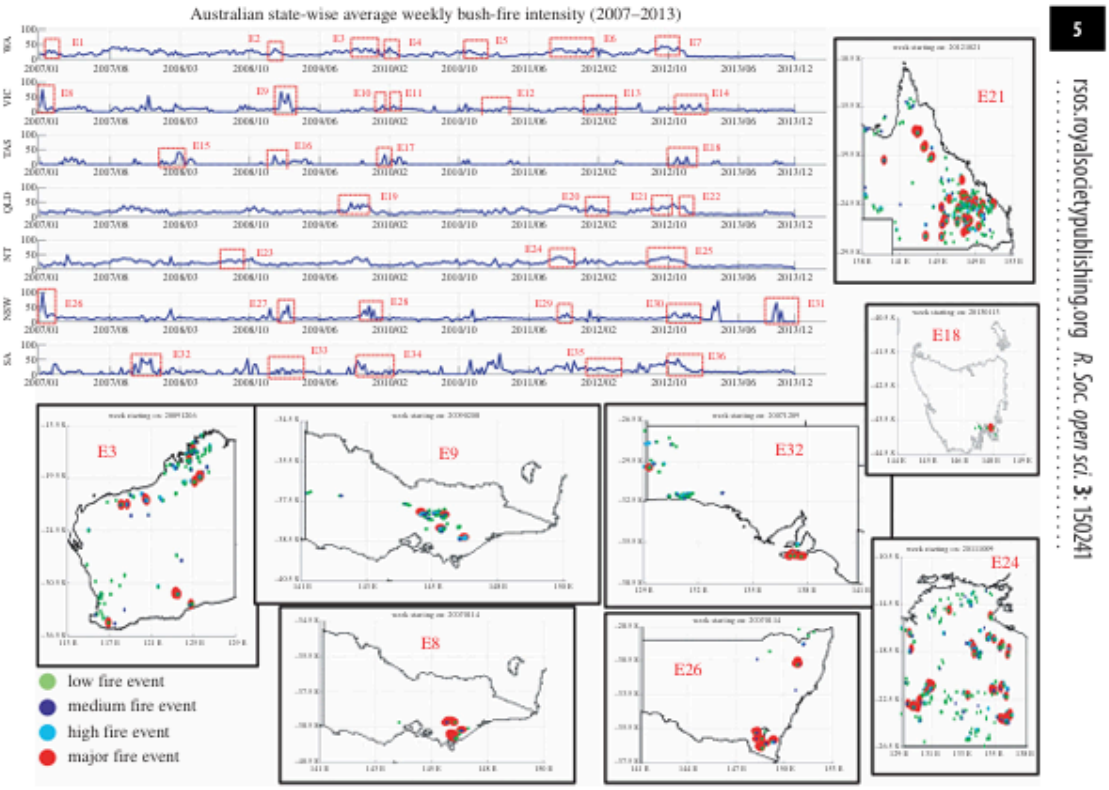


Inputs included NASA MODIS active fire data, burn area data, and weekly climatic surfaces.
Variables included: Incoming Solar Irradiance (SolarMJ), Maximum Temperature (TempMax), Minimum Temperature (TempMin), Precipitation (FWPrec), Soil Moisture Upper Layer (WRel1), Soil Moisture Lower Layer (WRel2), Evaporation Soil+Vegetation (FEW), Transpiration (FWTra), Soil Evaporation (FWSoil), Potential Evaporation (FWPT), Local Discharge Runoff + Drainage (FWDis), Surface Runoff (FWRun), Deep Drainage (FWLch2), Sensible Heat Flux (PhiH), Latent Heat Flux (PhiE), NASA bush-fire ground truth hot spots (wf_gt_awap) and wind speed (wf_gt_wind).
They tested over 336 weeks of data. Their model achieved 91% global accuracy in predicting bush-fire hotspots. They found that bagging and conventional k-Nearest Neighbour (KNN) classifier were the two best classifiers, with 94.5% and 91.8% accuracy, respectively.[5] They concluded that ML successfully captured complex relationships between climate and fire incidence, providing reliable predictive capability.[12]

The figure below represents Australian state-wise average weekly bush-fire intensity from NASA MODIS Active Fire data and NASA Burned Area data. Weekly averaged and normalized fire intensities (in a range of [0-100]) are plotted for seven Australian states (WA, VIC, TAS, QLD, NT, NSQ, SA). Historical verification and cross-validation of the derived bush-fire intensity data from NASA data products were performed. They found 36 major fire events occurred during 2007-2013 (indicated by red squares and labelled

[12] Dutta R., Das A., and Aryal J. 2016. Big data integration shows Australian bush-fire frequency is increasing significantly. R. Soc. Open Sci. 3(2): 150241.

E1-E36) from various data sources. To make the ground truth data used in this study more reliable and meaningful, they visualized eight of the most severe bush-fire events in Australian history [E21, E18, E24, E3, E32, E8, E26].



KNN algorithm is a very effective supervised classification algorithm based on intuitive premise that similar data points are in close proximity according to some metric.[5]

A study out of British Columbia built and evaluated a data-driven wildfire prediction system tailored to BC's specific conditions. They constructed a high-resolution dataset integrating 5 years of wildlife incident records from the Canadian Wildland Fire Information System (CWFIS), climate data from ERA5 reanalysis as well as environmental, meteorological and geospatial variables. They tested five predictive models: Random Forest, XGBoost, Light GBM, CatBoost, RNN + LSTM (DL). They found CatBoost achieved the highest metrics with an accuracy of 93.4%. Random Forest also performed well with ~92.6% accuracy. The most influential environmental and climatic features for wildfire occurrence included: surface temperature, humidity, wind speed and soil moisture.[13]

CatBoost is a ML algorithm for classification, regression, and ranking that is especially strong at handling categorical variables directly, without extensive preprocessing. CatBoost is based on gradient boosting,

[13] Wildfire Prediction in British Columbia Using Machine Learning and Deep Learning Models: A Data-Driven Framework [https://www.mdpi.com/2504-2289/9/11/290]

which builds an ensemble of decision trees sequentially. Each new tree attempts to correct the errors of previous trees.[14]

A study in Russia set out to build a unified data-driven pipeline for predicting wildfire occurrence. They developed a unified pipeline for data acquisition and subset ML-based algorithm development. They analyzed the following algorithms:[15]

- Deep Learning: ConvLSTM, RegNetX (CNN), Attention MLP (AMLP)
- Machine Learning: Random Forest and XGBoost
- Anomaly Detection: Autoencoder (AE)

They collected a unique dataset covering several large regions in central Russia, incorporating more than 17,000 verified wildfire events over a period of 10 years.

The areas studied were as follows:

Amur:

- 2 different climate zones, dominated by monsoon-influenced subarctic climate
- Population / Land: ~750,000 / 361 900 km^2
- Avg Jan Temperature: -23.5 to -21.8°C
- Avg July Temperatures: +18 to +21.2°C
- Avg Annual Precipitation: ~674 mL

Irkutsk:

- Characterized by subarctic climate
- Population / Land: ~ 2,330,000 / 774 800 km^2
- Avg Jan Temperature: -20.6 to -19.6°C
- Avg July Temperature: +18.1 to +20°C
- Avg Annual Precipitation: ~454 mL

Rostovo:

- Hot humid continental climate
- Population / Land: ~4,150,000 / 101 000 km^2
- Avg Jan Temperature: -3.5 to -1.9°C
- Avg July Temperature: +24.2 to +24.9°C
- Avg Annual Precipitation: ~460 mL

Sverdlovsk:

- Dominated by a moderately continental climate
- Population / Land: ~4,230,000 / 194 300 km^2
- Avg Jan Temperature: -14.7 to -14.3°C

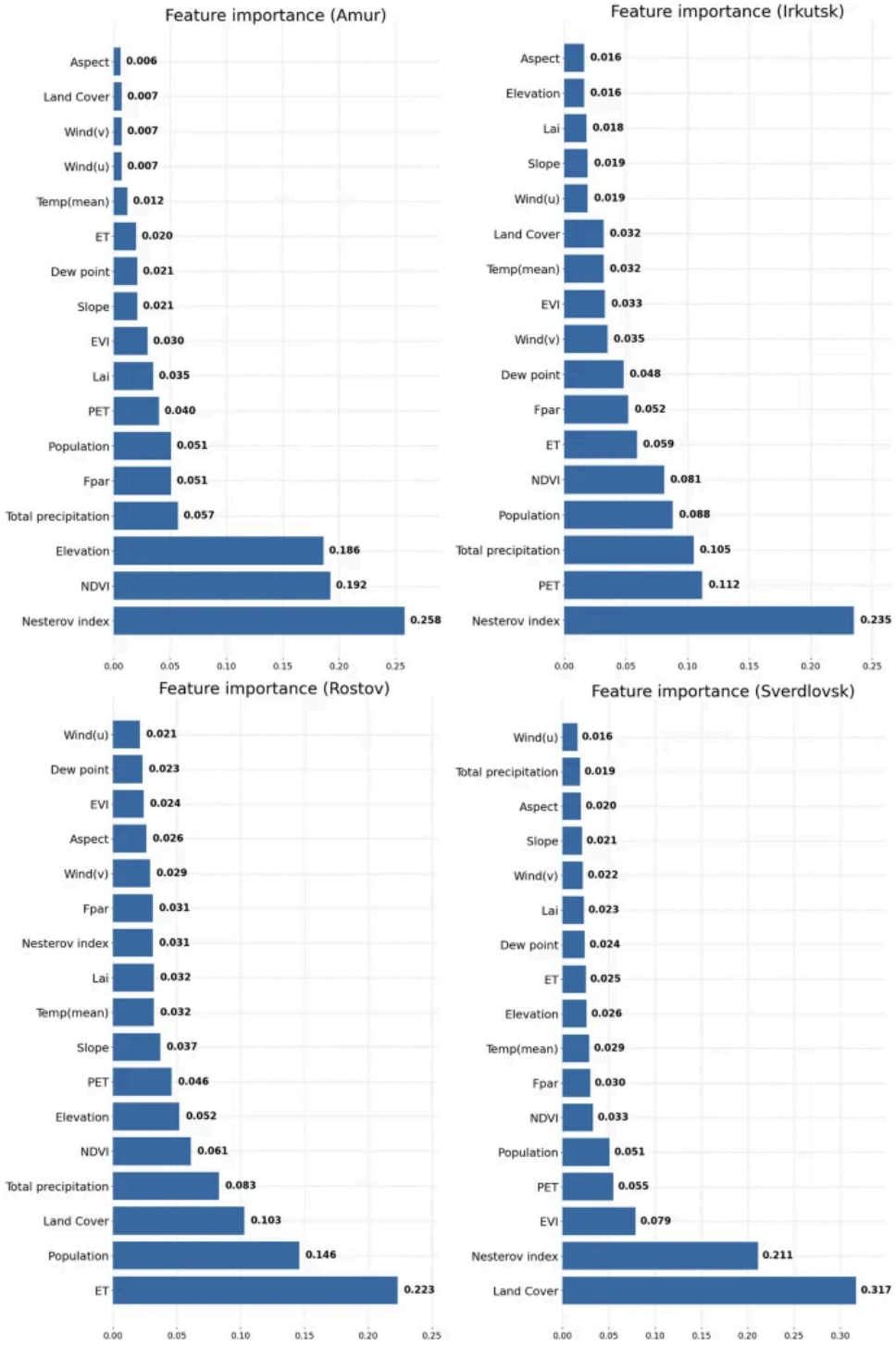
---

[14] *Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin — CatBoost: unbiased boosting with categorical features* (2017)

[15] Illarionova S, Shadrin D, Gubanov F, Shutov M, Tasuev U, Evteeva K, Mironenko M, Burnaev E. Exploration of geo-spatial data and machine learning algorithms for robust wildfire occurrence prediction. Scientific Reports. 2025 Mar 28;15(1):10712.

- Avg July Temperature: +17.9 to +19.2°C
- Avg Annual Precipitation: ~601 mL

XGBoost featured the importance for different regions:



Their findings underscore the necessity of developing individual ML models tailored to each region, taking into account the specific environmental features correlated with the probability of fire occurrence. No single approach consistently yielded the best results across all regions. Each region, with different

climatological zones, requires a tailored approach. DL models were generally better at predicting non-fire (fewer false positives). Classical ML models were better at predicting fire (higher recall). The autoencoder was consistently worse.

For Amur and Irktsk (Siberia), Nesterov index, Normalized Difference Vegetation Index (NDVI), Potential Evapotranspiration (PET), precipitation were dominant features.

For Rostove, Evapotranspiration (ET), population density and land cover matter most.

For Sverdlovsk (Ural forest), Nestervo index and land cover dominate.

XGBoost consistently delivered strong results and is easier to interrupt. Using XG Boost's Gain metric, Nestervo index is the most important feature in 3 of 4 regions. Vegetation indices (NDVI, EVI, LAI, FPARP) are consistently important. Precipitation and PET/ET matter in moist-sensitive regions. Furthermore, population density is a strong predictor in human-dominated landscapes.

Their meteorological data analysis compared weather distributions at different offsets from fire days (0, 1, 3, 5 days before. Some features (i.e., Nestervo index, precipitation) showed statistically significant differences. However, weather alone could not precisely pinpoint the fire day, but still provided strong predictive signals. *The Nesterov index* is a key indicator, capturing the cumulative impact of heat and drought on wildfire risk.

The authors concluded that a unified wildfire prediction model is not feasible and regional models perform better. They found that environmental features vary in importance by region, driven by climate, vegetation and human activity. Classical ML models (RF, XGBoost) remained highly competitive. DL models excel in spatial generalization and reducing false alarms.

XGBoost is a distributed, open-source ML library that uses gradient boosted decision trees, a supervised learning boost algorithm that makes use of gradient descent. XGBoost is known for its speed, efficiency and ability to scale well with large datasets. Boosting combines multiple individual weak trees (i.e., models that perform slightly better than random chance), to form a strong learner. Each weak learner is trained sequentially to correct the errors made by the previous models.[16]

Definitions:[15]

Evapotranspiration (ET)

- ET measures the amount of water transpired by plants and evaporated from the soil surface. Low ET values indicate drier conditions, potentially leading to increased vegetation stress and higher fire risk.

Potential ET (PET)

- PET represents the maximum amount of water that could be evaporated from the soil and transpired by vegetation under prevailing environmental conditions. PET influences vegetation moisture stress, with higher PET values indicating greater water demand and potential vegetation desiccation.

Leaf Area Index (LAI)

[16] https://www.ibm.com/think/topics/xgboost

- LAI measures the total area of leaves per unit ground surface area. High LAI values suggest dense vegetation cover and greater fuel continuity.

Fraction of Photosynthetically Active Radiation (FPAR)

- FPAR quantifies the fraction of incoming solar radiation absorbed by vegetation canopy. High FPAR values indicate active vegetation growth and biomass accumulation, which can contribute to increased fuel loads and fire risk.

Enhanced Vegetation Index (EVI)
Normalized Difference Vegetation Index (NDVI)

- *NDVI and EVI are both widely used remote sensing indices that provide valuable information about vegetation health and density. Higher NDVI and EVI values typically indicate denser vegetation, which can serve as fuel for wildfires.*

Practical Implications:

The BC framework could be scalable to other provinces, including Alberta. A well engineered ML pipeline can deliver accurate, province-wide wildfire occurrence predictions and support proactive fire management. Although Alberta has AI tools, I could not find a scientific study equivalent to the BC study. Alberta (as well as other provinces) could benefit from BC's study methodology, feature engineering and model comparison framework. The same pipeline could be retrained on Alberta specific data. Alberta would have to consider other factors though, such as:

- Spring temperature anomalies
- Overwinter snowpack melt timing
- Drought code and soil moisture deficit
- Lightning density
- Chinook-driven humidity drops and wind patterns

Conclusions:

Machine Learning models can accurately predict wildfire occurrence. CatBoost and XBoost are strong candidates for operational early-warning systems. Random Forest performed relatively well as well. However, a unified wildlife prediction model is not feasible. Environmental features avery in importance by region, driven by climate, vegetation, and human activity. Therefore, regional models would perform better. High-resolution environmental data enables localized risk assessment. Common predictors appear to include temperature, relative humidity and precipitation.