A Comparative Analysis of Machine Learning Models for Wildfire Prediction

Brent Kong

Received April 22, 2024
Accepted June 19, 2024
Electronic access June 30, 2024

Amid the increasing threat of wildfires driven by extreme weather events, the imperative development of rapid and efficient fire prediction models becomes evident for successful evacuation, suppression efforts, and air quality forecasts. This paper thoroughly analyzes ten popular machine learning models to evaluate their effectiveness in distinguishing meteorological and topographical data as conducive or non-conducive to fire occurrence. To overcome the limitation of a restricted dataset, a Conditional Tabular Generative Adversarial Network (CTGAN) is employed to enhance the dataset by generating synthetic samples. Detailed assessments are conducted on the original and augmented datasets. Across varying data quantities and distributions, CatBoost and AutoGluon consistently demonstrated exceptional performance, establishing themselves as robust models for wildfire prediction. The noteworthy predictive accuracy displayed by these two models underscores their suitability for real-world fire forecasting applications, particularly in the high- risk regions of California and Canada.

Introduction

Wildfires have the potential to cause extensive devastation, engulfing entire communities and contributing substantially to climate change. From 2017 to 2021, wildfires in the United States have caused about $80 billion in damages[1]. Due to the effects of global warming, there is an anticipated growth in both the number and intensity of wildfires. Intense heat and prolonged drought lead to evaporation of moisture from the soil and vegetation, creating an environment conducive to fire outbreaks. The interplay among topography, fuel, and weather is illustrated in Fig. 1. Elevated global temperatures and extreme weather events contribute to changes in topography, affecting the abundance of biomass fuels. This, in turn, results in an escalation of forest fires, which emitted a substantial 1.7 billion tons of carbon emissions globally in 2023[2]. Consequently, wildfires play a significant role in exacerbating climate change[3], contributing to a substantial release of greenhouse gases. The alteration in climate patterns may render certain areas more vulnerable to wildfires.

A possible solution to the influx of wildfires is the application of machine learning (ML), a technique that perceives patterns in datasets to forecast future statistics or other consequences of interest[3]. In the context of environmental considerations, the primary drivers of forest fire ignition are fuel, weather, and topography. By leveraging these key features, authorities can swiftly anticipate forest fires' potential location, size, and intensity, enabling proactive preparation and response measures. Many ML models have been developed for wildfire prediction, each offering distinct advantages and disadvantages under various data conditions. However, wildfire prediction datasets are relatively scarce due to the infrequency of forest fires and the complexities involved in data collection[4].

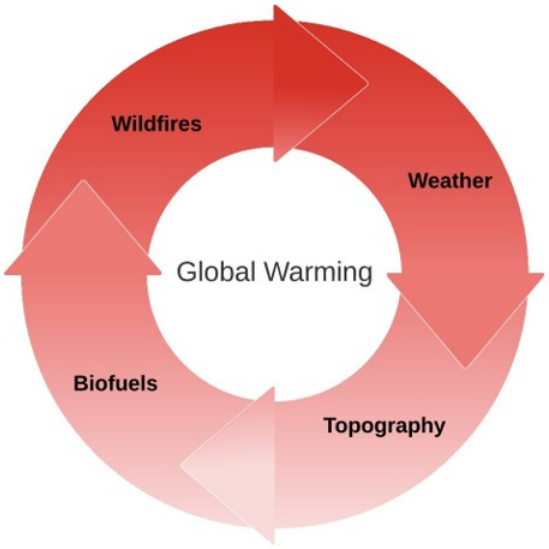

Fig. 1 Weather, topography, biofuels, and wildfires.

Literature Review

Forest fire monitoring increasingly relies on machine learning modeling, incorporating key factors such as forest composition,

weather patterns, and topographical characteristics. Cortez et al. introduced support vector machine (SVM) for forest fire prediction in the northeast region of Portugal by utilizing temperature, relative humidity, wind, and rain as features[5]. Their study compared SVM to neural networks, multiple regression, decision tree (DT), and random forest (RF). The study showed that the SVM model was the most effective in predicting the size of small forest fires.

Preisler et al. proposed statistical models to provide spatially explicit forecasts for the expected numbers, locations and costs of large fires in California[6]. Vegetation, topography, and hydroclimate were potential features, with spatial location, month-in-year, elevation, and percentage of forested land identified as significantly related to historic probabilities of large fires. Sayad et al. took a comprehensive approach by combining big data, remote sensing, and data mining to predict forest fires in Canada[7]. Their integrated methodology represents a multidimensional approach to fire prediction, leveraging diverse data sources and advanced algorithms. Data were acquired from the Moderate Resolution Imaging Spectroradiometer (MODIS), with three attributes considered as features for forest fires: normalized difference vegetation index (NDVI), land surface temperature (LST), and the fire indicator "Thermal Anomalies." The performances of both the Artificial Neural Network (ANN) and SVM were assessed, revealing outstanding results.

Chowdhury et al. introduced a data augmentation approach based on deep learning for wildfire risk prediction. They tested the approach on two datasets: one from the MERRA-2 satellite in California while the other from Sayad et al.[7]. These datasets were synthetically augmented using Conditional Tabular Generative Adversarial Networks (CTGAN). The authors incorporated weather, topographical, and vegetation information into their analysis. Their proposed model, a CTGAN-based neural network (NN), outperformed four other models (decision tree, random forest, gradient boosting machine, and support vector machine) on both the original and augmented datasets, achieving a peak accuracy of 79.31%. Lastly, Langford et al. advocated for deep neural networks (DNN) with a validation-loss (VL) weight selection strategy for predicting forest fires in Alaska[8]. The comparison involved XGBoosting, VL-DNN, and standard DNN using wildfire and non-wildfire pixels. VL-DNN exhibited notable performance, showcasing a significant improvement.

Our Contribution

This paper conducts a comprehensive comparison of ten commonly utilized machine learning models for forest fire prediction, including ANN, CatBoost, SVM, Logistic Regression (LogitReg), AutoGluon, LightGBM, Random Forest, K- Nearest Neighbors, Ridge Classification, and Decision Tree. Given the multitude of ML models proposed by researchers across various datasets, our goal is to identify the optimal models for forest fire prediction through a standardized evaluation across four datasets. The assessment of model performance is based on datasets provided by Chowdhury et al.[9] (849 instances) and Sayad et al.[7] (1713 instances), which were later synthetically augmented with a Conditional Tabular Generative Adversarial Network (CTGAN). Notably, CatBoost and AutoGluon demonstrated exceptional performance across the four datasets analyzed in this study, achieving peak accuracies over 90%.

Methods

Machine Learning Models

This research considers the performance of Ridge Classification, K-Nearest Neighbors (KNN), DT, RF, CatBoost, LightGBM, LogitReg, SVM, ANN, and AutoGluon. They can be classified into three categories: baseline models, tree-based regressors, and deep learning.

1. **Baseline models** are well-established models in the field of machine learning. They are ideal for quickly identifying patterns and relationships within data. Ridge classification combines conventional classification techniques and ridge regression ideas to prevent overfitting[10]. K-Nearest Neighbors (KNN) uses proximity to make predictions about the grouping of an individual point[11]. Logistic regression (LogitReg) fits data to an "S" shaped logistic function and outputs the probability of an instance belonging to a particular class[12]. Support vector machine (SVM) finds a hyperplane in N-dimensional space (N is the number of features) that distinctively classifies the points[13].

2. **Tree-based regressors** are models composed of numerous decisions trees. A Decision Tree (DT) appears as a flowchart, featuring a tree structure composed of a root node, branches, internal nodes, and leaf nodes[14]. Random Forest contains decision trees separately trained on small subsets of the data, and a majority vote determines the final output. CatBoost adopts a level-wise tree growth approach, where each decision tree aims to improve upon the errors of the previous one by level[15]. Lastly, LightGBM is a gradient-boosting model that employs a leaf-wise tree growth strategy, selecting the leaf with the maximum delta loss for expansion.

3. **Deep learning** solutions use multi-layered networks to generate predictions. Artificial Neural Network (ANN) resembles the structure of a human brain[16], containing an input layer, an output layer and several hidden layers. A layer can have many neurons (units) depending on the complexity of the dataset, citing it as a powerful model. AutoGluon is an automated ML library[17] that can quickly

prototype deep learning and classical machine learning solutions. AutoGluon experiments with various SOTA tabular data prediction models and chooses the best model for a given task.

Dataset Information

Fuel, weather, and topography are the most substantial drivers of forest fire ignition[3]. Thus, it is important to capitalize on these features when training ML models to predict the occurrence of forest fires.

Accurate fire predictions necessitate careful consideration of fuel density and moisture content. Normalized difference vegetation index (NDVI) and enhanced vegetation index (EVI) are important vegetation indices. NDVI and EVI can capture chlorophyll concentration and the extent of green coverage from satellite images. Areas rich in green vegetation are less prone to catching fire, as living organisms are less combustible than dead ones. Conversely, the concentration of vegetation influences its flammability, as closely packed fuel particles can ignite each other, leading to rapid fire growth[18].

Weather plays a pivotal role in influencing the spread and escalation of forest fires. Wind, a significant factor, provides oxygen and propels the fire towards new fuel sources. Atmospheric and surface temperatures impact the heat of fuels, influencing their susceptibility to ignition[18]. Additionally, humidity, the amount of water vapor in the air, affects the moisture content of fuels. Fuels with higher moisture levels impede fire growth, as the fire must first evaporate the water. Lastly, atmospheric and surface pressure patterns are linked to terrain-induced foehn winds[19], leading to sudden warming and drying, further impacting fire conditions. Topography also influences the growth and dispersion of wildfires. Therefore, it is essential to consider the digital elevation model (DEM). Higher elevations tend to be drier than lower elevations, heightening the risk of fire.

Finally, a common cause of forest fire ignition is electrical equipment and power line failures. When a power line fails due to weather or other factors, it may remain energized until the utility company completely shuts it down. The potential for a fire to easily ignite arises when vegetation comes into contact with a downed conductor.

The features mentioned above are incorporated into the two datasets under consideration for this study:

- **Calif-Fire** dataset[9] pertains to the two most devastating wildfires, 'Camp' and 'Tubbs,' which occurred in Northern California during 2017-2018. The data points were collected from the MERRA-2 and the Landsat 8 satellites, and California Energy Commission (CEC). It has ten features: temperature, DEM, distance to power lines, eastward wind, northward wind, NDVI, EVI, surface pressure, water vapor, and specific humidity. Weather parameters, including temperature, specific humidity, precipitable water vapor, surface pressure, and eastward and northward wind, were collected from the MERRA- 2 dataset. Vegetation indices (NDVI and EVI) and topographic data (DEM) were obtained from the USGS, using data from the Landsat 8 satellite. Additionally, the locations of electric transmission lines were sourced from the CEC. The dataset exhibits balance, with 420 instances for negative class-0 and 429 instances for positive class-1.
- **Can-Fire** dataset[7] is composed of multiple zones located in the center of Canada, approximately 2 million hectares of Canada's largest forests. These zones vary in size, burn period, and extent. The data was collected from the Terra and Aqua satellites via MODIS. It includes three key features: NDVI, LST, and thermal anomalies. NDVI was obtained from the Terra satellite, while LST was provided by MODIS. Lastly, thermal anomalies, which serve as a fire indicator measuring the likelihood of a fire on a scale from 1 to 10, were also collected by MODIS. The dataset is imbalanced: 1327 instances belong to the negative class-0 while 386 instances for the positive class- 1.

Considering the aforementioned two datasets, some biases can be observed. First, the Calif-Fire dataset is noticeably small, which may prevent machine learning models from effectively learning the relationship between features. Second, the Can-Fire dataset is imbalanced, with a 3:1 ratio of negative to positive classes. The skewed distribution would favor negative predictions, hindering the efficacy of ML models. These limitations were addressed with the addition of synthetic instances via CTGAN, as discussed in the following section.

Data Augmentation via CTGAN

Modeling the intricate connection between environmental conditions and wildfire occurrences demands abundant data, yet the Calif-Fire and Can-Fire datasets offer limited instances. This study additionally explores the impact of data augmentation on all ten machine learning models. Data augmentation applies various techniques to expand the dataset artificially. This can include adding noise, introducing variations, or perturbing existing values within the data points. The goal is to enhance the diversity of the tabular dataset, providing ML models with a more comprehensive and varied set of examples to learn from.

Generating synthetic data presents inherent challenges, particularly with heterogeneous datasets encompassing categorical and continuous values[20]. The inclusion of categorical features complicates the task, as models encounter difficulties in capturing the characteristics of real data due to imbalances inherent in categorical data. Moreover, tabular datasets frequently deviate from a Gaussian distribution and may exhibit multiple modes, introducing complexities associated with distinct local maxima and minima.

Table 1 Optimal Values of Hyper-parameters

Model \ Dataset	Calif-Fire	Calif-Fire-Aug	Can-Fire	Can-Fire-Aug
Ridge	Alpha (0.3)	Alpha (0.1)	Alpha (0.001)	Alpha (0.1)
KNN	N_Neighbors (7) Weights (uniform) Metric (Manhattan)	N_Neighbors (21) Weights (uniform) Metric (Manhattan)	N_Neighbors (11) Weights (distance) Metric (Manhattan)	N_Neighbors (15) Weights (distance) Metric (euclidean)
CatBoost	Learning_rate (0.1) Depth (6) Iterations (100) Loss function (Logloss)	Learning_rate (0.5) Depth (3) Iterations (100) Loss function (Logloss)	Learning_rate (0.5) Depth (6) Iterations (50) Loss function (Logloss)	Learning_rate (0.1) Depth (3) Iterations (150) Loss function (Logloss)
ANN	Neurons (85) Hidden layers (10) Dropout (0.1) Learning rate (0.001)	Neurons (65) Hidden layers (2) Dropout (0.1) Learning rate (0.01)	Neurons (80) Hidden layers (2) Dropout (0.1) Learning rate (0.001)	Neurons (95) Hidden layers (8) Dropout (0.1) Learning rate (0.01)
LogitReg	Max_iter (100) C (1.0) Penalty (l1) Solver (liblinear)	Max_iter (2500) C (0.0001) Penalty (l2) Solver (sag)	Max_iter (100) C (0.0001) Penalty (l2) Solver (liblinear)	Max_iter (100) C (0.01) Penalty (l2) Solver (sag)
DT	Max depth (3)	Max depth (9)	Max depth (3)	Max depth (6)
LightGBM	Max_depth (5) Num leaves (6)	Max_depth (5) Num leaves (6)	Max_depth (5) Num leaves (10)	Max_depth (5) Num leaves (10)
RF	Max_depth (3) Max_features (log2) Max_leaf_nodes (3) N estimators (100)	Max_depth (6) Max_features (log2) Max_leaf_nodes (9) N estimators (150)	Max_depth (6) Max_features (none) Max_leaf_nodes (9) N estimators (150)	Max_depth (6) Max_features (sqrt) Max_leaf_nodes (9) N estimators (50)
SVM	Kernel (rbf) γ (0.1) C (10)	Kernel (rbf) γ (0.01) C (100)	Kernel (rbf) γ (0.1) C (100)	Kernel (rbf) γ (0.1) C (100)
AutoGluon	WeightedEnsemble L2	WeightedEnsemble L2	WeightedEnsemble L2	WeightedEnsemble L2

Description: *A 5-fold cross validation was conducted on all ten models using the validation subsets. The table summarizes the optimal hyperparameters for each model.*

To address these challenges, we employed a conditional tabular generative adversarial network (CTGAN) for generating synthetic data points. CTGAN, a deep learning-based approach, discerns patterns within the existing dataset and creates new synthetic samples[9]. Notably, CTGAN proves robust in handling heterogeneous datasets compared to alternative methods like WGAN and PacGAN [20], specifically designed to tackle the intricacies associated with tabular data. CTGAN utilizes mode-specific normalization to capture the multimodal distributions within datasets by incorporating a conditional generator to address imbalanced categorical values. As noted Xu et al.[21], the conditional generator handles imbalanced categorical values by evenly sampling all categories, ensuring the representation of minor classes. The model produces high- quality and realistic samples, demonstrating versatility across various domains.

Despite the efficacy of CTGAN, the method exhibits several potential drawbacks. The quality of synthetic data largely depends on the characteristics of the training data. High- cardinality features pose challenges to CTGAN as it is difficult to capture the properties of datasets with a plethora of unique classes. Furthermore, CTGAN struggles with distributions with a large number of constant values[20]. Lastly, CTGAN's accuracy may be limited on smaller datasets, as deep learning models require extensive data.

Although the performance of CTGAN largely depends on the quality and quantity of data, the versatility and effectiveness of its synthetic data complement small and imbalanced datasets, notably the wildfire datasets provided in[7,9]. For synthetic data generation, the Calif-fire and Can-fire datasets were split by class. Each class was independently processed by CTGAN, generating 5000 synthetic instances per class. The choice of 5000 synthetic instances was determined to ensure statistical significance and enhance model stability and generalization across various datasets. Previous studies and experimental trials indicated that this quantity strikes a balance between generating enough data to effectively train the models and avoiding excessive computational costs. The addition of these new instances was also sufficient to accurately assess the performance of the

models and balance the skewed class distribution. The resulting synthetic data, combined with the original datasets, produced augmented datasets with sizes of 10849 and 11713 instances, denoted as **Calif-Fire-Aug** and **Can-Fire-Aug** datasets, respectively.

Data Preprocessing and Hyperparameter Tuning

Before fitting the models, the original and CTGAN augmented datasets were standardized using StandardScaler from the sklearn module. The data points were standardized according to the formula:

$$z = \frac{x - \mu}{\sigma}, \tag{1}$$

where μ is the mean and σ is the standard deviation. The process is done to ensure a similar scaling of the input values. All four datasets were then split into training, validation, and testing subsets in the ratio of 6:2:2.

Using the validation subset, a 5-fold cross-validation grid search was conducted on each of the ten models to determine the optimal hyperparameters. Table I details the set of optimal hyperparameters for both the original and augmented datasets. Depending on the model, various parameters were adjusted for maximum accuracy, including learning rate (rate of optimization), max depth (depth of decision tree), and iterations (batches of data). After optimal hyperparameters were selected, all models were fit with the training data. Next, the testing dataset was inputted into all ten models, and their performances were compared across the four datasets. The results are discussed in the following section.

Modules

We utilized various tools to conduct extensive simulations. Specifically, Google Colab served as the primary web-based, interactive computing environment. Key modules and tools included NumPy, Pandas, and Matplotlib. Scikit-learn provided a comprehensive library of data preprocessing methods and machine learning models. Additionally, the study incorporated LightGBM, AutoGluon, TensorFlow, and CatBoost for their extensive ML models. Collectively, these tools formed the computational foundation for the analysis and modeling of this research.

Results

Performance Metrics

Given that the task is a classification problem, we evaluate the models based on accuracy. Accuracy is defined as the fraction of correct predictions:

Table 1 Model Performance on CALIF-FIRE with 849 Instances

Model	Accuracy	Precision	Recall	F1	ROC-AUC
CatBoost	0.8059	0.7476	0.9167	0.8235	0.8072
SVM	0.8000	0.7500	0.8929	0.8152	0.8011
AutoGluon	0.7765	0.7054	0.9405	0.8061	0.7784
ANN	0.7588	0.7654	0.7381	0.7515	0.7586
LightGBM	0.7529	0.7059	0.8571	0.7742	0.7542
KNN	0.7471	0.7113	0.8214	0.7624	0.7479
RF	0.7471	0.6694	0.9643	0.7902	0.7496
LogitReg	0.7353	0.6970	0.8214	0.7541	0.7363
Ridge	0.7353	0.6822	0.8690	0.7644	0.7368
DT	0.7235	0.6581	0.9167	0.7662	0.7258

Description: CatBoost and SVM achieved accuracies above 80%, with competitive Precision, Recall, and F1. ROC-AUC values are closely correlated with accuracy values.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \tag{2}$$

Furthermore, we take additional performance metrics into account, including precision, recall, and F1 score, defined as follows:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{3}$$

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{4}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Precision measures the accuracy of positive predictions while recall measures the completeness of positive predictions. F1 score is the harmonic mean of both precision and recall, and it captures the traits of each measure.

Due to the small and imbalanced nature of the datasets, we introduce ROC-AUC as an additional metric. ROC-AUC is defined as the area under the receiver operating characteristic curve (ROC), which summaries the performance of binary classification models on the positive class[22]. This metric evaluates the ability of ML models to distinguish between positive and negative instances on a scale from 0 to 1. Values close to 1 indicate a perfect model, while values near 0.5 suggest models with no discriminative capability (i.e., always predicting the same class). Values below 0.5 reveal models that perform worse than random guessing. This metric complements the accuracy metric, particularly for imbalanced datasets where accuracy alone can be misleading. Finally, a high- performing model should exhibit high values for accuracy, precision, recall, F1-score, and ROC-AUC.

Table 2 Model Performance on CAN-FIRE with 1713 Instances

Model	Accuracy	Precision	Recall	F1	ROC-AUC
AutoGluon	0.8163	0.7059	0.3117	0.4324	0.6370
KNN	0.8017	0.6286	0.2857	0.3929	0.6184
CatBoost	0.7872	0.6111	0.1429	0.2316	0.5583
DT	0.7813	1.0000	0.0260	0.0506	0.5130
RF	0.7813	0.5556	0.1299	0.2105	0.5499
SVM	0.7813	0.6667	0.0519	0.0964	0.5222
ANN	0.7784	0.6000	0.0390	0.0732	0.5157
LightGBM	0.7668	0.4595	0.2208	0.2982	0.5728
LogitReg	0.7668	0.3333	0.0390	0.0698	0.5082
Ridge	0.7668	0.3333	0.0390	0.0698	0.5082

Description: AutoGluon and KNN achieved accuracies above 80% but scored low on Recall and F1. There is a disparity between accuracy and ROC-AUC values.

Table 3 Model Performance on CALIF-FIRE-AUG with 10849 Instances

Model	Accuracy	Precision	Recall	F1	ROC-AUC
CatBoost	0.9244	0.9035	0.9502	0.9263	0.9244
AutoGluon	0.9180	0.8974	0.9437	0.9200	0.9180
LightGBM	0.9051	0.8804	0.9373	0.9080	0.9051
SVM	0.8442	0.8336	0.8598	0.8465	0.8443
ANN	0.8387	0.8522	0.8192	0.8354	0.8387
DT	0.8309	0.7803	0.9207	0.8447	0.8310
KNN	0.8295	0.8000	0.8782	0.8373	0.8295
RF	0.8235	0.8291	0.8146	0.8218	0.8235
Ridge	0.7567	0.7896	0.6993	0.7417	0.7566
LogitReg	0.7493	0.7755	0.7011	0.7364	0.7493

Description: CatBoost, AutoGluon, and LightGBM achieved accuracies above 90% and high Precision, Recall, and F1. ROC-AUC values match accuracy.

Table 4 Model Performance on CAN-FIRE-AUG with 11713 Instances

Model	Accuracy	Precision	Recall	F1	ROC-AUC
CatBoost	0.8212	0.8441	0.7493	0.7939	0.8158
AutoGluon	0.8195	0.8457	0.7428	0.7909	0.8137
LightGBM	0.8195	0.8421	0.7474	0.7919	0.8141
DT	0.8075	0.8623	0.6917	0.7676	0.7989
RF	0.8045	0.8443	0.7047	0.7682	0.7971
ANN	0.7964	0.9298	0.6026	0.7313	0.7819
KNN	0.7896	0.8252	0.6880	0.7504	0.7820
SVM	0.7776	0.8441	0.6332	0.7236	0.7669
LogitReg	0.7559	0.7507	0.7020	0.7255	0.7518
Ridge	0.7559	0.7440	0.7149	0.7292	0.7528

Description: CatBoost, AutoGluon, and LightGBM achieved accuracies about 82%. Half of the models performed above 80% accuracy. ROC-AUC values correspond with accuracy scores.

Model Performance

1. **Original Datasets:** Table II outlines the performance of the ten ML models on the Calif-Fire dataset. Accuracy, precision, recall, F1, and ROC-AUC are evaluated for each model. Clearly, CatBoost and SVM achieve the highest accuracy at 80.59% and 80.00%, suported by ROC-AUC scores of 0.8072 and 0.8011. AutoGluon exhibits balanced performance, while ANN, LightGBM, and KNN consistently perform well. RF, LogitReg, Ridge classifier, and DT contribute to the overall analysis, revealing diverse strengths and weaknesses across the models in the context of wildfire prediction on this dataset. In general, accuracy is closely correlated with ROC-AUC, with a minimum ROC-AUC value of 0.7258. This indicates that the models possess discriminatory capability and can effectively understand the dataset. Table III provides the performance of ten models on the **Can-Fire** dataset. Notably, AutoGluon exhibits the highest accuracy at 81.63%. However, there is a disparity between accuracy and ROC-AUC: ROC-AUC is much lower at 0.6370. Furthermore, KNN and CatBoost demonstrate good overall accuracy but lacked in recall, F1, and ROC-AUC. The same trend is observed with DT, RF, SVM and ANN, with ROC-AUC values approaching 0.5. LogitReg and Ridge classifiers have poor results, indicating they are unable to learn the dataset. This dataset demonstrates that accuracy can be a misleading metric: despite having accuracies similar to those of the **Calif-Fire** dataset, the models exhibit little to no discriminatory power. Therefore, it is crucial to consider other metrics before making a sound judgment.

2. **Augmented Datasets:** Table IV presents the performance evaluated on the augmented dataset **Calif-Fire-Aug**. CatBoost and AutoGluon are top-performing models with accuracy scores of 92.44% and 91.80%, respectively, showcasing high precision, recall, F1, and ROC-AUC scores. LightGBM and SVM exhibit commendable performance, while ANN, DT, KNN, and RF demonstrate competitive results. LogitReg and Ridge classifiers have the lowest accuracy. The minimum ROC-AUC is 0.7493 while the maximum is 0.9244, suggesting most models are able to learn and achieve fine performance. Furthermore, ROC-AUC values closely align with accuracy, reinforcing the effectiveness of top-performing models. These results underscore the benefits of synthetic data. Finally, Table V displays the results tested on the augmented dataset **Can-Fire-Aug**. For this dataset, CatBoost, AutoGluon and LightGBM are the leading models, achieving accuracy scores of 82.12%, 81.95%, and 81.95%, respectively. These models exhibit a balance of precision, recall, and F1 scores,with ROC-AUC values matching accuracy scores. DT also delivers competitive results, demonstrating its efficacy. RF, ANN, KNN

and SVM exhibit varying performance levels, emphasizing the importance of model selection for effective predictions. Again, Ridge and LogitReg have lower accuracy, underlining a potential limitation of these two benchmarks. In contrast to the **Can-fire** dataset, the ROC-AUC values reveal that the models are able to learn and generate accurate predictions, indicating the benefits of data augmentation.

CTGAN Effects

Figure 2 illustrates the difference in model accuracy between the Calif-fire and the Calif-fire-aug datasets. The CTGAN synthetic had varying effects on each of the ten models. Models such as CatBoost, AutoGluon, and LightGBM experience a 10-12% increase in accuracy on the augmented datasets. However, for models such as Ridge and LogitReg, the improvement is limited, roughly 1-2% increase in performance. Despite this, Figure 2 suggests that synthetic data has potential in the field of machine learning. ROC-AUC values are closely tied to accuracy, demonstrating that the models are able to better understand the patterns within the data.

Figure 3 depicts the change in model accuracy between the Can-fire and Can-fire-aug datasets. In contrast to the Calif-fire dataset, the effects of synthetic data initially appeared to be mixed. Certain models, such as CatBoost, DT, and ANN, show a slight increase in accuracy, roughly 1-2%. On the other hand, models such as KNN, SVM, and LogitReg experience a decrease in accuracy. Yet, the figure is misleading as it only considers accuracy: ROC-AUC values indicate a major improvement on the augmented over the original datasets. On average, ROC-AUC values increase from 55% to 79%, implying the models better understand the dataset with the aid of synthetic data reducing class imbalance. This situation reflects the importance of considering various metrics when evaluating model performance.

Discussion

Implications and Significance

CatBoost and AutoGluon consistently emerged as the top-performing models across different datasets, achieving peak accuracies of 92.44% and 91.80%, respectively. Their exemplary performance across all four datasets can be attributed to their unique characteristics and design. CatBoost is specifically designed for handling categorical features. Likewise, AutoGluon easily adapts to various datasets as it prototypes numerous ML solutions. Their unique capabilities give them an edge over other models, even on imbalanced datasets such as Can-fire.

Interestingly, the effectiveness of top baseline models, notably KNN and SVM, varies depending on the dataset. While these models were top performers on the original datasets, their efficacy relative to other models diminishes on the augmented datasets. Overall, KNN demonstrates suitability for the Can-fire dataset, while SVM proves proficient with the Calif-fire dataset. The fluctuating performance of baseline models implies they are dataset dependent and useful under certain circumstances. These models are relatively simple and perform well on smaller datasets.

Conversely, LogitReg and Ridge show sub-optimal performance across all four datasets, likely due to their inherent simplicity and the nature of the datasets. Furthermore, despite its widespread acclaim, ANN displays mediocre performance, highlighting a contrast between the underwhelming performance of deep-learning models and the exceptional performance of tree-based models: an observation seen in general machine learning[23]. While deep-learning solutions have shown promise in image and text recognition, their effectiveness in tabular datasets remains uncertain. Perhaps the conditions of these particular datasets posed inherent challenges to ANN: the imbalanced, heterogeneous, and deficient nature of the dataset reduced its learning capacity, especially on the Can-fire dataset.

Regarding data augmentation, the models clearly demonstrated improved performance on the augmented data for the Calif-fire dataset. Figure 2 shows the noticeable increase in accuracy for each model, reflecting a maximum increase of 10-12%. However, the effects of synthetic data on the Can-fire dataset appear unclear. Figure 3 displays mixed benefits of data augmentation, notably certain models (CatBoost and Light- GBM) show improved performance while the others (SVM, LogitReg, and Ridge) have decreased performance. Despite this, ROC-AUC values sustain significant model improvement due to data augmentation. ROC-AUC scores increased by roughly 0.2 from the Can-fire to the Can-fire-aug datasets, implying reduced model bias and increased model learning. These results showcase the merits of synthetic data while highlighting a potential drawback of solely relying on accuracy as the performance metric.

Upon reflection, several implications can be drawn. First, CatBoost and AutoGluon have shown merit in forest fire prediction, achieving accuracies above 90%. Second, synthetic data has shown promise in improving model performance by increasing the dataset size and reducing class imbalance. Finally, it is crucial to consider a range of metrics before drawing conclusions, particularly because accuracy can be misleading when dealing with imbalanced datasets. Furthermore, several enhancements can be considered. The datasets used in this study lacked time-related information, such as day of the week or month. It has been established that a direct correlation exists between season (wildfires are most common in summer and fall) and the day of the week (large fires commonly occur on weekdays). The inclusion of these additional features could enhance prediction accuracy.

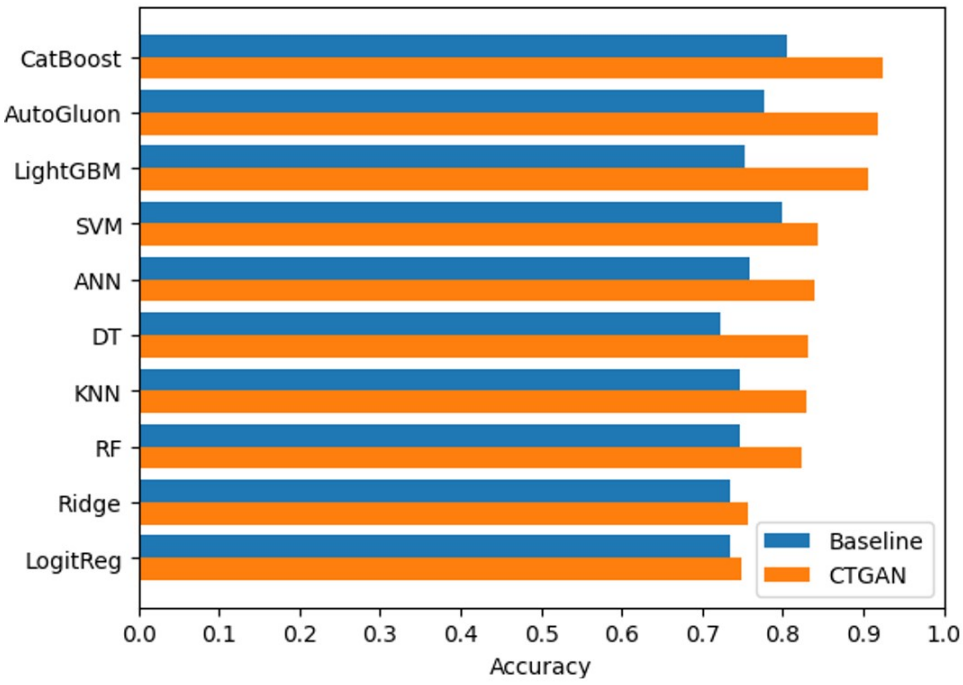

Fig. 2 Testing accuracy for Calif-Fire and Calif-Fire-Aug datasets. Cat- Boost, AutoGluon, and LightGBM exhibited a significant boost in accuracy.

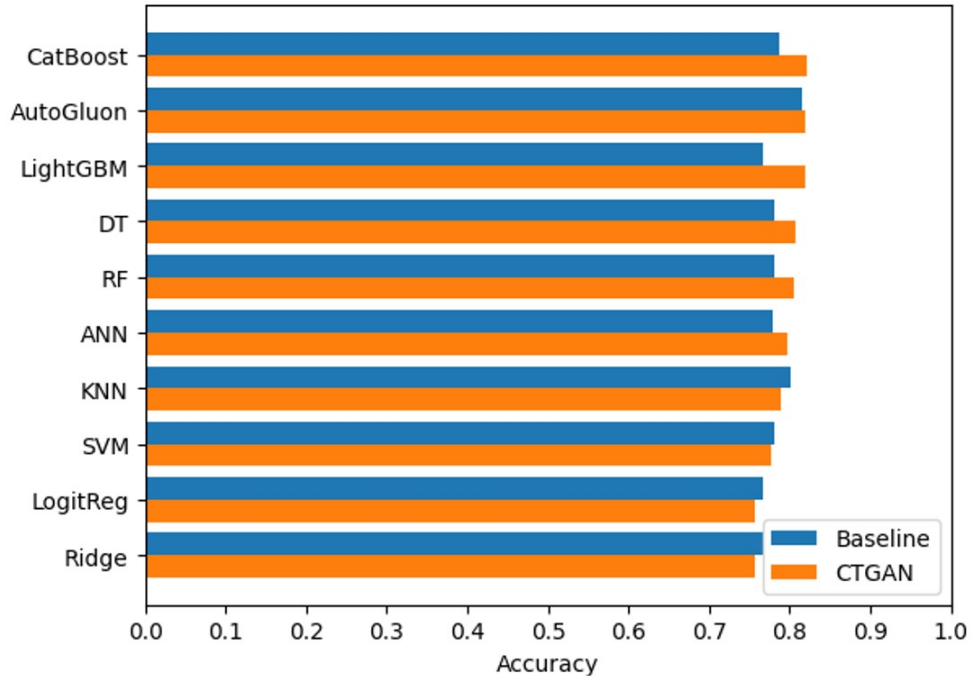

Fig. 3 Testing accuracy for Can-Fire and Can-Fire-Aug datasets. Synthetic data had mixed effects on model accuracy.

Conclusion

Wildfires pose severe threats to both communities and the natural environment, particularly in regions like California and Canada with prolonged periods of arid climate and limited rainfall. This paper contributes a comparative analysis of ten popular machine learning models for wildfire prediction and showcases the merits of synthetic data. The initial datasets[7,9] underwent cleaning and standardization before being augmented to larger datasets, each enriched with an additional 10,000 instances. CatBoost and AutoGluon demonstrated highly similar and accurate predictions for forest fires. Their success stems from their ability to manage heterogeneous datasets and adapt to varying dataset conditions. Data augmentation notably improved model understanding for both Calif-fire and Can-fire datasets, as shown by ROC- AUC values. This highlights the benefits of synthetic data in enhancing forest fire detection, especially in the face of limited available data.

A possible application of this research is wireless sensor networks (WSNs), a system designed to remotely monitor a phenomenon or event of interest[24]. In the proposed detection system, WSNs are deployed to cover the target region, with sensors monitoring environmental parameters such as air temperature, gases, and soil moisture. The acquired data can be analyzed via machine learning models for accurate forest fire prediction. Compared with other traditional methods, WSNs have shown promise for efficiently detecting forest fires. Both supervised and unsupervised ML models find application in WSNs for various purposes. Lee et al. integrated a RF model for feature selection and a Minmax Probability Machine for real-time Network Intrusion Detection System (NIDS) classification[25]. Dampage et al.[26] proposed a K-means technique in WSNs to detect forest fires, by using features like temperature, relative humidity, light intensity, and carbon monoxide levels. The models discussed in this paper can be adapted for WSNs by adjusting input features and output requirements. Models such as CatBoost and AutoGluon have shown promise in accurately classifying fire and no fire instances, indicating their potential utility in WSN applications.

Future endeavors will focus on exploring the impact of synthetically generated datasets on machine learning models. Additionally, there is an intention to develop robust forest fire prediction models incorporating temporal information.

Acknowledgements

The author expresses gratitude to Dr. Yu Zhang, Sifat Chowdhury, Shourya Bose, Kejun Chen, and Maina Dhar at UC Santa Cruz for their valuable resources and insightful discussions related to the topic.

References

1 *Wildfires*, https://www.edf.org/climate/heres-how-climate-change-affects-wildfires, Available:.

2 O. Milman and A. Witherspoon, *After a record year of wildfires, will canada ever be the same again?" Nov. 2023*, https://www.theguardian.com/world/2023/Nov./09/canada-wildfire-rec, Available:.

3 S. Singh, *Frontiers*, **5**, year.

4 G. Tylor, *Unveiling the challenges of ai data in wildfire detection: Exci's breakthrough*, https://www.exci.a, Available:.

5 P. Cortez and A. Morais, *A data mining approach to predict forest fires using meteorological data*.

6 H. Preisler, A. Westerling, K. Gebert, F. Munoz-Arriola and T. Holmes, *Intl. J. of Wildland Fire*, **20**, 508,.

7 Y. Sayad, H. Mousannif and H. Moatassime, *Fire Safety Journal*.

8 Z. Langford, J. Kumar and F. Hoffman, *2018 IEEE Intl. Conf. on Data Mining Workshops (ICDMW*, IEEE Computer Society, Los Alamitos, CA, USA, p. 770–778.

9 S. Chowdhury, K. Zhu and Y. Zhang, *Applied Energy Symposium: MIT A+B*.

10 A. Kumar, *Oct*.

11 *What is the k-nearest neighbors algorithm?" 2023*, https://www.ibm.com/topics/knn, Available:.

12 *Understanding logistic regression*, https://www.geeksforgeeks.org/understanding-logistic-regression/, Online]. Available:.

13 R. Gandhi, *Support vector machine - introduction to machine learning algorithms*, https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorit, Online]. Available:.

14 A. Saini, *Aug*.

15 https://catboost.ai/en/docs/concepts/parameter-tun, Online]. Available:.

16 *Artificial neural networks and its applications*, https://www.geeksforgeeks.org/artificial-neural-networks-a, Online]. Available:.

17 X. Yin, *Mar*.

18 *Wildland Fire Behavior (U.S. National Park Service*, https://www.nps.gov/articles/wildland-fire-behavior.htm.

19 *Critical fire weather*, https://www.nwcg.gov/pub, Available:.

20 M. Santos, *Apr*, 130– 146,.

21 L. Xu, M. Skoularidou, A. Cuesta-Infante and K. Veeramachaneni, *Modeling tabular data using conditional GAN*, Curran Associates Inc, Red Hook, NY, USA.

22 J. Brownlee, *Roc curves and precision-recall curves for imbalanced classification*, https://machinelearningm, Online]. Available:.

23 L. Grinsztajn, E. Oyallon and G. Varoquaux, *Why do tree-based models still outperform deep learning on tabular data?*

24 A. EDIS, *Ae521/ae521: What is a wireless sensor network?*, `https://edis.ifas.ufl.edu/publication/AE521`, Available:.

25 S. Lee, D. Kim and J. Park, *A hybrid approach for real-time network intrusion detection systems*.

26 U. Dampage, L. Bandaranayake, R. Wanasinghe, K. Kottahachchi and B. Jayasanka, *Scientific Reports*, **12**, year.