Revista Eletrônica de Iniciação Científica em Computação, 2026, 24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.7053

Research Paper

Predicting Active Fire Occurrence in the Brazilian Cerrado Using ConvLSTM Networks, Multi-Source Environmental and Anthropogenic Data

Guilherme G. de A. Vieira [Universidade do Distrito Federal Professor Jorge Amaury Maia Nunes – UnDF | *guilherme.vieira@undf.edu.br*]
Igor M. Santos Magalhães [Universidade do Distrito Federal Professor Jorge Amaury Maia Nunes – UnDF | *igor.magalhaes@undf.edu.br*]
Lucas Rocha de Santana [Universidade do Distrito Federal Professor Jorge Amaury Maia Nunes – UnDF | *lucas.santana@undf.edu.br*]
Elvis Miranda Teixeira [Universidade do Distrito Federal Professor Jorge Amaury Maia Nunes – UnDF | *elvis.teixeira@undf.edu.br*]
Kauã Maciel Veit [Universidade do Distrito Federal Professor Jorge Amaury Maia Nunes – UnDF | *kaua.veit@undf.edu.br*]
Jeovan Assis da Silva [Universidade do Distrito Federal Professor Jorge Amaury Maia Nunes – UnDF | *jeovan.silva@undf.edu.br*]

Universidade do Distrito Federal Professor Jorge Amaury Maia Nunes (UnDF), CEP 71503-502, Brasília, DF, Brazil.

Abstract. Wildfires in the Brazilian Cerrado combine high frequency, large spatial extent, and strong anthropogenic influence, resulting in substantial ecological, climatic, and social impacts. This paper investigates the short-term prediction of daily active fire occurrence in the Cerrado using convolutional long short-term memory (ConvLSTM) networks driven by multi-source environmental and anthropogenic data, with emphasis on satellite-based products. We build a spatiotemporal dataset on a regular grid by integrating active fire detections from the BDQueimadas system (INPE), meteorological variables from INMET and ERA5-Land, MODIS vegetation indices and land surface temperature, MapBiomas land-use and land-cover maps, SRTM topography, and distance-to-infrastructure layers. Each training sample consists of a short window of recent environmental conditions and a binary target map indicating whether at least one new active fire ignition occurs in each grid cell on the following day, thus framing ignition risk as a spatiotemporal forecasting problem. Although raw satellite and meteorological data are available from 2014 to 2024, limitations in memory and computation led us, in this initial prototype, to construct the full 5 km feature cube only for two representative years (2014 and 2017). The ConvLSTM model is trained on windows extracted from 2014 and evaluated on an independent test set from 2017. On this extremely imbalanced test set ($\approx 0.03\%$ positive cell–days), the model attains an area under the ROC curve (AUC-ROC) of about 0.89 at the pixel level, while precision–recall analysis and top-k evaluation highlight both nontrivial skill and the challenges of predicting rare fire ignitions. The study documents the data-engineering pipeline, the ConvLSTM configuration, and the evaluation protocol as a step toward operational early-warning tools for the Cerrado biome.

Keywords: Wildfire modeling; ConvLSTM; Spatiotemporal deep learning; Cerrado; Fire risk.

Received: 29 November 2025 • **Accepted:** 30 January 2026 • **Published:** 06 March 2026

1 Introduction

Wildfires in the Brazilian Cerrado combine high frequency, large spatial extent, and strong anthropogenic influence, generating significant ecological, climatic, and social impacts.

The Cerrado is recognized as one of the world's main biodiversity hotspots, combining high species richness, elevated endemism, and intense human pressure associated with agricultural expansion and unplanned land occupation [Colli and Vieira, 2020]. More than half of its original vegetation cover has already been converted to pasture and cropland, while legally protected areas remain limited relative to its ecological importance [Sano *et al.*, 2019]. This combination of high biodiversity, rapid land conversion, and limited protection makes the biome especially vulnerable to changes in its fire regime.

Fire plays a central ecological role in the Cerrado, but fire-related adaptations are associated with specific regimes of frequency, intensity, and seasonality. Syntheses of the literature indicate that policies of full suppression, combined with widespread use of fire for agricultural management, have altered historical regimes and produced negative impacts on vegetation and ecological processes [Durigan and Ratter, 2016]. Remote-sensing analyses show that high fire recurrence is associated with declines in vegetation indices and primary productivity, as well as increases in land-surface temperature, indicating progressive degradation [Santana *et al.*, 2020].

Beyond local ecological impacts, Cerrado wildfires contribute to greenhouse gas emissions, soil degradation, loss of ecosystem services, and deterioration of air quality in nearby urban centers. During years of severe drought, large areas of the Cerrado and neighboring biomes experience peaks in fire activity, with substantial social and economic consequences. In this scenario, tools capable of anticipating short-term active fire occurrence can support management decisions, enforcement actions, resource allocation, and preventive measures by environmental agencies and civil protection authorities.

Recent advances in satellite remote sensing and cloud-based platforms such as Google Earth Engine enable continental-scale integration of active-fire detections, vegetation indices, land-surface temperature, meteorological reanalyses, and land-use information [Gorelick *et al.*, 2017; Didan, 2021; Wan *et al.*, 2021; Muñoz-Sabater *et al.*, 2021]. In Brazil, the BDQueimadas system, maintained by the National Institute for Space Research (INPE), consolidates active-fire detections from multiple satellite sensors [INPE, 2023], while initiatives such as MapBiomas document land-use and land-cover change over recent decades [MapBiomas Project, 2025; Souza *et al.*, 2020].

In parallel, the wildfire-prediction literature has incorporated a growing number of machine learning (ML) and deep learning (DL) approaches. Supervised models have been used to forecast fire danger, ignition risk, spread, and burned area by combining historical fire records with meteorological, topographic, vegetation, and anthropogenic variables [Pan and Zhang, 2023]. In Brazilian and Latin American contexts, studies such as Sharma and Khanal [2024] illustrate the potential of spatial ML algorithms to estimate short-term risk maps from joint modeling of fire occurrence and environmental covariates.

At the same time, deep learning models originally developed for precipitation and air-quality forecasting have shown promise for spatiotemporal environmental tasks. Shi et al. [Shi *et al*., 2015] introduced the ConvLSTM architecture by formulating precipitation *nowcasting* as prediction of radar-image sequences from previous observations. Subsequent applications, such as DeepRain [Kim *et al*., 2017], and hybrid CNN–LSTM and ConvLSTM models for atmospheric-pollutant forecasting [Putri *et al*., 2024], reinforce the potential of spatiotemporal networks in complex environmental settings. More broadly, reviews of deep learning for time-series forecasting highlight the role of recurrent, convolutional, and attention-based architectures, while emphasizing challenges related to non-stationarity, class imbalance, and interpretability [Casolaro *et al*., 2023].

Despite these advances, most studies focus on temperate biomes and high-income regions, whereas applications in seasonally dry tropical biomes such as the Cerrado remain relatively scarce and rarely make explicit use of spatiotemporal deep architectures. Important gaps persist regarding the systematic assessment of ConvLSTM networks for predicting active fires in biomes strongly influenced by human activities, based on joint integration of environmental satellite products, meteorological reanalyses, and geospatial information.

Against this backdrop, this study addresses the following research question: to what extent can ConvLSTM networks, driven by historical sequences of environmental and anthropogenic maps derived predominantly from remote sensing, predict daily active-fire occurrence in the Brazilian Cerrado?

To address this question, we investigate the use of ConvLSTM networks for spatiotemporal prediction of active fires in the Cerrado, integrating data from multiple public sources. More specifically, we aim to: (i) construct a spatiotemporal dataset that combines BDQueimadas active-fire detections with meteorological, vegetation, topographic, and infrastructure variables on a regular grid over the biome; (ii) formulate active-fire ignition prediction as a spatiotemporal forecasting task using ConvLSTM; and (iii) design an experimental protocol based on metrics appropriate for highly imbalanced classification, with a view to future development of early-warning systems for Cerrado fire risk.

This article is structured as follows. Section˜2 reviews related work on the Cerrado fire regime, wildfire prediction models, and deep learning for environmental time series. Section˜3 presents the study area, datasets, and modeling framework. Section˜4 describes code and data availability. Section˜5 reports results and discussion, and Section˜6 concludes with main findings and future directions.

This study relies exclusively on publicly available remote-sensing and geospatial datasets and does not involve human or animal subjects, personal data, or interventions; therefore, no specific ethics-committee approval was required.

2 Related Work

The studies reviewed in this section were selected through a structured literature search conducted between 2024 and 2025 using Google Scholar and Scopus. Search strings included combinations of “wildfire prediction”, “fire ignition modeling”, “ConvLSTM”, “spatiotemporal deep learning”, and “Cerrado”. Priority was given to peer-reviewed studies published between 2015 and 2024.

In this section, we review related studies along three main strands: (i) work on the Cerrado, its fire regime, and ecological impacts; (ii) models for wildfire prediction using machine and deep learning; and (iii) applications of deep learning to environmental time series, with an emphasis on spatiotemporal architectures such as ConvLSTM.

2.1 Cerrado, fire regime and ecological impacts

The Cerrado exhibits pronounced spatial heterogeneity in climate, soils, topography, vegetation physiognomies, and land-use patterns. Proposals for subdividing the biome into ecoregions provide a spatial framework for assessing and prioritizing conservation actions, linking different combinations of biophysical factors to varying levels of threat and protection [Sano *et al*., 2019]. This spatial structure is highly relevant for fire-prediction models, as environmental and anthropogenic drivers vary across ecoregions and affect both ignition probability and fire spread.

From an ecological standpoint, fire has a structuring role in the Cerrado. Although many species display fire-related adaptations, such as thick bark and subterranean buds, these traits are linked to particular regimes of frequency and intensity. Durigan and Ratter [Durigan and Ratter, 2016] argue that policies of total fire suppression, when combined with uncontrolled anthropogenic ignitions, generate fire regimes that are incompatible with the biome’s evolutionary history, contributing to biodiversity loss and the replacement of native vegetation by impoverished pasture formations.

Remote-sensing time series further reinforce these concerns. Santana et al. [Santana *et al*., 2020] show that high fire recurrence in the Cerrado and adjacent forests is associated with declines in vegetation indices and primary productivity, as well as increases in land-surface temperature, indicating progressive degradation. Other approaches, such as the stochastic model for simulating Cerrado wildfires based on cellular automata [Ferreira *et al*., 2023], highlight the combined influence of vegetation physiognomies, moisture, and wind on fire spread, underscoring the spatiotemporal complexity of the phenomenon in the biome.

2.2 Wildfire prediction using machine and deep learning

Recent work on wildfire prediction encompasses a wide range of supervised machine-learning approaches. The bibliometric analysis in Pan and Zhang [2023] documents a marked increase in publications on fire modeling and prediction from

the mid-2000s onward, with a concentration of studies in countries such as the United States, Canada, and China and in the Mediterranean region. The review highlights the predominance of tree-based algorithms (Random Forest, Gradient Boosting), support vector machines (SVM), and neural networks, as well as the more recent emergence of deep architectures for spatiotemporal data.

Sharma and Khanal [Sharma and Khanal, 2024] propose a spatial framework for wildfire prediction in South Carolina, integrating fire records with meteorological, topographic, vegetation, land-use, and infrastructure variables. The authors compare several models—including decision trees, random forests, logistic regression, SVM, artificial neural networks, and convolutional networks—and use feature importance and correlation analysis to combine geospatial layers into danger maps. Results with accuracy above 80% suggest that ML approaches can capture complex relationships between environmental variables and fire occurrence.

Alongside occurrence prediction, a parallel line of research addresses fire and smoke detection in imagery. Deep convolutional networks pre-trained on large image datasets, combined with transfer learning and *learning without forgetting* strategies, have been applied successfully to the classification of images into fire, smoke, and background categories [Sathishkumar *et al.*, 2023]. Other studies explore deep-learning architectures to detect fire in RGB images and video data [Vamsi *et al.*, 2023]. Although these works focus on detection rather than the prediction of future active fires, they underscore the potential of deep architectures for fire-related problems, particularly in scenarios with large volumes of visual data.

2.3 Deep learning for environmental time series

Deep-learning models have become an important tool for forecasting environmental time series, in which variables such as precipitation, temperature, and air quality exhibit complex dependencies in time and space. Shi et al. [Shi *et al.*, 2015] introduced the ConvLSTM network in a precipitation *nowcasting* setting, showing improvements over conventional LSTM models and operational baselines. Later applications, such as DeepRain [Kim *et al.*, 2017], reinforce the ability of ConvLSTM networks to capture spatiotemporal patterns in multichannel meteorological data.

Hybrid CNN–LSTM and ConvLSTM architectures have also been employed to predict concentrations of atmospheric pollutants such as $PM_{2.5}$, integrating time series of meteorological variables, emission estimates, and spatial information [Putri *et al.*, 2024]. From a broader perspective, reviews of deep learning for time-series forecasting emphasize that, although recurrent, convolutional, and attention-based architectures have achieved promising performance, important challenges remain, including non-stationarity, class imbalance, scarcity of labels, and interpretability [Casolaro *et al.*, 2023].

By framing the prediction of active fires as a short-term spatiotemporal forecasting problem, the present study builds on these methodological developments, and in particular on ConvLSTM architectures, to estimate from sequences of environmental and anthropogenic maps the probability of future fire occurrence in each grid cell of the Cerrado biome.

3 Materials and Methods

3.1 Study area, analysis period and spatial grid

The study focuses on the Brazilian Cerrado biome, as delineated by the official biome map of the Brazilian Institute of Geography and Statistics (IBGE) [IBGE, 2019]. All datasets were clipped to this polygon, thereby restricting the analysis to the Cerrado domain.

The raw data used in this work span the period from 2014 to 2024, allowing us to capture both seasonal patterns and interannual variability in fire occurrence. To bring the different data sources into a common framework, we adopt a regular grid with an approximate spatial resolution of 5 km, onto which all environmental variables and the target variable are aggregated. Each grid cell is identified by an index pair (i, j), corresponding to an approximate geographic location within the biome.[1] Due to computational and memory limitations, however, the complete 5 km feature tensors used in the ConvLSTM experiments reported here were generated only for two representative years, 2014 and 2017. The model is trained on windows extracted from 2014 and evaluated on an independent test set based on 2017.

The choice of 2014 for training and 2017 for testing follows a realistic forecasting rationale, in which models are trained on historical data and applied to future periods. Constructing full 5 km spatiotemporal tensors for the entire 2014–2024 period would require memory resources beyond those available in the Google Colab environment used in this prototype, given the dimensionality of tensors of size $T \times H \times W \times C$. Therefore, two representative years were selected to ensure computational feasibility while preserving temporal independence between training and testing.

3.2 Active fire data (BDQueimadas/INPE)

Active fire detections used in this study were obtained from the BDQueimadas database maintained by the National Institute for Space Research (INPE) [INPE, 2023]. BDQueimadas consolidates active-fire detections from different orbital sensors aboard multiple satellites, processed and made available as a historical database of thermal anomalies.

Each record includes, among other attributes, the date and time of detection, the latitude and longitude of the pixel, and the satellite and sensor used. In the first step, fire points were filtered using the Cerrado biome polygon. The records were then aggregated to the regular $\sim$5 km grid adopted in this study. For each cell (i, j) and each day t in the analysis period, we recorded whether at least one active fire was detected within that cell during the interval $[t, t + 1]$.

Based on this aggregation, we defined a binary presence variable:

$$y_{i,j,t} = \begin{cases} 1, & \text{if at least one active fire is detected in } (i,j) \\ & \textit{during } [t, t+1], \\ 0, & \text{otherwise.} \end{cases}$$

[1] The spatial resolution may be adjusted in future experiments without altering the general problem formulation.

To focus on new ignitions rather than persistent burning, we derived a second label representing ignition events. For each cell (i, j) and day $t > 1$ we define

$$z_{i,j,t} = \begin{cases} 1, & \text{if } y_{i,j,t} = 1 \text{ and } y_{i,j,t-1} = 0, \\ 0, & \text{otherwise,} \end{cases}$$

so that $z_{i,j,t} = 1$ marks the first day on which fire appears after at least one day without fire in that cell. The ConvLSTM experiments reported in this paper use this ignition label at a one-day prediction horizon.

3.3 Meteorological data

Surface meteorological variables were obtained from stations of the National Institute of Meteorology (INMET) located within the Cerrado, accessed via the BDMEP (*Banco de Dados Meteorológicos para Ensino e Pesquisa*) database [INMET, 2025]. The variables considered include, among others, air temperature, relative humidity, accumulated precipitation, wind speed and direction, and surface pressure. Hourly or daily series from these stations underwent basic quality control and were interpolated to the study grid using spatial interpolation techniques appropriate to the problem scale.

In addition, we employed the ERA5-Land reanalysis dataset, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and accessed through Google Earth Engine [Muñoz-Sabater *et al*., 2021; Gorelick *et al*., 2017]. ERA5-Land provides hourly fields of meteorological variables at high spatial resolution, such as air temperature, specific humidity, precipitation, and surface fluxes. These fields were resampled to the Cerrado grid and aggregated to the daily resolution adopted for the fire data.

3.3.1 Selection of INMET stations within the Cerrado

Catalogs of automatic and conventional INMET stations, provided as CSV files, were first harmonized into a common tabular format. We addressed differences in column naming across files and converted coordinate fields, originally stored with comma decimal separators, into floating-point values. Each station was then represented as a georeferenced point in a geographic coordinate system (CRS EPSG:4326).

The spatial extent of the Cerrado was obtained from an IBGE shapefile containing the official biome polygon, reprojected when necessary to match the station coordinate reference system. A spatial join was performed between the set of station points and the Cerrado polygon using the *within* predicate, thereby selecting only stations located inside the official biome boundary.

This procedure yielded two filtered subsets: automatic stations in the Cerrado and conventional stations in the Cerrado. The resulting tables were exported as CSV files without geometry columns, to be used in subsequent steps of interpolating meteorological fields to the regular grid adopted in this study.

3.4 Land cover, vegetation and topography

Land-use and land-cover information was obtained from Collection 10 of MapBiomas Brasil, which provides annual national-scale maps of land use and cover [MapBiomas Project, 2025; Souza *et al*., 2020]. Classes of interest include different natural Cerrado formations, agricultural areas, pastures, urban areas, and water bodies. For each year in the study period, the land-cover map was clipped to the biome and resampled to the regular grid, allowing each cell (i, j) to be associated with a dominant land-use class or with fractions of different classes, depending on the chosen representation.

Beyond the categorical classes, we derived continuous variables from vegetation indices and land-surface temperature. The Normalized Difference Vegetation Index (NDVI) was obtained from the MODIS MOD13A2 product, which provides 16-day composites at 1 km spatial resolution [Didan, 2021]. Land-surface temperature (LST) was obtained from the MODIS MOD11A1 product, with a nominal spatial resolution of 1 km and daily frequency [Wan *et al*., 2021]. Both products were accessed through Google Earth Engine [Gorelick *et al*., 2017], clipped to the Cerrado, resampled to the study grid (5 km), and aggregated to the temporal resolution adopted here (e.g., daily means or maxima).

Topographic variables were derived from the Shuttle Radar Topography Mission (SRTM) digital elevation model (DEM) [Farr *et al*., 2007]. From the DEM we computed mean elevation per grid cell, slope, and aspect, with the latter decomposed into "northerness" and "easternness" components. These variables aim to capture orographic and exposure effects that may influence local microclimate, fuel distribution, and fire spread.

3.5 Infrastructure, anthropogenic use and water bodies

Variables related to infrastructure and anthropogenic pressure were derived from vector layers of roads, urban areas, and associated infrastructure available from MapBiomas [MapBiomas Project, 2025; Souza *et al*., 2020] and other compatible public datasets. From these layers we computed, for each cell (i, j), the minimum Euclidean distance to the nearest road segment and to the main urban centers.

Similarly, hydrographic layers were used to derive distance to surface water bodies (rivers, lakes, and reservoirs). Distance-to-infrastructure and distance-to-water variables seek to represent both the facilitation of ignition by human activities in more accessible areas and potential modulation of fire spread associated with proximity to urban areas and water resources.

3.5.1 Road proximity maps

To more explicitly explore the relationship between road accessibility and fire risk in the Cerrado, we constructed heat maps of road proximity on the same $\sim$5 km grid used in the model. Based on a consolidated road network, we computed, for each cell (i, j), the minimum distance to the nearest road segment, including different road types (paved highways, unpaved roads, and secondary roads), and rescaled this field into ranges suitable for modeling.

These proximity maps are used both as continuous input variables (e.g., normalized distance) and as a basis for defining risk categories associated with distance bands. Figures 1 and 2 present two examples of road-distance heat maps generated for the biome, highlighting regions located closest to the road network. Shorter distances to roads (warmer colors)

are concentrated along major highway corridors, while longer distances (cooler colors) occur in more remote regions, underscoring areas where anthropogenic ignition pressure is likely to be stronger.

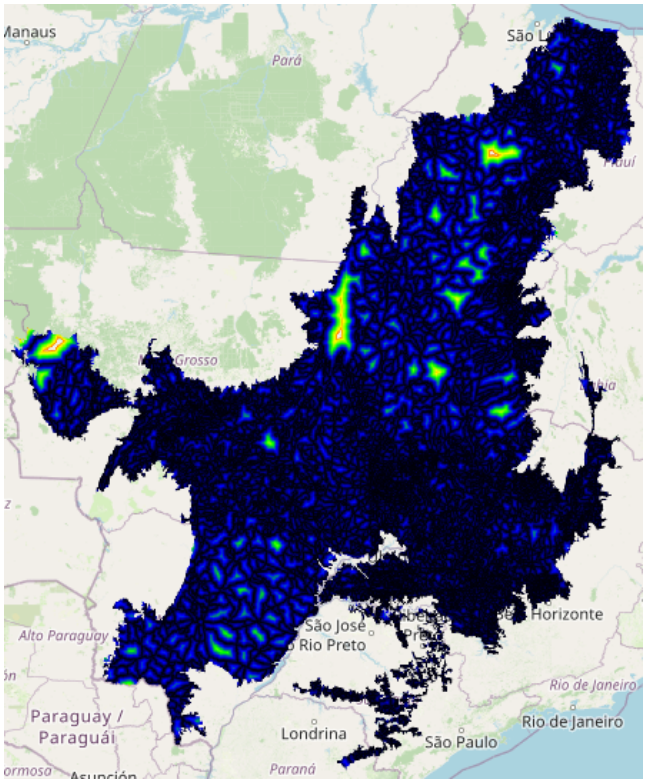

Figure 1. Road-distance heat map for the Cerrado biome, showing shorter distances to the road network (warmer colors) concentrated along major highway corridors and longer distances (cooler colors) in more remote regions.

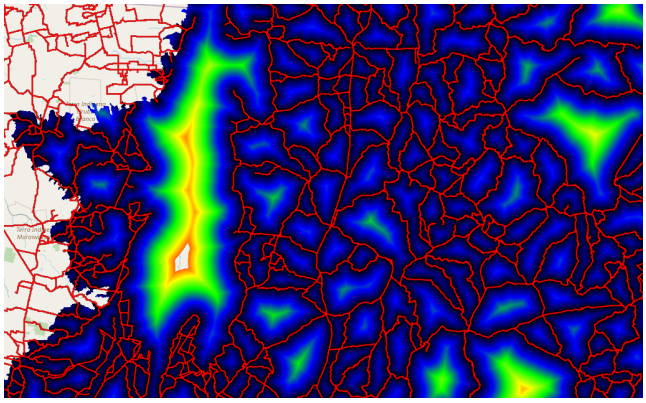

Figure 2. Zoomed road-distance heat map highlighting fine-scale variation in proximity to major highways and local road networks in a portion of the Cerrado.

3.6 Construction of the spatiotemporal dataset

To make the problem tractable for deep learning, the data were organized into a regular spatiotemporal structure. We adopted the daily scale as the basic time unit: for each day t, each continuous variable was aggregated from its intra-daily observations using simple statistics such as mean, maximum, minimum, or sum, depending on its nature. For instance, precipitation was accumulated over the day, whereas temperature and relative humidity were represented by daily averages or extremes.

At an intermediate stage, variables were first computed on a 1 km grid and then aggregated to 5 km by block coarsening, so that each 5 km cell combines information from the underlying 1 km cells. Static variables (e.g., elevation, slope, distances, land-use classes) were kept constant in time, while dynamic variables (meteorology, NDVI, LST) were stored as daily maps. After deriving the ignition label $z_{i,j,t}$ described earlier, we constructed for each day an ignition map on the 5 km grid.

For the ConvLSTM experiments, the input to the model is built from sliding windows of length T, containing the recent history of environmental maps. In this prototype we consider a window of $T = 14$ days, so that each training sample comprises a spatiotemporal block

$$\mathbf{X}_{t-T+1:t} \in \mathbb{R}^{T\times H\times W\times C},$$

where H and W denote the height and width of the spatial grid and C is the number of channels (environmental variables). The list of channels includes, among others, NDVI, land-surface temperature, accumulated precipitation, air temperature and humidity, wind components, elevation, slope, aspect, land-use class (encoded as binary channels), and distances to roads, urban areas, and water bodies. Each channel was normalized appropriately (for example, standardization by mean and standard deviation or rescaling to the $[0, 1]$ interval) based on the training set.

The target associated with each window is the ignition map $z_{i,j,t+1}$ for the day immediately following the observed history. Although the primary focus is on daily prediction (one-day horizon), the same framework can be extended to multi-day horizons in future work.

3.7 ConvLSTM architecture and training configuration

The proposed model is based on 2D ConvLSTM layers, which are capable of simultaneously capturing spatial and temporal dependencies in sequences of maps [Shi *et al.*, 2015]. In the configuration used for the experiments reported here, we employ two stacked ConvLSTM layers, each with 64 filters of size 3×3. The last hidden state is passed to a 2D convolutional layer with a 1×1 kernel and sigmoid activation that outputs a probability map $\hat{z}_{i,j} \in [0, 1]$ for each grid cell.

Let θ denote the model parameters and $p_{i,j,t}$ the predicted probability of ignition for cell (i, j) at time t. We adopt a binary cross-entropy loss with a positive-class weight to mitigate the impact of class imbalance:

$$\mathcal{L}(\theta) = -\frac{1}{N}\sum_{i,j,t}\left[w_+ z_{i,j,t}\log p_{i,j,t} + (1 - z_{i,j,t})\log(1 - p_{i,j,t})\right],$$

where N is the number of training examples and w_+ is a scalar weight proportional to the ratio between negative and positive pixels in the training set. In practice we implement this loss using the `pos_weight` parameter of `BCEWithLogitsLoss` in PyTorch, so that the network is trained directly on logits.

Training is carried out with the Adam optimizer, using a learning rate of 10^{-3} and mini-batches of 4 windows. In this prototype we train for three epochs, which was sufficient to

stabilize validation performance while keeping computational demands compatible with the available hardware (Google Colab environment). More extensive hyperparameter searches, including different depths, numbers of filters, and regularization schemes, are left for future work.

3.8 Evaluation procedures

The spatiotemporal dataset was partitioned into training, validation, and test subsets preserving temporal integrity. Windows ending in 2014 were used for training and validation (with an 80/20 split at the window level), while windows ending in 2017 formed an independent test set. This design mimics a forecasting scenario in which a model trained on an earlier year is applied to a different year with distinct meteorological conditions.

Because this is a strongly imbalanced problem, in which most grid cells do not experience ignitions on most days, evaluation does not rely on overall accuracy but instead on a set of more informative metrics. The main metric adopted is the F1-score, defined as the harmonic mean of precision and recall, which summarizes the trade-off between false positives and false negatives.

In addition to the F1-score, we compute the following metrics:

- **Precision**: the proportion of predicted positives that actually correspond to cell–days with ignitions;
- **Recall (sensitivity)**: the proportion of cell–days with ignitions that are correctly identified by the model;
- **Area under the ROC curve (AUC-ROC)**: measures the model's ability to separate cell–days with and without ignitions across different decision thresholds;
- **Area under the precision–recall curve (AUC-PR)**: particularly relevant in imbalanced settings, as it emphasizes performance on the minority class;
- **Intersection-over-Union (IoU)**: computed from binary prediction and observation maps, quantifying the spatial overlap between cells with predicted and observed ignitions.

The use of precision, recall, F1-score, and IoU follows established evaluation practices for imbalanced classification problems [Powers, 2011].

We first evaluate ConvLSTM outputs as continuous probabilities using ROC and precision–recall curves. We then study two derived decision schemes: (i) global thresholds on $p_{i,j,t}$ (e.g., 0.1, 0.2, 0.3, 0.5), and (ii) a top-k strategy in which only the k pixels with highest predicted risk in each daily map are flagged as positives, for different percentages of the grid (1%, 2%, 5%). The first scheme approximates conventional binary classification, while the second emulates an operational setting in which only a small fraction of cells can be monitored or inspected in detail.

4 Code and data availability

The Python code, notebooks, and main processed datasets used in this study are organized in a public Google Drive[2].

[2]Accessible at: `https://drive.google.com/drive/folders/1TNlziLPCC6vMf0PtD4YhKuuSBxJ8TzHe?usp=drive_link`

The directory is structured as follows. The folder `data/raw` contains raw and intermediate data exported from Google Earth Engine and other sources, including BDQueimadas fire records and intermediate station catalogs. The subfolder `INMET_Cerrado` holds station catalogs and subsets of INMET meteorological stations alongside the Cerrado biome shapefile. The `notebooks` directory concentrates Jupyter notebooks for data acquisition, preprocessing, and spatiotemporal aggregation, organized into a main pipeline for building the 5 km feature and label cubes, notebooks dedicated to static distance layers (roads, urban areas, water bodies) generated with Google Earth Engine, and legacy or teaching notebooks. This organization is intended to facilitate reproducibility and reuse of the proposed approach in future studies.

5 Results and Discussion

In this section, we present preliminary results obtained with the ConvLSTM prototype trained on 2014 and evaluated on 2017, focusing on one-day-ahead ignition prediction. The aim is not to provide a definitive operational model, but rather to quantify what level of pixel-scale skill can be achieved with a relatively simple spatiotemporal architecture under extreme class imbalance.

5.1 Dataset characteristics and class imbalance

As described in the previous sections, the fire dataset is constructed by aggregating BDQueimadas active-fire detections and environmental predictors onto a regular 5 km grid over the Cerrado. In principle, the underlying satellite and meteorological sources cover a ten-year period (2014–2024), and all years were downloaded and preprocessed at the point and 1 km scales. However, due to memory and runtime constraints, the construction of full 5 km feature cubes was only feasible for two representative years: 2014 (used for training and validation) and 2017 (used as an independent test year). The results reported in this paper therefore correspond to a reduced but internally consistent subset of the planned spatiotemporal domain.

On the daily 5 km grid, ignitions are extremely rare. In the 2017 test set, only about 2.8×10^{-4} of the cell–day instances correspond to ignition events, i.e., fewer than 0.03% of grid cells exhibit fire on a given day. The 2014 training year displays a very similar order of magnitude, leading to an effective negative-to-positive ratio of roughly $3{,}500{:}1$. This degree of imbalance is typical of ignition-level fire modeling and has important implications for model design and evaluation: naive baselines that always predict "no fire" achieve very high accuracy, and many conventional metrics become uninformative.

To mitigate these issues, we explicitly cast the problem as ignition prediction (transition from no fire to fire) rather than generic fire occurrence, use a weighted loss function during training, and rely on metrics that are more appropriate for rare events, such as the area under the ROC curve (AUC-ROC) and top-k hit rates. Throughout the analysis, all results are reported at the daily 5 km scale for the 2017 test year, with models trained and tuned on 2014 data under the same spatial

resolution.

5.2 Overall model performance

We first evaluate one-day-ahead ignition prediction on the 2017 test year, using models trained and validated on 2014. Even under the severe class imbalance described above, the proposed ConvLSTM model with a 14-day input window (two stacked ConvLSTM layers with 64 filters each and class-weighted binary cross-entropy loss) achieved an AUC-ROC of 0.89 on the independent test set, indicating a strong ability to rank cells by relative ignition risk.

To obtain a single operating point in the probability space, we scan decision thresholds on the test set and compute the corresponding precision, recall, and F1-score. The best trade-off between precision and recall, measured by the F1-score, occurs at a probability threshold of 0.15. At this threshold, the model attains 0.48% precision, 4.45% recall, and an F1-score of 0.0073. While these absolute values may appear small, they are more than one order of magnitude higher than those of a random classifier under the same prevalence, reflecting genuine predictive skill in an extremely unfavorable class-imbalance regime.

It is important to emphasize that, in such a rare-event setting, pixel-level F1 is heavily penalized by the very large number of negative cells that must be flagged in order to capture a modest fraction of ignitions. For this reason, threshold-independent or ranking-based metrics—such as AUC-ROC, precision–recall curves, and hit rates in the top-k highest-risk cells—provide a more informative view of model performance, especially from an operational standpoint.

Although additional machine-learning baselines were not implemented in this initial prototype due to computational constraints, previous studies have reported strong performance of tree-based models such as Random Forest and Gradient Boosting for wildfire prediction tasks [Pan and Zhang, 2023; Sharma and Khanal, 2024]. However, many of these approaches rely on tabular representations or coarser spatial units, whereas the present study formulates ignition prediction explicitly as a spatiotemporal forecasting problem on a regular grid. Therefore, the reported AUC-ROC of 0.89 suggests that ConvLSTM provides competitive ranking performance under extreme class imbalance and fine spatial resolution.

5.3 Top-k analysis

From an operational point of view, environmental agencies are often less interested in classifying every grid cell than in prioritizing a limited subset of locations for intensified monitoring or preventive action. In this context, ranking cells by predicted ignition probability can be more informative than applying a single global threshold, since only a small fraction of the landscape can be inspected on any given day.

To explore this perspective, we evaluate a top-k decision rule in which, for each daily probability map in the 2017 test year, only the k cells with highest predicted risk are flagged as positives. We then compute the fraction of all observed ignitions that fall within this restricted set. Results show that the top 1% highest-risk cells already contain about 12% of all observed ignitions in 2017. The top 2% contain 22% of ignitions, and the top 5% contain 45% of ignitions.

In other words, the ConvLSTM is able to concentrate nearly half of the observed fires in only 5% of the grid cells, despite the underlying event prevalence being below 0.03%. This behavior indicates a genuine ability to prioritize regions for monitoring and enforcement, and suggests that probability maps generated by the model could be useful as an additional layer in early-warning systems that need to allocate limited resources over large areas.

5.4 Spatial and temporal patterns

Qualitative inspection of the predicted probability maps suggests that the model tends to assign higher risk to portions of the Cerrado with greater historical fire recurrence and higher levels of human activity, such as agricultural and pasture areas. Transition zones between forest and savanna formations generally exhibit intermediate probabilities, consistent with their frontier position in terms of both vegetation structure and land use.

Temporally, the highest predicted probabilities occur predominantly during the dry season and the transition from dry to wet conditions, when low relative humidity, accumulated water deficit, and intensification of agricultural activities combine to increase ignition risk. This seasonal pattern is consistent with the fire climatology described for the Cerrado in previous studies [Durigan and Ratter, 2016; Santana *et al.*, 2020].

Typical failure cases include isolated ignitions in areas with low historical recurrence and events associated with very localized anthropogenic ignitions under only moderately favorable environmental conditions. These patterns are expected given the coarse spatial resolution of 5 km and the absence of explicit information about fine-scale human activities (e.g., burning permits, local management practices).

5.5 Implications and limitations

The obtained results clearly illustrate the difficulty of predicting ignitions in an extremely rare-event setting: fewer than 0.03% of grid cells exhibit fire on a given day in the 5 km daily grid. In this context, it is expected that precision remains very low even at tuned probability thresholds, since any attempt to increase recall requires flagging a large number of "suspicious" cells that will not burn. For this reason, ranking-based metrics such as the AUC-ROC (0.89 in our experiments) and the top-k analysis are more informative than accuracy or F1 alone. The ConvLSTM model is able to concentrate about 45% of the observed ignitions within only 5% of the highest-risk cells, which indicates a genuine ability to prioritize regions for monitoring despite the very low event prevalence.

From an applied perspective, these results suggest that daily ignition-probability maps could serve as a complementary information layer for environmental agencies and civil protection services. Even though performance at the individual-cell level is far from perfect, aggregating probabilities over larger spatial units (for example, microregions) and temporal windows (such as ten-day periods) tends to produce more stable patterns, which may be useful for planning prevention, surveillance, and resource allocation.

There are, however, important limitations that must be acknowledged. First, although satellite and meteorological data were collected for the period 2014–2024, the full 5 km

feature cubes could only be built for two years (2014 and 2017) due to memory and computational constraints. As a consequence, the ConvLSTM was trained and evaluated on a reduced temporal subset, and potential interannual variability in fire–climate relationships is not fully explored. Second, the ignition labels are derived from satellite thermal-anomaly detections and do not perfectly distinguish new ignitions from the spread of existing fires, introducing label noise that likely underestimates the true predictive potential of the model.

These factors suggest that the reported numbers should be interpreted as a conservative lower bound on performance. Future work will focus on extending the feature cubes to additional years, incorporating extra predictors such as recent fire history, and refining the definition of ignition labels. In parallel, systematic comparisons with simpler tabular machine-learning models and with alternative deep architectures (for example, attention-based spatiotemporal networks) will help clarify the cost–benefit trade-offs of ConvLSTM in operational wildfire early-warning settings.

6 Conclusion

This article presents the design and first results of an approach for spatiotemporal prediction of active fire ignitions in the Brazilian Cerrado biome, based on ConvLSTM networks and on the integration of environmental and anthropogenic data from multiple sources. Framing the problem as a grid-based forecasting task allows us to draw on recent advances in deep learning for environmental time series while incorporating remote-sensing, topographic, infrastructure, and land-use information.

Methodologically, the study explores the use of ConvLSTM networks to predict daily ignition events in each grid cell, treating the problem as a highly imbalanced binary classification task. The evaluation protocol, based on metrics such as AUC-ROC, precision, recall, F1-score, and top-k analyses, is designed to capture more informatively the model's performance on the minority class and the spatial concentration of risk.

On an independent 2017 test set with an empirical positive rate of approximately 0.03%, the ConvLSTM prototype attained an AUC-ROC of about 0.89, indicating substantial skill in ranking pixels by ignition probability. Nevertheless, pixel-level F1-scores remained low under both global thresholds and top-k decision schemes, underscoring the intrinsic difficulty of forecasting rare ignitions and the importance of considering spatial and temporal aggregation when designing operational products.

As future work, we plan to: (i) extend the 5 km feature and label cubes to the full 2014–2024 period, enabling more robust temporal generalization tests; (ii) refine the ConvLSTM architecture and hyperparameters, exploring deeper networks, alternative loss functions, and regularization strategies; (iii) systematically compare ConvLSTM with tabular machine-learning models and alternative deep architectures, such as attention-based models for spatiotemporal series; and (iv) incorporate model-interpretation techniques, such as SHAP values, to investigate the relative contribution of meteorological, vegetation, infrastructure, and topographic variables to predicted risk in different regions of the biome. From an applied perspective, we intend to evaluate, in collaboration with environmental agencies, the potential of spatially aggregated ignition-risk maps as a complementary tool to existing satellite-based monitoring systems, thereby contributing to fire-prevention and control policies in the Cerrado and other Brazilian biomes.

Declarations

Authors' Contributions

Guilherme Gomes de Araújo Vieira led the work, contributing to the overall conception of the study, methodological design, data processing, and writing of the manuscript. Igor Marques Santos Magalhães, Lucas Rocha de Santana, Elvis Miranda Teixeira, and Kauã Maciel Veit contributed to data acquisition and preprocessing, implementation of the spatiotemporal dataset, literature review, and writing – review and editing of the manuscript. Jeovan Assis da Silva acted as the supervising professor, contributing to the conceptualization, refinement of the research problem, methodological guidance, critical review of the text, and validation of the scientific content. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The main datasets and codes used in this study are based on publicly available sources (BDQueimadas/INPE, INMET, ERA5-Land, MODIS, MapBiomas, and SRTM). Processed data and scripts will be made available upon request to the corresponding author.

Further relevant information

This study relies exclusively on publicly available remote-sensing and geospatial datasets and does not involve human, animal subjects or personal data.

References

Casolaro, A. *et al*. (2023). Deep learning for time series forecasting: Advances and open problems. *Information*, 14(11):598. DOI: 10.3390/info14110598.

Colli, G. and Vieira, C. (2020). Biodiversity and conservation of the cerrado: recent advances and old challenges. *Biodiversity and Conservation*, 29. DOI: 10.1007/s10531-020-01967-x.

Didan, K. (2021). Mod13a2 modis/terra vegetation indices.

Durigan, G. and Ratter, J. A. (2016). The need for a consistent fire policy for cerrado conservation. *Journal of Applied Ecology*, 53(1):11–15. DOI: 10.1111/1365-2664.12559.

Farr, T. G. *et al*. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, 45(2). DOI: 10.1029/2005RG000183.

Ferreira, H. F., Tinoco, C. R., Martins, L. G. A., and Oliveira, G. M. B. (2023). Stochastic model for wildfire simulation based on the characteristics of the brazilian cerrado. In *Artificial Intelligence and Soft Computing*, pages 487–496. Springer.

Gorelick, N. *et al*. (2017). Google earth engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27. DOI: 10.1016/j.rse.2017.06.031.

IBGE (2019). Brazilian biomes: map and vector dataset.

INMET (2025). Meteorological database for teaching and research (bdmep).

INPE (2023). Bdqueimadas: Brazilian fire database.

Kim, S. *et al*. (2017). Deeprain: Convlstm network for precipitation prediction using multichannel radar data.

MapBiomas Project (2025). Collection 10 of the annual land use land cover maps of brazil.

Muñoz-Sabater, J. *et al*. (2021). Era5-land: a state-of-the-art global reanalysis dataset for land applications. *Earth System Science Data*, 13:4349–4383. DOI: 10.5194/essd-13-4349-2021.

Pan, M. and Zhang, S. (2023). Visualization of prediction methods for wildfire modeling using citespace: A bibliometric analysis. *Atmosphere*, 14(6):1009. DOI: 10.3390/atmos14061009.

Powers, D. M. W. (2011). Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*. DOI: 10.9735/2229-3981.

Putri, T. H. *et al*. (2024). Fine-tuning of predictive models cnn-lstm and conv-lstm for nowcasting pm2.5 level. *IEEE Access*, 12:28988–29003. DOI: 10.1109/ACCESS.2024.3368034.

Sano, E. E., Rodrigues, A. A., Martins, E. S., Bettiol, G. M., Bustamante, M. M. C., Bezerra, A. S., Couto, A. F., Vasconcelos, V., Schüler, J., and Bolfe, E. L. (2019). Cerrado ecoregions: A spatial framework to assess and prioritize brazilian savanna environmental diversity for conservation. *Journal of Environmental Management*, 232:818–828. DOI: 10.1016/j.jenvman.2018.11.108.

Santana, N. C., Carvalho Júnior, O. A., Gomes, R. A. T., and Guimarães, R. F. (2020). Comparison of post-fire patterns in brazilian savanna and tropical forest from remote sensing time series. *ISPRS International Journal of Geo-Information*, 9(11):659. DOI: 10.3390/ijgi9110659.

Sathishkumar, V. E., Cho, J., Subramanian, M., and Naren, O. S. (2023). Forest fire and smoke detection using deep learning-based learning without forgetting. *Fire Ecology*, 19(1):9. DOI: 10.1186/s42408-022-00165-0.

Sharma, S. and Khanal, P. (2024). Forest fire prediction: A spatial machine learning and neural network approach. *Fire*, 7(6):205. DOI: 10.3390/fire7060205.

Shi, X. *et al*. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting.

Souza, C. M. *et al*. (2020). Reconstructing three decades of land use and land cover changes in brazilian biomes with landsat archive and earth engine. *Remote Sensing*, 12(17):2735. DOI: 10.3390/rs12172735.

Vamsi, P., Chakravarthy, D., and Neelish, I. (2023). Forest fire prediction using image processing through deep learning. *International Journal of Scientific Research in Engineering and Management*, 7. DOI: 10.55041/IJSREM18364.

Wan, Z., Hook, S., and Hulley, G. (2021). Mod11a1 modis/terra land surface temperature.