Wildfire Monitor: Real-Time Prediction and Interpretable Alerts from Fused Satellite and Meteorological Data

Luís Fabrício de Freitas Souza[1,5], **Guilherme F. B. Severiano**[2,5],
José Jerovane da C. Nascimento[3,5], **Pedro Hugo Ursulino Fernandes**[1,4,5],
Ícaro de Sousa Rodrigues[3,5], **Jesus Ossian Cunha Silva**[3],
Francisco Italo G. da Silva[3,4,5],
Osvaldo Soares Landim Junior[1,2,5]

[1] Universidade Federal do Cariri (UFCA)
Av. Tenente Raimundo Rocha, – Cidade Universitária,
Juazeiro do Norte-CE - Brazil

[2]Instituto Federal do Ceará – Fortaleza, CE – Brazil

[3]Universidade Federal do Ceará - (UFC) – Fortaleza, CE – Brazil

[4]Instituto Atlântico - Instituição de Ciência e Tecnologia

[5]Laboratório de Inovação em Sistemas de Inteligência Artificial – LISIA

fabricio.freitas@ufca.edu.br,
guilherme.freire.brilhante03@aluno.ifce.edu.br,
{jerovane, icaros1105}@alu.ufc.br,
pedro.hugo@aluno.ufca.edu.br, jesus.ossian@ufc.br,
{italoguilherme18.1,
osvaldo.jrjuniorlinhares}@gmail.com

***Abstract. Research Context**: Wildfires cause significant environmental, economic, and social damage, requiring intelligent systems capable of supporting prevention and mitigation actions. **Scientific and/or Practical Problem**: The increasing frequency of wildfires in Brazil and worldwide demands innovative technological approaches for real-time monitoring and accurate prediction of fire outbreaks using reliable and official data sources. **Proposed Solution and/or Analysis**: This paper presents the Cariri Wildfire Monitor, a system for mapping and monitoring wildfires through a web portal that integrates Artificial Intelligence (AI) and official datasets. The proposal includes a technological solution for continuous monitoring and a predictive module for fire hotspots. **Related IS Theory**: The system design is grounded in Information Systems (IS) theories related to data-driven decision support, predictive analytics, and socio-environmental information integration for sustainable management. **Research Method**: The system performs data fusion from INPE and INMET sources, applies regression models to predict wildfire hotspots, and integrates a decision layer with Large Language Models (LLMs) to produce auditable alerts and qualitative insights. **Summary of Results**: Among the results, the ExtraTrees regressor achieved* R^2=0,9843*. In the stratified audit of the LLM module with live threshold* $0{,}758$ *(proxy* $0{,}500$*;* N_{pop}=1000*;* M_{audit} *used* = 12*), we obtained Accuracy (w)* 0,7973±0,1165*, and p95 latencies between 4,75s and 5,95s. **Contributions and Impact to IS area**: The system demonstrates the potential of predictive* *computational models and LLMs in environmental monitoring, contributing to the development of intelligent, auditable, and sustainable Information Systems solutions for wildfire management.*

1. Introduction

The climate crisis, driven by current patterns of production and consumption, has become increasingly evident in daily life. In recent years, there has been a rise in heat waves and severe storms, resulting in wildfires, droughts, and the risk of displacement or extinction of regional flora and fauna, among other consequences of this global scenario [Nieuwenhuijsen 2024].

Recent data on wildfires have raised concerns among scientists. According to the [Weisse and Goldman 2023], the estimated loss of primary tropical forests in 2023 totaled 3,7 million hectares equivalent to nearly 10 soccer fields of forest lost per minute. This represents a 9% reduction compared to 2022. Also in 2023, this forest loss contributed approximately 2,4 gigatons of carbon dioxide emissions, an amount equivalent to half the annual fossil fuel emissions of the United States.

In 2023, Brazil recorded a 6% increase in the total area affected by wildfires compared to the previous year, totaling approximately 17,3 million hectares about 2% of the national territory. September and October were particularly severe, each with 4 million hectares burned [MapBiomas 2025]. In the state of Ceará, the Cariri region was the most impacted, leading wildfire occurrences in September 2023, with about 2,116 records. The cities of Juazeiro do Norte, Crato, and Barbalha faced dense smoke clouds, resulting in respiratory problems in the population and local evacuations due to the proximity of the flames [Povo 2023].

In this context, intelligent systems capable of predicting and detecting wildfire hotspots are of fundamental importance for environmental protection planning. These systems support better planning of actions, monitoring, and damage mitigation [Xu et al. 2021]. Several approaches have been proposed, from *deep learning* models [Park et al. 2022] to cellular automata [Murilo et al. 2024], regression-based forecasts [Peng et al. 2023], and fuzzy methods for early detection. However, most studies focus on a single algorithm or lack integration with real-time monitoring systems [da Costa Nascimento et al. 2025].

The adoption of Natural Language Processing (NLP) methods capable of interpreting large-scale data and translating them into logical, well-contextualized insights may represent a significant advancement in the study and understanding of environmental profiles in wildfire-affected areas. A well-trained NLP model, tailored to the regional context and data, can accurately identify key contributing factors to fires and suggest the most effective prevention and recovery measures. The model may also highlight indicators typically only recognized by experts and flag emergency thresholds in the dataset values that might otherwise be just another input for regression but can qualitatively represent critical circumstances [Mohnish et al. 2023].

This work presents the *Cariri Wildfire Monitor*, a system already implemented for the Cariri region in Ceará. The system performs data fusion between INPE hotspots and INMET meteorological variables, applies regression models for prediction, and integrates a decision module with LLMs to generate auditable alerts. Unlike previous works, our approach explicitly evaluates the reliability of LLMs with structured metrics and stratified auditing, ensuring robustness in critical scenarios.

The main contributions of this study are:

- Integration of INPE and INMET datasets through a fusion process aimed at wildfire prediction.
- Comparative study of regression methods with statistical validation, highlighting the performance of ensembles such as ExtraTrees.
- Deployment of an auditable LLM module that generates structured outputs, monitored for hallucination and reliability.
- Delivery of an operational system for intelligent wildfire monitoring in Cariri, Ceará.

2. Related Work

Wildfire prediction and monitoring have been approached through multiple modeling families, from classical statistics to modern machine learning, *deep learning*, simulation, and, more recently, decision support with Natural Language Processing (NLP) and Large Language Models (LLM) [Marques et al. 2024]. Early lines emphasized regression and time-series forecasting; recent works integrate remote sensing and meteorology to improve spatial and temporal sensitivity. Despite advances, much of the literature optimizes accuracy in isolated settings, with fewer efforts on real-time applicability, integration of heterogeneous sources, or auditability of decision layers.

Deep models have been explored for early detection and burned-area estimation from imagery, achieving strong spatial performance for emergency response [Park et al. 2022]. Other studies combine meteorological variables with remote sensing for hotspot prediction [Xu et al. 2021]. However, these approaches often demand large annotated datasets and can lack interpretability, complicating adoption by agencies that need transparent decision criteria.

Simulation-based methods such as Cellular Automata (CA) have been used to model fire spread and support prevention strategies [Murilo et al. 2024]. While useful to visualize propagation scenarios, CA is sensitive to initial/boundary conditions and can be computationally expensive over large areas.

Supervised regression and ensembles have also been applied to estimate spatiotemporal hotspot dynamics, with *ensemble learning* frequently improving robustness under heterogeneous environmental signals [Peng et al. 2023, Da Silva et al. 2019, de Andrades et al. 2019]. That said, many pipelines do not explicitly address incomplete/noisy data or provide mechanisms for online integration, hampering operational use.

Beyond vision-only pipelines, sensors and fuzzy-hybrid classifiers have been investigated. [Suklabaidya and Das 2023] evaluated machine learning algorithms for fire detection in smart-home contexts using IoT signals, comparing Support Vector Machine (SVM), Naïve Bayes, Decision Tree (DT), Random Forest (RF), and K-Nearest Neighbor (KNN), and reporting 84% accuracy; gains with hyperparameter optimization were noted as likely. [Umoh et al. 2022] combined interval Type-2 fuzzy logic preprocessing with normalization and principal component analysis for feature selection, training KNN, SVM, RF, Linear Discriminant Analysis (LDA), and CART. Reported Receiver Operating Characteristic (ROC) values approached 0.998 for SVM/RF, though additional tuning and broader metrics would strengthen conclusions.

Thermal imagery with fuzzy modeling has been tested in laboratory and field conditions. [Sousa et al. 2019] showed that dynamic temperature patterns can be characterized effectively; the 8-cluster configuration reached the highest Area Under the Curve (AUC) (0.9740), while a 3-cluster model offered improved interpretability with competitive performance. As usual with fuzzy systems, expert tuning remains important to sustain performance across environments [Severiano et al. 2024].

NLP and LLMs are beginning to appear as decision-support layers for wildfire response. [Zhang et al. 2024] proposed FireRobBrain, coupling a Knowledge Graph with an LLM to assist robots with stepwise decisions in dynamic scenes (864-sample evaluation). Their results suggest that combining structured knowledge and prompting can improve response quality. Nevertheless, current reports rarely include formal *reliability* audits (e.g., parse compliance, fallback rates, calibration/Brier, or latency), which are crucial in high-risk domains.

Time-series baselines based on ARIMA have been studied for Brazilian biomes using INPE/INMET variables; [Viganó et al. 2018] reported a 66.5% fit for occurrence prediction. We treat such numbers *contextually*: (i) ARIMA's univariate/low-dimensional paradigm differs from multivariate regression with derived climatic features, and (ii) cross-region comparisons (e.g., Pantanal vs. Cariri) are not directly fair due to geographic and climatic divergence. Accordingly, when external figures are cited, we refrain from claiming superiority and advocate fair comparisons on identical splits with chronological validation.

Positioning and gaps. In light of the above, the present work focuses on: (i) explicit fusion of heterogeneous official sources (INPE hotspots + INMET meteorology) into a reproducible feature space; (ii) a comparative regression study with statistical residual diagnostics; (iii) probability mapping of hotspot predictions for operational thresholds; and (iv) an auditable LLM decision layer with strict JSON outputs and stratified reliability metrics (weighted precision/recall/F1, ROC–AUC, Brier calibration, parse rate, fallback, and p95 latency). Rather than proposing a new learning algorithm, our contribution is an *operational*, end-to-end pipeline that emphasizes integration, interpretability, and measurable reliability for real-time monitoring dimensions that remain underrepresented in the literature.

3. Materials and Methods

This section describes the dataset, models, and evaluation methods used in the proposed system, combining regression analysis with an auditable LLM pipeline.

3.1. Dataset

Historical wildfire and meteorology data were obtained from two main sources: INPE (National Institute for Space Research) [Silva et al. 2024], responsible for wildfire hotspots, and INMET (National Institute of Meteorology) [Melo 2024], which provides climatic variables. A *DataFusion* process was applied to align fire records with meteorological observations by municipality and time step. Derived variables include the monthly difference in hotspots (Q) and normalized attributes using the mean (μ) and standard deviation (σ) of historical distributions. This integration produces a dataset with greater representativeness and robustness for regression modeling.

3.2. Regressors

We evaluated 13 regression methods, from linear models to ensembles: Linear Regression, Ridge, Lasso, ElasticNet, Bayesian Ridge, SVR, KNN, Decision Tree, Random Forest, Gradient Boosting, ExtraTrees, AdaBoost, and HistGradientBoosting. Decision Trees were included as a simple baseline, while ensembles such as RF and ExtraTrees exploit variance reduction to achieve higher accuracy. Hyperparameters were optimized with *GridSearchCV* [Mantovani et al. 2016], chosen due to the limited search space, in which exhaustive search is advantageous. For larger and more interdependent spaces, *RandomizedSearchCV* would be preferable.

3.3. LLM Orchestration

The decision layer employs LLMs via the OpenRouter API, with a *custom*, framework-free Python orchestrator designed for auditability and low latency. The runtime follows a three-stage graph Moderator→Risk→Decision with explicit contracts:

- Moderator — pre-validates ranges/units, attaches a compact *context bundle* (municipality, period, recent ΔQ, anomaly flags, regressor output and its mapped probability), injects a versioned JSON schema (`schema_version = "1.1"`). It also passes the current *risk policy* (live threshold 0.758, recall-prioritized rubric, and guardrail toggles).
- Risk — reasons over the bundle and emits a *draft* JSON with fields `alert` (boolean), `risk_probability` in $[0, 1]$, and `explanation` (concise, source-aware).
- Decision — validates against the schema, applies rule guards (e.g., edge cases near threshold or contradictory signals), optionally applies probability calibration (if enabled), and outputs the final JSON. All branches (success, repair, fallback) are logged for audit.

Prompt engineering and guardrails. We use a layered prompt design: (i) a strict *system* preamble ("reply *only* with JSON matching this schema", with field types/ranges and examples); (ii) a short *instruction* template with slot-filling ({`municipality`}, {`period`}, {`delta_Q`}, {`prob_mapped`}, {`risk_policy`}) to guarantee deterministic structure; and (iii) *few-shot* exemplars covering typical patterns and adversarial cases (e.g., high predicted risk with unusually high humidity; missing/flagged meteorology; sharp ΔQ with low baseline; dry-season amplification). We keep temperature low (e.g., 0.2), cap `max_tokens` to avoid verbosity, and set stop-sequences to prevent trailing text after JSON. A streaming JSON parser enforces schema while decoding; on violation we attempt a single *self-repair* prompt ("fix JSON only, do not re-reason"). Persistent failure triggers redundant-model fallback (Qwen↔LLaMA); if both fail, a rule-based template guarantees a minimal, valid JSON.

Architecture and reliability. The orchestrator is modular (pure Python, typed), with a router that ranks providers by recent *health* (parse rate, 429 incidence, mean/p95 latency). We employ speculative parallelism (fire primary and secondary with tight timeouts; accept first valid JSON) to keep p95 < 6s. Every call records: model id, latency, token usage, `parse_ok`, `fallback_used`, and guardrail triggers (e.g., UNIT_MISMATCH). Logs feed the stratified audit and support threshold tuning. We do not persist PII; analytics aggregate at municipality level. Full metrics are in Results.

3.4. Evaluation Metrics

We report metrics at three layers: (i) *regression* quality; (ii) *decision* quality after thresholding probabilities into alerts; and (iii) *operational* reliability for the LLM service. Chronological splits are respected end-to-end.

Regression and diagnostics - MAE, RMSE, MSE, MAPE, and R^2 quantify fit; Shapiro–Wilk (normality), paired t-test (bias vs. zero), and Pearson r (association) diagnose residuals. Headline numbers are computed on the chronological test split; validation folds are used only for selection.

Decision metrics and calibration - Counts are mapped to probability via a logistic link to enable cost-sensitive thresholds. We prioritize Recall (safety) and track Precision, F1, ROC–AUC, PR–AUC, and the Brier score for probability quality. Reliability diagrams and ECE guide optional post-hoc calibration (isotonic/Platt) fitted on validation only. The *live* operating point ($\tau = 0.758$) is chosen by expected-cost analysis (FN $>$ FP) and validated in the stratified audit.

Stratified audit and reporting - Municipality/month strata have different prevalences; we therefore compute *weighted* metrics and report the effective sample size (n_{eff}) to avoid overconfident intervals when a few strata dominate. Uncertainty bands (95% CI) are obtained via stratified bootstrap. We also include ablations comparing *rule-only* vs. *LLM-orchestrated* alerts to isolate the LLM contribution (precision/recall trade-offs and Brier improvements).

Operational (LLM) metrics and SLAs - The LLM layer is treated as a production component with explicit SLOs: JSON parse rate (first-pass schema compliance), fallback rate (redundant model or rules invoked), and latency (mean/p95). We also track 429/backoff incidence and guardrail violations. These indicators quantify reliability and responsiveness and are reported per-model in the Results.

4. Methodology

The proposed method consists of an eight-step *pipeline*, illustrated in Fig. 1, designed to integrate data acquisition, regression-based prediction, statistical validation, and interpretability via LLM in a complete wildfire monitoring system.

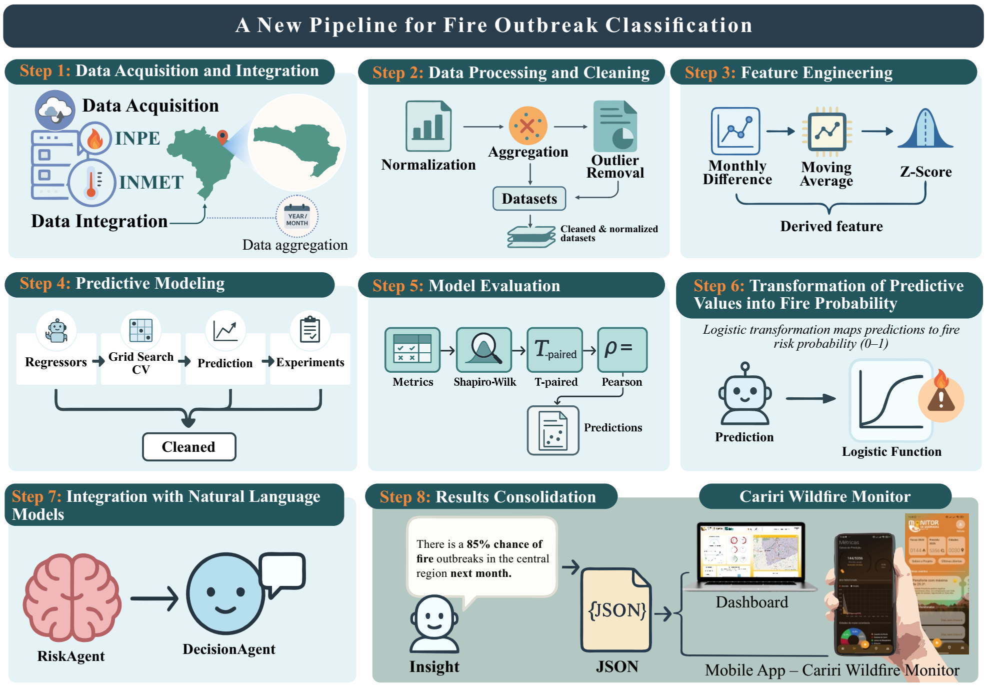
A New Pipeline for Fire Outbreak Classification
Step 1: Data Acquisition and Integration
Data Acquisition
INPE
INMET
Data Integration
YEAR/MONTH
Data aggregation
Step 2: Data Processing and Cleaning
Normalization
Aggregation
Outlier Removal
Datasets
Cleaned & normalized datasets
Step 3: Feature Engineering
Monthly Difference
Moving Average
Z-Score
Derived feature
Step 4: Predictive Modeling
Regressors
Grid Search CV
Prediction
Experiments
Cleaned
Step 5: Model Evaluation
Metrics
Shapiro-Wilk
T_{paired}
T-paired
$\rho =$
Pearson
Predictions
Step 6: Transformation of Predictive Values into Fire Probability
Logistic transformation maps predictions to fire risk probability (0–1)
Prediction
Logistic Function
Step 7: Integration with Natural Language Models
RiskAgent
DecisionAgent
Step 8: Results Consolidation
There is a 85% chance of fire outbreaks in the central region next month.
Insight
{JSON}
JSON
Cariri Wildfire Monitor
Dashboard
Mobile App – Cariri Wildfire Monitor

Figure 1. Proposed methodology divided into eight steps: Step 1. Data Acquisition and Integration; Step 2. Data Processing and Cleaning; Step 3. Feature Engineering; Step 4. Predictive Modeling; Step 5. Model Evaluation; Step 6. Transforming Predictive Values into Fire Probability; Step 7. LLM Orchestration and Decision Integration; Step 8. Results Consolidation.

Step 1: Data Acquisition and Integration

In this initial step, data are collected from two official and authorized sources: INPE, which provides historical records of wildfire hotspots, and INMET, which provides meteorological variables including precipitation, temperature, humidity, and wind speed. Acquisition is automated by Python scripts that first check the existence of local files to avoid unnecessary requests, reducing computational cost and server load. Downloaded files are often compressed, requiring systematic extraction and validation of expected content. By relying on official Brazilian institutions, the dataset ensures credibility and contextual relevance for Cariri, avoiding dependence on synthetic or global bases that may not capture local dynamics. Integration between sources is achieved via a shared temporal identifier (`year_month`), aligning fire incidents with corresponding climatic conditions at a monthly scale. This fusion is critical, as it creates a dataset that reflects both the physical occurrence of fires and their climatic context, enabling more reliable predictive modeling.

Step 2: Data Processing and Cleaning

After acquisition, the data undergo rigorous processing to ensure reliability. Dates are standardized in `datetime` format, ensuring consistent temporal indexing. Column headers from different sources are renamed under a unified scheme, reducing ambiguity in subsequent analyses. Inconsistencies such as missing values, duplicates, and anomalous outliers (e.g., extreme temperature spikes due to sensor defects) are detected and removed or imputed with temporal means. These operations are crucial because systematic errors, even small ones, can propagate and strongly distort regression results. The processed dataset is structured into tables in which hotspots are directly linked to climatic descriptors, facilitating subsequent aggregations. By imposing strict preprocessing, the pipeline ensures that modeling steps are based on high-quality inputs, addressing a frequent weakness noted by reviewers: dataset reliability.

Step 3: Feature Engineering

In this step, raw attributes are transformed into derived variables that capture temporal dynamics and statistics of fire occurrence. Monthly means are computed as:

$$\bar{x}_m = \frac{1}{N_m} \sum_{i=1}^{N_m} x_i,$$

where N_m is the number of daily measurements in month m. To capture temporal shifts, the monthly difference in hotspots is calculated as:

$$\Delta Q_m = Q_m - Q_{m-1}.$$

Long-term dependencies are incorporated via 12-month moving averages, while short-term climatic dynamics are modeled by 7-day moving averages:

$$\text{MA}_7(x)_d = \frac{1}{7} \sum_{i=d-6}^{d} x_i.$$

Finally, normalization is applied to all attributes:

$$x' = \frac{x - \mu}{\sigma},$$

where μ and σ are historical mean and standard deviation, respectively. This prevents variables on larger scales from dominating training. The combination of temporal aggregations, difference operators, moving averages, and normalization creates a feature space sensitive to short-term climatic anomalies (e.g., drought spikes) and long-term seasonal cycles. This step addresses criticisms from earlier versions regarding the justification of features.

Step 4: Predictive Modeling

With features prepared, regression models are trained to predict the number of hotspots. We tested a broad spectrum of regressors: linear (Linear, Ridge, Lasso, ElasticNet, BayesianRidge), kernel-based (SVR), instance-based (KNN), trees (Decision Tree), and ensembles (Random Forest, Gradient Boosting, ExtraTrees, AdaBoost, HistGradientBoosting). Decision Tree serves as an interpretable baseline, while ensembles provide robustness against variance and overfitting. Hyperparameters are optimized via *GridSearchCV*, an exhaustive choice suitable for the reduced search space, avoiding the stochasticity of *RandomizedSearchCV* and ensuring full exploration. Cross-validation is used to ensure performance is not an artifact of a single data split. The main performance measure is the coefficient of determination:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2}.$$

The broad comparative analysis allows us to identify not only the best model (ExtraTrees) but also to quantify the relative trade-offs between simple and complex regressors, addressing the question of why so many algorithms were tested.

Step 5: Model Evaluation

Performance is evaluated with MAE, RMSE, MAPE, and R^2. To enhance robustness, we apply hypothesis tests to residuals. Shapiro–Wilk checks normality, the paired *t*-test assesses whether residuals differ significantly from zero, and Pearson measures the linear relationship between predicted and observed. These tests verify not only accuracy but also bias and statistical consistency. This step addresses reviewers' comments for deeper validation and justification of statistical tests.

Step 6: Transformation into Fire Probability

Regression outputs (predicted number of hotspots) are not directly interpretable for decision-making. To address this, we map predictions into occurrence probabilities via a logistic function:

$$P(\text{fire}) = \frac{1}{1 + e^{-k(x-x_0)}},$$

where k controls the slope and x_0 is the inflection point, defined as the 75th percentile of predictions. Thus, values above the critical threshold translate into high probability, while low values indicate minimal risk. The logistic transformation improves interpretability for actors such as environmental agencies and firefighters, converting numerical outputs into probabilistic alerts.

Step 7: LLM Orchestration and Decision Integration

To complement predictions with contextual interpretability, we developed LLM orchestration implemented directly in Python, without external frameworks. The architecture follows the Moderator→Risk→Decision graph, integrated with the OpenRouter API. Two state-of-the-art models are employed in redundant variants (Qwen and LLaMA). All prompts enforce strict JSON with fields `alert`, `risk_probability`, and `explanation`. Operational reliability is measured via stratified audit (Section 5.3).

Step 8: Results Consolidation

The final step consolidates all outputs regression predictions, statistical validation, and LLM decisions into a unified JSON schema. This format facilitates integration with dashboards, monitoring systems, and the Cariri Wildfire Monitor mobile app. For example, a prediction of "85% fire risk" is directly exportable as a real-time alert to civil defense. Beyond practical usability, consolidation favors reproducibility: other researchers can re-run the pipeline and compare with the logs, ensuring transparency. This step clarifies that "results consolidation" is not an additional algorithmic phase but packaging for deployment, addressing comments about redundancy.

In summary, the methodology describes a complete operational pipeline that fuses heterogeneous data, applies validated regression models, integrates statistical and probabilistic transformations, and uses LLM orchestration for interpretability. Unlike fragmented works, the system is end-to-end, reproducible, and directly applicable to wildfire monitoring in Cariri.

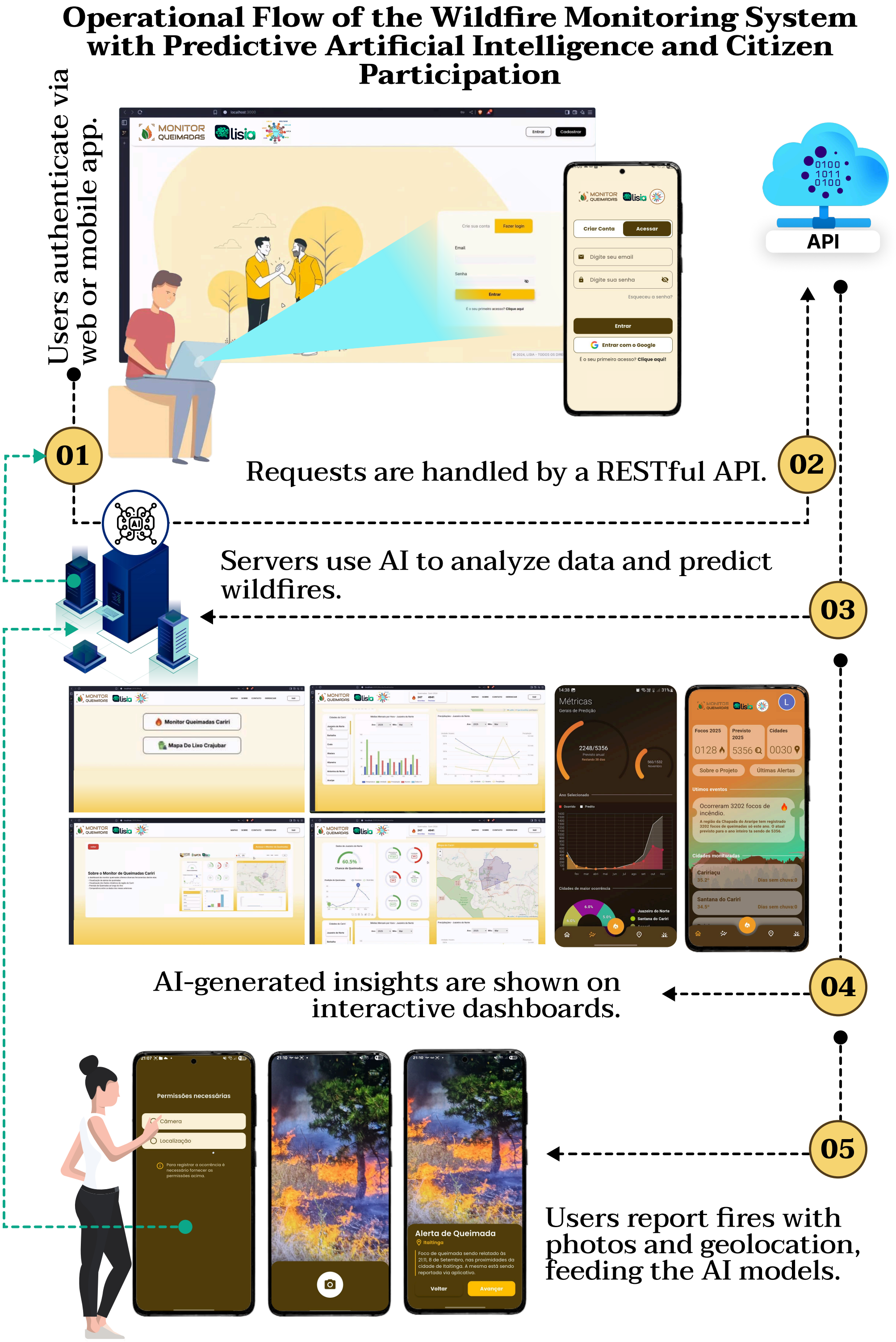
Operational Flow of the Wildfire Monitoring System with Predictive Artificial Intelligence and Citizen Participation
Users authenticate via web or mobile app.
MONITOR QUEIMADAS
lisia
Entrar
Cadastrar
Criar Conta
Acessar
Digite seu email
Digite sua senha
Esqueceu a senha?
Entrar
Entrar com o Google
É o seu primeiro acesso? Clique aqui!
API
01
Requests are handled by a RESTful API.
02
AI
Servers use AI to analyze data and predict wildfires.
03
Monitor Queimadas Cariri
Mapa Do Lixo Crajubar
Métricas
AI-generated insights are shown on interactive dashboards.
04
Permissões necessárias
Câmera
Localização
Alerta de Queimada
Voltar
Avançar
05
Users report fires with photos and geolocation, feeding the AI models.

Figure 2. Operational flow of the wildfire monitoring system with predictive AI and citizen participation. (01) Users authenticate via web portal or mobile app; (02) requests are handled by a RESTful API; (03) servers execute the AI pipeline to analyze data and predict wildfires; (04) results and AI-generated insights are displayed on interactive dashboards; (05) citizens report hotspots with photos and geolocation, feeding back into the models.

System Operational Flow (Fig. 2)

(01) Authentication and access. Users access the system via the web portal or mobile app. Login creates an authenticated session and enables both the visualization of dashboards and the submission of georeferenced reports with images. (02) Data exchange via RESTful API. All interactions pass through an API that normalizes requests and responses, applies access control, rate limiting, and versioning. This layer serves (i) the integrated official INPE/INMET data and (ii) events from citizen participation for analytical processing. (03) Intelligent processing and prediction. On the servers, Steps 1–6 are executed: data fusion, feature engineering, fitting of regressors (with emphasis on ensembles such as ExtraTrees), logistic transformation into fire probability, and LLM orchestration (Moderator→Risk→Decision) to produce auditable alerts. (04) Visualization and decision support. Outputs are made available on interactive dashboards (maps, time series, and risk cards) for near real-time inspection and integration with the routines of civil defense and environmental agencies. (05) Citizen participation and feedback. The application allows reporting hotspots with photos and coordinates. After verification, these reports return to the API (step 02) and enrich the dataset, continuously improving predictions.

5. Results and Discussion

This section presents the results obtained with the fused INPE+INMET dataset and discusses them with statistical rigor. We display (i) a comparative analysis of regressors with classic error metrics and residual-based statistical tests; (ii) a critical reading of suspiciously perfect scores; and (iii) contextualized discussion vis-à-vis the literature, avoiding misleading claims of superiority.

5.1. Wildfire Hotspot Regression (Comparative Results)

Table 1 summarizes residual diagnostics (Shapiro–Wilk, paired *t*-test, Pearson), while Table 2 and Fig. 3 report MAE, RMSE, MAPE, and R^2 for all models.

Table 1. Statistical validation results (20% of 2003–2023) for the methods discussed.

Model	Shapiro–Wilk		Paired *t*-test		Pearson Correlation	
	Statistic	*p*-value	Statistic	*p*-value	*r*	*p*-value
LinearRegression	9.7505e-01	6.3085e-01	-1.2669e+00	2.1433e-01	9.0229e-01	7.3624e-13
Ridge	9.4696e-01	1.0843e-01	-1.2860e+00	2.0766e-01	9.0425e-01	5.4558e-13
Lasso	9.6004e-01	2.5918e-01	-1.3790e+00	1.7745e-01	9.0324e-01	6.3758e-13
ElasticNet	9.4290e-01	8.2528e-02	-1.2948e+00	2.0466e-01	9.0325e-01	6.3614e-13
BayesianRidge	9.6261e-01	3.0587e-01	-1.2830e+00	2.0871e-01	9.0558e-01	4.4320e-13
SVR	9.0836e-01	8.8372e-03	-1.7423e-01	8.6278e-01	9.4612e-01	9.8839e-17
KNeighbors	9.2983e-01	3.4613e-02	7.4966e-01	4.5894e-01	9.4221e-01	2.8491e-16
DecisionTree	8.5477e-01	4.3431e-04	-5.1589e-02	9.5918e-01	7.0922e-01	3.8317e-06
RandomForest	8.1513e-01	6.3588e-05	1.9748e-01	8.4470e-01	9.3915e-01	6.1927e-16
GradientBoosting	8.6449e-01	7.2170e-04	-6.9587e-01	4.9154e-01	9.3294e-01	2.6764e-15
ExtraTrees	**7.8569e-01**	**1.7421e-05**	**-5.1779e-02**	**9.5903e-01**	**9.3623e-01**	**1.2548e-15**
AdaBoost	8.6776e-01	8.5920e-04	5.4793e-01	5.8755e-01	9.2506e-01	1.4170e-14
HistGradientBoosting	8.1677e-01	6.8544e-05	-9.0043e-01	3.7462e-01	9.1879e-01	4.7100e-14

According to Shapiro–Wilk, which assesses whether prediction errors follow a normal distribution, Table 1 shows several *p*-values below 0,05, except for Linear, Ridge, Lasso, ElasticNet, and BayesianRidge. This indicates that, in general, errors do not follow perfect normality, something common in real wildfire data. RandomForest and ExtraTrees exhibit strong deviations from normality ($p \approx 6\times10^{-5}$ and 2×10^{-5}), suggesting that they still do not fully capture the phenomenon.

As for the paired *t*-test, which assesses statistical significance between predictions and actual values, Table 1 indicates that no model shows statistically significant mean differences (all $p > 0{,}05$).

The Pearson coefficient reflects the strength of the linear relationship between predicted and observed (Table 1). The closer to 1, the better the regressor explains variability. Here, SVR achieves the highest r=0,946 (p<10^{-16}), indicating predictive capability—without ruling out overfitting.

Table 2. Regression metrics for models applied to the INPE + INMET dataset.

Model	MAE	RMSE	MAPE	R^2
LinearRegression	14.9780	21.8605	247.31%	0.6904
Ridge	14.7860	21.9775	252.50%	0.6871
Lasso	14.7280	21.9506	245.28%	0.6878
ElasticNet	14.7690	22.0112	253.88%	0.6861
BayesianRidge	14.8539	21.8997	248.00%	0.6893
SVR	9.1447	18.4861	82.55%	0.7786
KNeighbors	**0.0000**	**0.0000**	**0.00%**	**1.0000**
DecisionTree	9.2108	17.1043	87.06%	0.8104
RandomForest	5.8701	9.3245	79.66%	0.9437
GradientBoosting	4.2109	7.9961	51.24%	0.9586
ExtraTrees	**2.4069**	**4.9148**	**32.43%**	**0.9843**
AdaBoost	6.7553	10.1605	99.13%	0.9331
HistGradientBoosting	**0.0183**	**0.0483**	**0.16%**	**1.0000**

Two rows warrant caution: KNeighbors and HistGradientBoosting display near-perfect scores (R^2=1,00; MAE/RMSE $\approx$ 0). Such results are atypical for real wildfire data and are *strong indications* of leakage or non-chronological evaluation (e.g., random splits instead of temporal blocks). We retain them for transparency but treat them as *overfitting* and exclude them from generalization claims. Future evaluations will enforce chronological hold-out (train on early years, test on later years) and/or blocked cross-validation.

Among *credible* contenders, ensembles stand out. ExtraTrees balances error and stability (MAE $=$ 2,4069, RMSE $=$ 4,9148, MAPE $=$ 32,43%, R^2=0,9843). GradientBoosting also performs well, though with slightly larger errors. DecisionTree serves as an interpretable baseline but, as expected, falls short of ensembles. The findings align with the literature: random sampling and averaging reduce variance and tend to yield better out-of-sample behavior in heterogeneous environmental data.

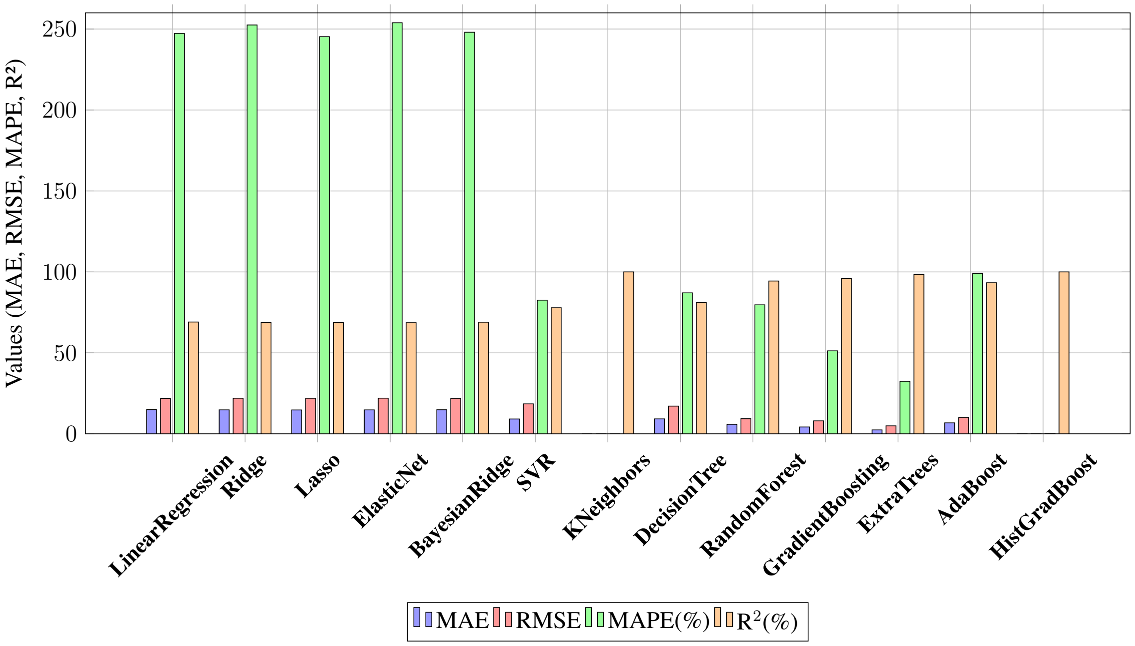

Figure 3. Comparative chart of regression metrics (MAE, RMSE, MAPE, and R^2) for each model in Table 2. To display R^2 on the same axis as MAPE, it was converted to a percentage (e.g., $R^2 = 0.6904 \rightarrow 69.04\%$).

5.2. Positioning Relative to Time-Series Studies

Previous studies, such as [Viganó et al. 2018], applied ARIMA for wildfire prediction in the Pantanal. We include their results strictly as *contextual reference* to situate approaches. It is not a baseline for our system because (i) the paradigms differ (our method is multivariate regression with meteorology and derived temporal features; ARIMA is univariate/low dimensionality) and (ii) geographic/temporal scopes differ (Cariri vs. Pantanal).

Under these caveats, ExtraTrees on the Cariri dataset attains higher R^2 (0,9843 vs. $0{,}665$) and comparable MAPE ($32{,}43\%$ vs. $31{,}60\%$). We do not claim superiority over [Viganó et al. 2018]; the evidence suggests that, in Cariri, richer features combined with ensembles explain a larger fraction of variance. A fair test would require retraining ARIMA/SARIMA/S-Naïve on the *same* Cariri splits with chronological validation left as future work.

5.3. LLM Module: Stratified Audit, Compliance, and Cost

Audit design. Population: N_{pop}=1000. Audit size: M_{audit} (required)= 20; M_{audit} (used)= 12;effective sample size n_{eff}=8,16. Calls: $\sim$40 per model/round; total $\sim$80 (2 models $\times$ 2 variants $\times M$). Threshold: proxy/train $0{,}500$; *live* (OOF) $0{,}758$.

Compliance and operational cost. Across the audited window we issued 48 calls (Table 4), enough to stress the pipeline under realistic load while controlling spend. The JSON parse rate of $0{,}980$ in Table 3 indicates that nearly all responses were schema-compliant on first pass; the residual $\sim 2\%$ were auto-recovered via the fallback path without operator intervention. The Qwen fallback $0{,}020$ and the rule-based fallback $0{,}020$ quantify these automatic recoveries: the first re-queries a redundant model, while the second coerces well-formed JSON deterministically both preventing alert loss. The rate of

Table 3. Weighted metrics (stratified audit) at *live* threshold $0{,}758$ (values ± deviation; 95% CI in brackets).

Metric (w)	Value ± deviation	95% CI
Accuracy	$0{,}7973 \pm 0{,}1165$	$[0{,}5428,\ 1{,}0000]$
Precision	$0{,}7438 \pm 0{,}1826$	$[0{,}3660,\ 1{,}0000]$
Recall	$0{,}9070 \pm 0{,}1018$	$[0{,}7044,\ 1{,}0000]$
F1	$0{,}8062 \pm 0{,}1380$	$[0{,}4895,\ 1{,}0000]$
ROC–AUC	$0{,}7123 \pm 0{,}1719$	$[0{,}3189,\ 1{,}0000]$
Brier	$0{,}3411 \pm 0{,}1135$	$[0{,}1379,\ 0{,}5624]$

`<think>` artifacts ($0{,}020$) remained negligible and was filtered before logging, preserving clean audit trails. Per-model service indicators in Table 4 show low 429_rate and mean latencies between 3.08–3.70 s with p95 < 6 s, which bounds the tail and keeps dashboards responsive; by definition, p90 falls below that bound, so typical alert rendering occurs well within the 6 s envelope. Collectively, these indicators evidence operational stability high machine-readability, rare retries, and controlled latency tails crucial for timely and dependable alerts in civil-protection settings.

Table 4. Per-model operation (audited window).

Model	Calls	429_rate	lat_mean	lat_p95
qwen-3-235b-a22b-instruct-2507	24	0,038	3,70s	5,95s
llama-4-scout-17b-16e-instruct	24	0,038	3,08s	4,75s

Implications - At the live threshold $0{,}758$, the stratified audit in Table 3 preserves high weighted recall ($\sim 0{,}91$), which directly reduces false negatives and is the primary safety requirement for civil-defense workflows (missing a true risk is costlier than inspecting a false alert). The observed precision (w) $\sim 0{,}74$ keeps the triage workload compatible with on-call teams, and the resulting F1 (w) $\sim 0{,}81$ summarizes a balanced operating point. The Brier score (w) $\sim 0{,}34$ indicates probabilistic quality that is usable in production and can be further improved via calibration (Platt or isotonic) to tighten thresholding by expected cost. Service-level indicators in Table 4 show sparse 429 rates with backoff and p95 < 6 s, meaning that 95% of alerts are emitted within the 6-second envelope required for near real-time dashboards. Although p90 was not separately logged in this audit window, by construction p90 $<$ p95; operationally, most responses concentrate between the mean latency (3.08–3.70 s) and the reported p95, sustaining a responsive operator experience. The JSON parse rate of $0{,}980$ (Table 3) ensures schema-compliant messages with minimal human intervention, reducing retries and alert drops.

5.4. LLM Module: Reliability and Operational Trade-offs

Consolidating Tables 3 and 4, the operating point yields: Accuracy (w) $\sim 0{,}80$, Precision (w) $\sim 0{,}74$, Recall (w) $\sim 0{,}91$, F1 (w) $\sim 0{,}81$, ROC–AUC (w) $\sim 0{,}71$, Brier (w)

$\sim$ 0,34, JSON parse $=$ 0,98, and p95 $<$ 6 s. Each metric adds a distinct guarantee to the problem: (i) Recall governs safety by limiting false negatives; (ii) Precision bounds field workload; (iii) F1 captures the operating balance; (iv) ROC–AUC supports post-hoc threshold retuning when agencies change cost priorities; (v) Brier quantifies the reliability of the probabilities that drive cost-sensitive policies; (vi) p95 anchors worst-case responsiveness for live dashboards, while the (implicit) p90 below that level reflects typical user experience; and (vii) parse rate ensures machine-readability for automated dispatch. In high-risk contexts, this profile favors conservative triage (high recall) without saturating teams, and it remains amenable to calibration and threshold optimization to reduce false positives when desired.

6. Conclusion

The research presented an information system based on artificial intelligence for predicting forest fires. The fire monitoring system proved to be an effective tool for predicting fires using different computational models. The research presents different comparisons between predictive models combined with LMM for forest fire solutions and monitoring using real-time satellite data.

We presented an end-to-end operational pipeline for wildfire monitoring in Cariri, based on data fusion of INPE hotspots and INMET meteorological variables, comparative regression modeling with statistical validation, probabilistic mapping, and an auditable decision layer with LLMs. Among credible contenders, ExtraTrees showed strong performance (R^2=0,9843; favorable MAE/RMSE/MAPE).

The stratified audit of the LLM module with *live* threshold 0,758 and n_{eff}=8,16—recorded Recall (w) 0,9070 $\pm$ 0,1018, Accuracy (w) 0,7973 $\pm$ 0,1165, AUC (w) 0,7123 $\pm$ 0,1719, and Brier (w) 0,3411 $\pm$ 0,1135, with JSON parse rate 0,980 (raw 1,0) and p95 $<$ 6s, supporting a safe profile (low FNs) and operational stability.

As future work, we propose (i) probabilistic calibration and threshold optimization by cost; (ii) strict chronological validation with temporal baselines (S-Naïve, SARIMA) on the same splits; and (iii) transfer studies across regions.

Acknowledgements

Acknowledgements to the Laboratory of Innovation in Artificial Intelligence Systems (LISIA), responsible for conducting the research through the commitment and dedication of its researchers, with special thanks to the lead researcher, Luís Fabrício de Freitas Souza, who led the research and the scientific, methodological, and technical development of this work. We also thank the Instituto Atlântico for its collaboration and institutional support throughout the execution of the activities, strengthening the integration between academic research and applied innovation. The authors also acknowledge the Interset-CE Network (MEC) for its institutional support and coordination.

References

da Costa Nascimento, J. J., Marques, A. G., do Nascimento Souza, L., de Mattos Dourado, C. M. J., da Silva Barros, A. C., de Albuquerque, V. H. C., de Freitas Sousa, L. F., et al. (2025). A novel generative model for brain tumor detection using magnetic resonance imaging. *Computerized Medical Imaging and Graphics*, 121:102498.

Da Silva, P. M., Lima, M. N., Soares, W. L., Silva, I. R., Fagundes, R. A. d. A., and De Souza, F. F. (2019). Ensemble regression models applied to dropout in higher education. In *2019 8th Brazilian conference on intelligent systems (BRACIS)*, pages 120–125. IEEE.

de Andrades, R. S., Grellert, M., and Fonseca, M. B. (2019). Hyperparameter tuning and its effects on cardiac arrhythmia prediction. In *2019 8th Brazilian conference on intelligent systems (BRACIS)*, pages 562–567. IEEE.

Mantovani, R. G., Horváth, T., Cerri, R., Vanschoren, J., and De Carvalho, A. C. (2016). Hyper-parameter tuning of a decision tree induction algorithm. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 37–42. IEEE.

MapBiomas (2025). Área queimada no brasil cresce 79% em 2024 e supera os 30 milhões de hectares. Acesso em: 28 abr. 2025.

Marques, A. G., de Figueiredo, M. V. C., da Costa Nascimento, J. J., de Souza, C. T., de Mattos Dourado, C. M. J., de Albuquerque, V. H. C., de Freitas Souzal, L. F., et al. (2024). New approach generative ai melanoma data fusion for classification in dermoscopic images with large language model. In *2024 37th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 1–6. IEEE.

Melo, M. A. d. (2024). Desenvolvimento de um plugin no qgis para tratamento de dados disponibilizados pelo inmet.

Mohnish, S., Kannan, B. D., Vasuhi, S., et al. (2023). Vision transformer based forest fire detection for smart alert systems. In *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 891–896. IEEE.

Murilo, L., Oliveira, G., and Martins, L. (2024). Evolutionary adjustment of a cellular automata-basedmodel for wildfire spreading. In *Anais da XXXIV Brazilian Conference on Intelligent Systems*, pages 260–275, Porto Alegre, RS, Brasil. SBC.

Nieuwenhuijsen, M. J. (2024). Climate crisis, cities, and health. *The Lancet*, 404(10463):1693–1700.

Park, M., Jeon, Y., Bak, J., Park, S., et al. (2022). Forest-fire response system using deep-learning-based approaches with cctv images and weather data. *IEEE access*, 10:66061–66071.

Peng, W., Wei, Y., Chen, G., Lu, G., Ye, Q., Ding, R., Hu, P., and Cheng, Z. (2023). Analysis of wildfire danger level using logistic regression model in sichuan province, china. *Forests*, 14(12):2352.

Povo, O. (2023). Incêndios florestais geram nuvem de fumaça em cidades do cariri. Acesso em: 28 abr. 2025.

Severiano, G. F. B., Marques, A. G., Nascimento, J. J. d. C., and Rodrigues, Y. O. A. (2024). Alpr system perspective adjustment: New automatic license plate. *Intelligent Systems Design and Applications: Real World Applications, Volume 5*, 1050:295.

Silva, A. P., Genaro, A. F., and Branco, R. H. (2024). A contribuição de sistemas de informação para o processo de gestão do portfólio de projetos e programas do inpe. In *Workshop de Computação Aplicada em Governo Eletrônico (WCGE)*, pages 210–221. SBC.

Sousa, M. J., Moutinho, A., and Almeida, M. (2019). Classification of potential fire outbreaks: A fuzzy modeling approach based on thermal images. *Expert systems with applications*, 129:216–232.

Suklabaidya, S. and Das, I. (2023). Processing iot sensor fire dataset using machine learning techniques. In *2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC)*, pages 1–7. IEEE.

Umoh, U. A., Eyoh, I. J., Murugesan, V. S., and Nyoho, E. E. (2022). Fuzzy-machine learning models for the prediction of fire outbreaks: A comparative analysis. In *Artificial intelligence and machine learning for EDGE computing*, pages 207–233. Elsevier.

Viganó, H. H. d. G., Souza, C. C. d., Reis Neto, J. F., Cristaldo, M. F., and Jesus, L. d. (2018). Prediction and modeling of forest fires in the pantanal. *Revista Brasileira de Meteorologia*, 33(2):306–316.

Weisse, M. and Goldman, E. (2023). Latest analysis of deforestation trends. Accessed on: March 13, 2025.

Xu, R., Lin, H., Lu, K., Cao, L., and Liu, Y. (2021). A forest fire detection system based on ensemble learning. *Forests*, 12(2):217.

Zhang, J., Cai, S., Jiang, Z., Xiao, J., and Ming, Z. (2024). Firerobbrain: Planning for a firefighting robot using knowledge graph and large language model. In *2024 10th IEEE International Conference on Intelligent Data and Security (IDS)*, pages 37–41. IEEE.