MDPI

Article

An Interpretable Artificial Intelligence Approach for Reliability and Regulation-Aware Decision Support in Power Systems

Diego Armando Pérez-Rosero[1], Santiago Pineda Quintero[1], Juan Carlos Álvarez Barreto[2], Andrés Marino Álvarez-Meza[1], and German Castellanos-Dominguez[1]

[1]Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales 170003, Colombia; spinedaq@unal.edu.co (S-P.-Q.); cgcastellanosd@unal.edu.co (G.C.-D.)
[2]Central Hidroeléctrica de Caldas - CHEC-Grupo EPM, Manizales 810003, Colombia; juan.alvarez.barreto@chec.com.co (J.C.A.-B.)
* Correspondence: dieaperezros@unal.edu.co (D.A.P.-R.) and amalvarezme@unal.edu.co (A.M.Á.-M.)

Abstract

Modern medium-voltage (MV) distribution networks face increasing reliability challenges driven by aging assets, climate variability, and evolving operational demands. In Colombia and across Latin America, reliability metrics such as the System Average Interruption Frequency Index (SAIFI), standardized under IEEE 1366, serve as key indicator for regulatory compliance and service quality. However, existing analytical approaches struggle to jointly deliver predictive accuracy, interpretability, and traceability required for regulated environments. Here, we introduces CRITAIR (Criticality Analysis through Interpretable Artificial Intelligence-based Recommendations), an integrated framework that combines predictive modeling, explainable analytics, and regulation-aware reasoning to enhance reliability management in MV networks. CRITAIR unifies three components: (i) a TabNet-based predictive module that estimates SAIFI using outage, asset, and meteorological data while producing global and local attributions; (ii) an agentic retrieval-and-reasoning stage that grounds recommendations in regulatory evidence from RETIE and NTC 2050; and (iii) interpretable reasoning graphs that map decision pathways for full auditability. Evaluations conducted on real operational data demonstrate that CRITAIR achieves competitive predictive performance—comparable to Random Forest and XGBoost—while maintaining transparency through sparse attention and sequential feature explainability. Also, our regulation-aware reasoning module exhibits coherent and verifiable recommendations, achieving high semantic alignment scores (BERTScore) and expert-rated interpretability. Overall, CRITAIR bridges the gap between predictive analytics and regulatory governance, offering a transparent, auditable, and deployment-ready solution for digital transformation in electric distribution systems.

Keywords: Artificial intelligence; Agentic RAG; Tabular data; Explainable AI; Power Systems; TabNet

Received:
Revised:
Accepted:
Published:

Citation: Pérez-Rosero, D.A.; Pineda-Quintero, S; Álvarez-Barreto J.C; Álvarez-Meza, A.M.; Castellanos-Dominguez, G. An Interpretable Artificial Intelligence Approach for Reliability and Regulation-Aware Decision Support in Power Systems. *Computation* **2025**, *1*, 0. https://doi.org/

1. Introduction

Modern Medium-Voltage (MV, 1–36 kV) distribution networks operate under heterogeneous and evolving conditions—aging assets, climate variability, and growing demand—that erode service continuity and, in turn, system-level reliability indicators [1]. Improving those indicators is a central objective for electric distribution companies seeking to elevate power supply quality [2]. In this sense, reliability is internationally assessed via the System Average Interruption Duration Index (SAIDI) and the System Average Interruption

Frequency Index (SAIFI), standardized in IEEE Std 1366, which harmonizes interruption-event data collection and categorization to ensure consistency in reporting [3]. These technical frameworks are further contextualized by trend and policy analyzes that track recent performance, alongside regional studies across Latin America and the Caribbean that use SAIDI/SAIFI to evaluate regulatory impacts on service quality [4,5]. In addition, the sector's growing emphasis on distribution-system resilience expands the remit of traditional indices by integrating preparedness, response, and recovery practices into planning and operations [6].

In Colombia, these international standards are instantiated through the regulatory framework established by the Comisión de Regulación de Energía y Gas (CREG), which in 2024 operationalized annual SAIDI/SAIFI targets for distribution system operators [7]. Oversight and enforcement fall to the Superintendencia de Servicios Públicos Domiciliarios (Superservicios), which publishes sector diagnostics. Meanwhile, XM—as the system and market operator—provides official data series that enable continuous quality monitoring [8]. This regulatory scaffolding is underpinned by a robust technical corpus: the Reglamento Técnico de Instalaciones Eléctricas (RETIE) and the Código Eléctrico Colombiano (NTC 2050) that ensure traceability and regulatory compliance in asset management and operations [9,10]. At the regional level, the Central Hidroeléctrica de Caldas (CHEC-Grupo EPM) exemplifies this scheme, with public reports on targets, outcomes, and investment plans aligned to SAIDI/SAIFI improvements that provide an operational substrate to connect analytics with capital planning and decision-making [11,12].

To meet these regulatory and operational demands, utilities are advancing digital-transformation agendas whose strategic aim is to convert large, multi-source datasets—outage logs, equipment metadata, and meteorological information—into actionable, regulation-aware decisions that strengthen resilience and transparency [13]. Significant hurdles persist; however, manual analyzes and static reports are insufficient to surface complex, cross-factor patterns at scale. In contrast, whereas "black-box" analytics face adoption barriers in regulated environments that require full traceability and auditability of results [14–16].

On this basis, Explainable Artificial Intelligence (XAI) and Interpretable Machine Learning (IML) provide a pragmatic bridge between high predictive performance and auditable decision support systems. Recent research has systematized explainability techniques and discussed pathways for their integration and governance in the power sector [17]. Empirical evidence supports this direction, with studies demonstrating successful applications of machine-learning models to predict outage duration and restoration time—leveraging transfer learning strategies and feature sets compiled from public data that enable reproducible forecasting pipelines [18,19]. Taken together, these advances facilitate a transition from opaque analytics to transparent, auditable recommendation systems, thereby improving risk management and the prioritization of operational actions in a highly regulated service environment [20].

In this context, the challenge coalesces around two complementary fronts that hinder proactive reliability management in MV networks: First, the lack of models with predictive and explanatory capabilities—approaches must estimate SAIFI while articulating the drivers of interruptions, explicitly incorporating external variables (e.g., meteorology, construction metadata) to capture cross-circuit and cross-season variability. Namely, they should provide consistent global and local explanations and remain stable under shifts in asset configurations so that forecasts can support maintenance scheduling and capital planning [21–24]. Second, the absence of integrated, interpretable decision-support systems, in which insights from heterogeneous data are fused with domain knowledge (e.g., RETIE, NTC 2050), leads to unclear, actionable, and trustworthy recommendations with full

traceability and explicit justification. Then, such systems should link analytical evidence to regulatory clauses and procedural artifacts while maintaining audit trails [25–27].

Existing approaches bifurcate into predictive modeling and decision-support. Linear and other classical regressors are simple but struggle with nonlinearities and exogenous drivers; ensemble methods improve accuracy yet provide limited transparency for regulated use [28]. Moreover, deep neural networks can be accurate yet opaque, while TabNet-based approaches offer a balanced alternative for tabular reliability modeling: sparse attention and sequential feature selection provide global/local attributions [29,30]. For decision-support, LLM-based QA improves access to RETIE/NTC but risks hallucinations and limited traceability [31]. Retrieval-Augmented Generation (RAG) grounds answers in retrieved evidence, though remains constrained for multi-source, tool-based reasoning [32]. Agentic and Multi-Agent RAG extend this by adding planning and tool orchestration across structured (outage logs) and unstructured (regulations, reports) sources, enabling auditable recommendations [33].

We propose CRITAIR (Criticality Analysis through Interpretable AI-based Recommendations), a hybrid, interpretable reliability framework that delivers accurate predictions, regulation-aware recommendations, and full auditability for MV operations. The core idea is to couple an interpretable TabNet pipeline with an agentic retrieval-and-reasoning layer and explicit reasoning graphs, unifying predictive attribution, verifiable evidence retrieval, and transparent decision paths. CRITAIR is implemented as an end-to-end architecture consisting of three key stages:

- Predictive and Interpretable Modeling (TabNet): Train a TabNet-based pipeline employing enhanced data outage records (endogenous and exogenous variables) to estimate SAIFI while producing global and local attributions for critical factors.
- Regulation-Aware Retrieval and Reasoning (Agentic RAG): Enable multi-step retrieval over RETIE/NTC and internal documents, grounding answers and suggested actions in cited clauses and context, with planning/tool-use for multi-source evidence integration.
- Interpretable Reasoning Graphs and Evidence Attribution: Transform the complete decision pathway—prioritized characteristics, extracted regulatory components, and inference processes—into auditable graphs that fulfill explainability standards in power system operations.

We evaluate CRITAIR on a real MV operational dataset from CHEC, comprising historical outage records, asset metadata, and 24-hour antecedent meteorological variables. For the predictive stage, TabNet is benchmarked against strong baselines (linear models, Random Forest, XGBoost), showing fast convergence and competitive reliability estimates while maintaining instance-level and global interpretability via sparse attention and sequential feature selection. In parallel, the agentic RAG subsystem is evaluated for querying structured outage tables, interpreting regulatory documents (e.g., RETIE, NTC 2050), and generating criticality-based recommendations; performance is measured using BERTScore across structured queries, normative interpretation, and recommendation synthesis, complemented by expert validation. Qualitative analysis—via TabNet attention masks and interpretable reasoning graphs—demonstrate clear inter-asset separability, stable feature salience across contexts, and regulation-aware semantic coherence in recommended actions, underscoring CRITAIR's suitability for deployment in audit-constrained utility environments.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 details the materials and methods. Sections 4 and 5 present the experiments and results. Finally, Section 6 provides concluding remarks.

2. Related Work

Research on reliability prediction in MV distribution networks has progressed substantially, transitioning from traditional statistical approaches to modern deep learning architectures tailored for tabular and heterogeneous data. Early linear models—including ordinary least squares, ridge regression, Lasso, and Elastic Net—remain appealing due to their low computational cost, interpretability, and ease of deployment in utilities with constrained analytical capabilities [34]. Nevertheless, their strictly linear structure hampers their ability to model complex interactions among grid components and to incorporate exogenous drivers such as precipitation, wind gust intensity, vegetation encroachment, or construction-related metadata. As a result, these methods often struggle to generalize under highly variable operational environments typical of real distribution systems [3].

To overcome the limitations of purely linear approaches, more flexible nonlinear models have been introduced into reliability prediction pipelines. Classical machine-learning algorithms—such as k-nearest neighbors and support vector regression (SVR)—offer improved expressiveness by capturing local patterns and nonlinear dependencies in outage behavior [35]. However, these methods often face scalability challenges when dealing with high-dimensional geospatial, environmental, and construction metadata, and their performance can degrade sharply under domain shifts or sparse event distributions, which are common in MV systems. Tree-based ensemble methods, particularly Random Forests and gradient-boosting algorithms like XGBoost, have demonstrated superior predictive accuracy by modeling nonlinear interactions and higher-order feature dependencies [19]. These approaches have been widely used to estimate key reliability indices—such as SAIDI, SAIFI, and CAIDI—across heterogeneous operating conditions. Despite their strong empirical performance, their limited interpretability remains a barrier to adoption in regulated utility environments where transparency, auditability, and explainability are mandatory [20,23]. Feature-importance heuristics, while informative, seldom provide the level of causal or mechanistic traceability required by domain experts and regulatory agencies.

The emergence of deep learning architectures has introduced an additional tier of predictive capability. Deep neural networks (DNNs), when trained on large outage logs enhanced with high-resolution meteorological, vegetation, and asset-condition data, have achieved state-of-the-art performance in predicting outage frequency, duration, and restoration time [15,16]. Yet, their inherently opaque "black-box" representations make them difficult to justify in high-stakes operational settings, particularly in contexts subject to regulatory oversight and safety-critical decision-making [30]. A more recent development is TabNet, a deep learning architecture explicitly designed for tabular data. By leveraging sparse attention and sequential feature selection, TabNet provides both global and local interpretability [29]. It integrates exogenous variables, highlights their relative contribution to outage risk, and preserves transparency in the decision process. This makes it particularly well suited for reliability studies in MV, where utilities must justify both predictive performance and regulatory compliance. Figure 1 summarizes the evolution of the predictive approaches reviewed.

In turn, large language models (LLMs) have evolved into three principal architectural families—only-encoder, only-decoder, and encoder–decoder—each tailored to specific natural language processing (NLP) task types. Understanding their respective strengths and limitations is essential to selecting models suitable for explainable, regulation-sensitive reliability systems. Only-encoder models, such as BERT, RoBERTa, and DistilBERT, rely on bidirectional transformers that contextualize input sequences without generating text [36,37]. They excel in extractive and discriminative tasks, including text classification, entity recognition, and span-based question answering. Their deep bidirectional attention enables fine-grained contextual understanding. However, the lack of generative capability limits

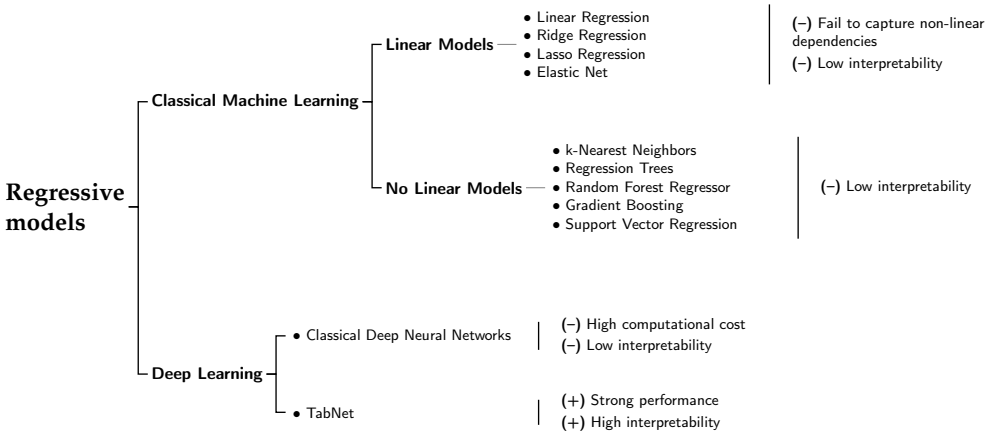

Figure 1. Reliability prediction methodologies: linear classical models, nonlinear machine-learning algorithms, deep neural networks, and attention-based architectures tailored for tabular data.

their use in tasks that require producing coherent explanations, summaries, or recommendations—functions central to decision-support systems.

Only-decoder models, typified by autoregressive architectures such as GPT, Gemini, LLaMA, Qwen, and DeepSeek, generate text token by token in a unidirectional manner [38–40]. This makes them inherently generative, excelling at tasks such as dialogue systems, reasoning, and contextual report synthesis. Their autoregressive design allows the progressive construction of fluent, semantically consistent text, making them especially suitable for explanatory and reasoning-oriented applications. Although only-decoder models lack the explicit bidirectional context of encoder–decoder architectures, their ability to handle long prompts and instruction-based conditioning compensates for this limitation in most real-world reasoning pipelines. Furthermore, through instruction tuning and reinforcement learning, these models can align text generation with domain-specific constraints—such as regulatory compliance or reliability terminology—while maintaining adaptability across diverse task types. Further, encoder–decoder models combine both paradigms, using a dedicated encoder to process the input and a decoder to generate outputs [41,42]. They are particularly effective for sequence-to-sequence tasks, such as translation or summarization, where input and output spaces differ. Despite their interpretability and structured conditioning, encoder–decoder models are typically more computationally demanding and slower during inference, which limits their applicability in interactive or multi-agent reasoning systems.

Beyond predictive modeling, another research frontier focuses on decision-support systems capable of translating analytical outputs into clear, auditable, and regulation-aware recommendations. Early approaches relied on LLM-based question answering (QA) systems, which allowed practitioners to query technical regulations such as RETIE or NTC 2050 directly [31,43]. These systems facilitated access to normative documents but suffered from hallucinations, lack of traceability, and limited contextual reasoning. To address these issues, Retrieval-Augmented Generation (RAG) architectures emerged, combining semantic retrieval with grounded text generation. RAG systems reduce hallucinations and improve factual consistency by explicitly citing retrieved passages [17,32]. However, most implementations remain constrained to single-step queries and are limited in their ability to integrate structured datasets (e.g., outage logs, asset metadata) or to reason over temporal dynamics. Recent advances have introduced Agentic RAG and Multi-Agent RAG architectures, where autonomous agents plan, decompose, and execute multi-step reasoning processes [33,44]. These agents can orchestrate multiple tools—such as SQL connectors for outage tables, vector search engines for technical manuals, and regulatory

parsers for RETIE/NTC clauses—to integrate heterogeneous evidence into contextualized recommendations.

Complementary to these developments, Knowledge Graphs (KGs) play a central role in improving interpretability and reasoning. KGs represent entities, attributes, and relationships explicitly, enabling structured reasoning that complements statistical models [45,46]. In the power sector, they have been applied to fault diagnosis and asset management, encoding equipment lifecycles, causal dependencies, and environmental stressors to guide maintenance and investment strategies [47,48]. From an explainability perspective, rule-enhanced cognitive graphs have been proposed to embed logical rules into graph structures, supporting transparent causal inference in grid operations [47]. Beyond domain-specific applications, KGs also enhance NLP-driven decision support. Recent frameworks such as GraphRAG extend standard RAG by embedding KGs alongside vector indices, grounding outputs in explicit relational structures rather than isolated fragments [49]. Other approaches, such as KG-SMILE, attribute specific entities and relations as explanatory evidence for generated recommendations [50]. Despite these advances, several open challenges persist, including the design of robust domain ontologies, mechanisms for continuous and dynamic graph updates, and the computational scalability of multi-hop reasoning over large, heterogeneous graphs. Nonetheless, contemporary literature increasingly converges on the view that KG–enhanced reasoning provides a promising pathway toward transparent, auditable, and regulation-compliant decision-support systems, particularly in domains where interpretability is as critical as predictive accuracy. The progression from classical NLP systems toward multi-agent and KG-enhanced reasoning architectures can be synthesized as in Figure 2.

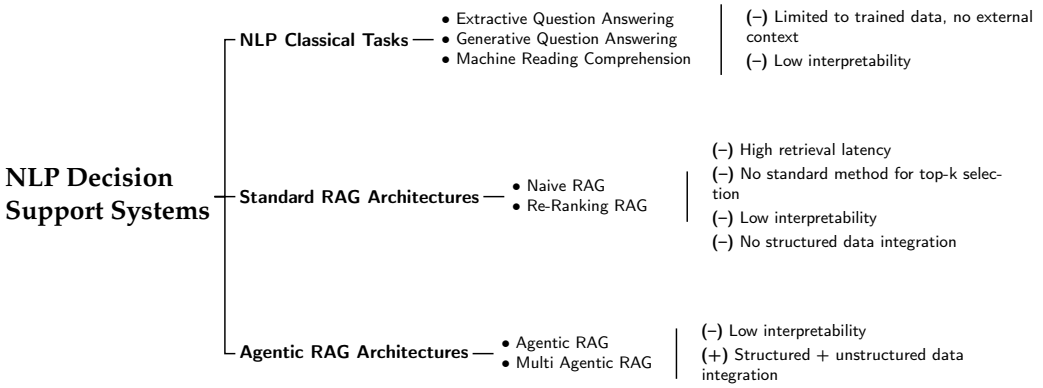

Figure 2. NLP-based decision-support system families: NLP Classic Tasks, Standard RAG Architectures, and Agentic RAG Architectures.

Taken together, the literature review highlights two complementary fronts in advancing reliability management for MV networks: predictive modeling with interpretability, where TabNet and related attention-based architectures combine predictive accuracy with global and local attributions [18,19,29]; and decision-support through NLP and KGs, where Agentic RAG and GraphRAG systems integrate heterogeneous evidence sources into contextualized and auditable recommendations [33,44,49]. These two fronts converge in the proposed CRITAIR methodology, which integrates interpretable predictive modeling, regulation-aware retrieval and reasoning, and explicit reasoning graphs. By unifying these advances, CRITAIR directly addresses the limitations of existing approaches and provides a hybrid, interpretable framework for reliability-oriented decision-making in MV networks under regulatory scrutiny.

3. Materials and Methods

3.1. CHEC Medium-Voltage Reliability Prediction Dataset

A comprehensive dataset was constructed for this study to support the prediction of electrical grid interruptions, utilizing statistical records from the CHEC from January 1, 2019, to June 30, 2024. The objective is to model the complex interaction between the structural characteristics of the network and dynamic environmental variables. The foundation of the dataset comprises interruption records, which document the operating protection device, the start and end times of the event, and service quality indices such as the SAIFI, formally defined as:

$$SAIFI = \frac{\sum_{i=1}^{\tilde{K}} N_i}{\tilde{N}}, \tag{1}$$

where N_i denotes the number of customers affected by interruption i, $\tilde{K}$ is the total number of interruptions considered in the analysis, and $\tilde{N}$ represents the total customer base served by the system. Each record was subsequently enriched with detailed structural information of the network assets, including poles, switches, transformers, and line sections. Following this, exogenous variables were integrated through spatiotemporal queries to contextualize each event. This enrichment process consists of three primary data blocks.

The first block comprises climatic variables, for which a dataset was incorporated using the Weatherbit API (https://www.weatherbit.io). For each interruption, hourly time series were extracted for the event's location over the 24-h period preceding the report time. An example of the time series extracted for a single event is illustrated in Figure 3.

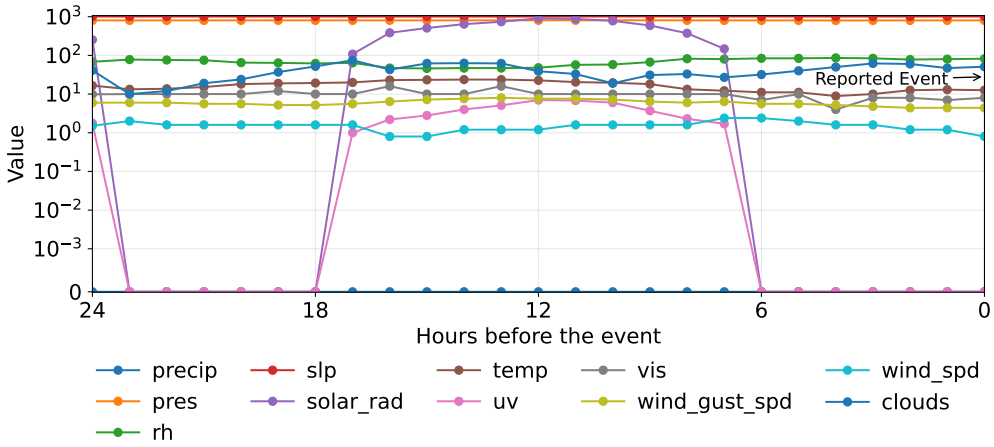

Figure 3. An example of a climatic variable time series extracted during the 24 hours preceding a reported event.

The variables integrated to characterize the operational environment include:

- Precipitation (precip): Associated with moisture-related risks for electrical components and grounding systems.
- Atmospheric Pressure (pres): Relevant at high altitudes, where it affects thermal dissipation and dielectric strength.
- Relative Humidity (rh): A critical indicator for corrosion and partial discharges.
- Sea Level Pressure (slp): Complements local pressure analysis and its impact on sensitive equipment.
- Solar Radiation (solar_rad): Accelerates material degradation under sunlight.
- Ambient Temperature (temp): Affects the thermal performance and lifespan of transformers and conductors.

- UV Index (uv): A determinant for the accelerated deterioration of polymeric materials.
- Visibility (vis): Relevant information for planning maintenance activities.
- Wind Gust Speed (wind_gust_spd): Related to additional mechanical loads on poles and conductors.
- Average Wind Speed (wind_spd): Affects the mechanical design and stability of overhead lines.
- Clouds (clouds): Satellite-based cloud coverage (%).

Furthermore, lightning strike activity was quantified by associating each event with discharges occurring within a 500 m radius during the preceding 24 hours. From this data, descriptive statistics for the current and altitude of the discharges were computed. Vegetation presence was determined by performing a spatial query within 30 m of each network section. This spatial enrichment process, depicted in Figure 4, culminates in the creation of the first primary structural database, where each event record is augmented with its immediate environmental context.

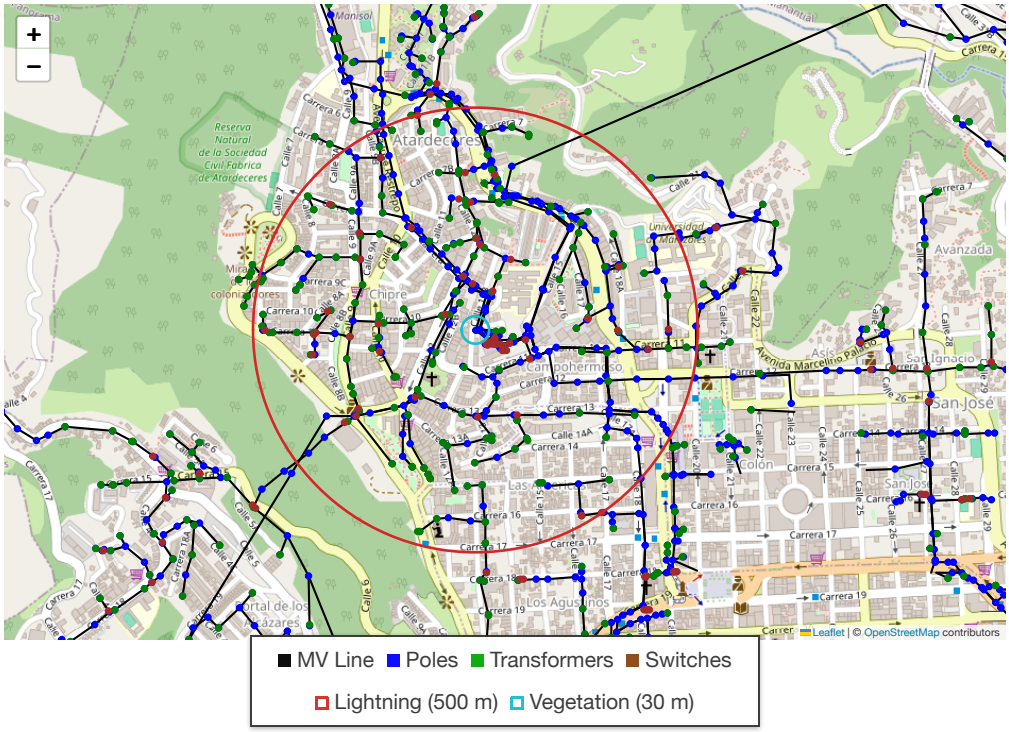

Figure 4. A visualization of the spatial data enrichment process. The figure displays network assets along with the query radii for lightning strikes and vegetation surrounding the network components.

To address the complexity of fault diagnostics, the dataset is constructed through the horizontal integration of multiple data sources, linked by operational keys (e.g., event ID, operating device, feeder). The structure of these data blocks and their preprocessing is summarized in Table 1.

Table 1. CHEC dataset structure by information block

Classification	Data Block (Columns)	Description
Structural	Events Data [0–9)	Core interruption metadata for incident identification and context.
Structural	Switches Data [9–17)	Operational and typological attributes of switching devices.
Structural	Transformers Data [17–28)	Nameplate and lifecycle attributes of power transformers.
Structural	MV Network Data [28–51)	Physical and topological properties of medium-voltage line sections.
Exogenous	Climatic Data [51–293)	Short-horizon local weather indicators around network assets.
Exogenous	Lightning Data [293–305)	Proximity-based indicators of lightning activity near assets.
Exogenous	Vegetation Data [305–306)	Surrounding vegetation and land-use typology near the network.
Structural	Supports Data [306–314)	Structural attributes of poles and associated components.

It is important to note that the total number of climatic features (242, corresponding to columns 51-293) is less than the theoretical maximum of 264 (11 variables over 24 hours). This difference arises from occasional unavailability during the data acquisition process. Also, a central difficulty in fault diagnostics is that the device that operates during an interruption is not necessarily the site of fault initiation. To address this, we implemented a downstream network-tracing algorithm that enumerates all assets electrically connected beyond the operated device. The event-level dataset was restructured into a component-level table tailored for root-cause analysis: each record corresponds to a candidate failing asset rather than an aggregated outage record. The associated metadata and climatic covariates were replicated across downstream assets for the relevant incident, whereas structural, lightning, and vegetation descriptors were assigned at the asset level. This representation enables the model to estimate, for each recorded event, the failure probability of every candidate asset independently.

Afterward, we assembled a regulation-focused corpus comprising RETIE, NTC, and CHEC technical standards. This corpus is augmented with a set of structured, asset-specific documents that map structural and exogenous variables to specific sections of each non-structural source. The resulting resource serves as input to an Interpretable Reasoning Graphs and Evidence Attribution module, which transforms the full decision pathway—prioritized characteristics, extracted regulatory clauses, and inference steps—into auditable graphs that satisfy explainability requirements for power-system operations.

3.2. Classical Regression Models

As a baseline for regression, Ordinary Least Squares (OLS) assumes a linear relationship between the input matrix $\boldsymbol{X} \in \mathbb{R}^{N\times P}$ (with N samples and P features) and the continuous target vector $\boldsymbol{y} \in \mathbb{R}^N$. The model coefficients $\boldsymbol{\theta} \in \mathbb{R}^P$ define this mapping as $\boldsymbol{y} = \boldsymbol{X\theta}$, estimated via the Moore–Penrose pseudoinverse:

$$\boldsymbol{\theta} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^{-1} \boldsymbol{X}^\top \boldsymbol{y}. \tag{2}$$

A regularized form is obtained by solving:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \|\boldsymbol{y} - \boldsymbol{X\theta}\|_2^2 + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2, \tag{3}$$

where $\lambda_1, \lambda_2 \geq 0$. When $\lambda_1 > 0$ and $\lambda_2 = 0$, the formulation yields LASSO regression [51]; when both $\lambda_1 > 0$ and $\lambda_2 > 0$, it becomes Elastic Net regression [35]. A key advantage of linear models is the direct interpretability of the coefficients $\boldsymbol{\theta}$. A schematic pipeline is shown in Figure 5.

Transcending linear constraints, Random Forests (RF) provide a powerful non-linear modeling approach by aggregating predictions from an ensemble of decision trees [52]. Operating on the same input data $\boldsymbol{X}$ and target $\boldsymbol{y}$, a non-linear prediction $\hat{\boldsymbol{y}} \in \mathbb{R}^N$ is formed by averaging the outputs from T_{RF} individual trees, where each tree function $f_t : \mathbb{R}^{N\times P} \to \mathbb{R}^N$ maps the input data to a vector of predictions [53]:

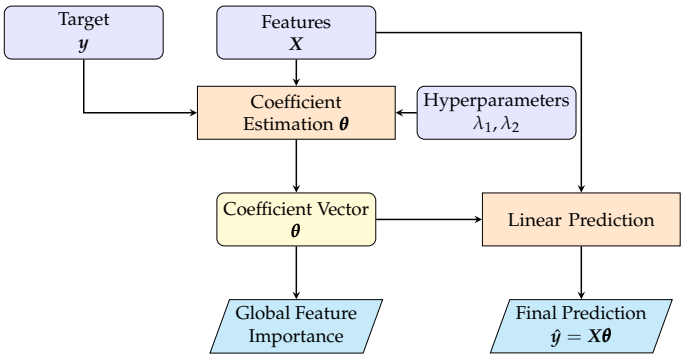

Figure 5. Schematic representation of a linear modeling workflow, summarizing inputs, parameter estimation, predictions, and global feature relevance.

$$\hat{\boldsymbol{y}} = \frac{1}{T_{RF}} \sum_{t=1}^{T_{RF}} f_t(\boldsymbol{X}). \tag{4}$$

Each tree t is trained on a bootstrap sample of indices $\mathcal{B}^{(t)} \subset \{1, \ldots, N\}$, and at each split, considers a random subset of feature indices $\mathcal{F}^{(t)} \subset \{1, \ldots, P\}$. Formally, let tree t have L_t leaves. The structure of the tree is captured by an indicator matrix $\Psi^{(t)} \in \{0,1\}^{N \times L_t}$ that routes each of the N observations to one of the L_t leaves. The prediction values for these leaves are stored in a vector $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^{L_t}$. The per-tree output is then:

$$f_t(\boldsymbol{X}) \;=\; \Psi^{(t)}(\boldsymbol{X}, C^{(t)})\, \boldsymbol{\beta}^{(t)}, \qquad \hat{\boldsymbol{y}} \;=\; \frac{1}{T_{RF}} \sum_{t=1}^{T_{RF}} \Psi^{(t)}(\boldsymbol{X}, C^{(t)})\, \boldsymbol{\beta}^{(t)}. \tag{5}$$

The set of split parameters for tree t, $C^{(t)}$ (comprising a feature index from $\{1, \ldots, P\}$ and a threshold in $\mathbb{R}$ for each internal node), is chosen greedily via recursive partitioning on the bootstrap sample $\mathcal{B}^{(t)}$, maximizing the reduction of node impurity [54]. Unlike single-tree CART pruning, RF typically grows unpruned trees (equivalently $\alpha = 0$ in the cost–complexity term):

$$(C^{(t)*}, \boldsymbol{\beta}^{(t)*}) = \arg \min_{C^{(t)}, \boldsymbol{\beta}^{(t)}} \left\| \boldsymbol{y}_{\mathcal{B}^{(t)}} - \Psi^{(t)}(\boldsymbol{X}_{\mathcal{B}^{(t)}}, C^{(t)})\, \boldsymbol{\beta}^{(t)} \right\|_2^2 \;+\; \alpha \left| T^{(t)} \right|, \tag{6}$$

where $|T^{(t)}| = L_t$ denotes the number of leaves and $\alpha \in \mathbb{R}^+$ is the cost–complexity coefficient. Out-of-bag (OOB) samples provide an internal, nearly unbiased generalization estimate (see Figure 6).

Building on the ensembling concept, XGBoost constructs an additive model in a stage-wise fashion. Key hyperparameters include the learning rate (shrinkage) $\eta \in (0, 1]$ and the number of boosting rounds T_{XGB} [55]. The prediction evolves as:

$$\hat{\boldsymbol{y}}^{(T_{XGB})} \;=\; \hat{\boldsymbol{y}}^{(T_{XGB}-1)} \;+\; \eta\, f_{T_{XGB}}(\boldsymbol{X}), \qquad \hat{\boldsymbol{y}}^{(T_{XGB})} \;=\; \sum_{t=1}^{T_{XGB}} \eta\, f_t(\boldsymbol{X}). \tag{7}$$

At iteration t, the learner f_t is found by minimizing a second-order approximation of the regularized objective $\mathcal{J}^{(t)} \in \mathbb{R}$:

$$\mathcal{J}^{(t)} \;=\; \sum_{n=1}^{N} \Big[g_n\, f_t(\boldsymbol{x}_n) + \tfrac{1}{2} h_n\, f_t^2(\boldsymbol{x}_n) \Big] \;+\; \Omega\big(\mathbf{w}^{(t)}\big), \qquad \Omega\big(\mathbf{w}^{(t)}\big) \;=\; \gamma L_t \;+\; \tfrac{\lambda}{2} \left\| \mathbf{w}^{(t)} \right\|_2^2. \tag{8}$$

Here, for each sample n, the scalars $g_n, h_n \in \mathbb{R}$ are the first and second-order derivatives of the loss with respect to the previous prediction $\hat{y}_n^{(t-1)}$. For a tree with L_t leaves, the

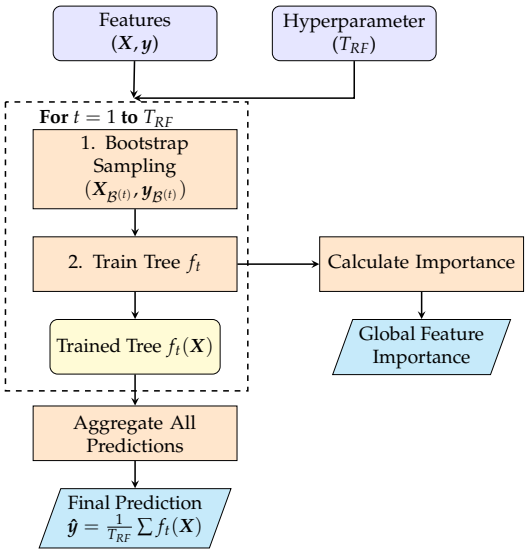

Figure 6. Conceptual pipeline for Random Forest regression: input data, bagging-based tree training, ensemble averaging for predictions, and derivation of global feature relevance

regularization Ω is controlled by the L2 coefficient $\lambda \in \mathbb{R}^+$ on leaf scores $\mathbf{w}^{(t)} \in \mathbb{R}^{L_t}$ and the complexity penalty $\gamma \in \mathbb{R}^+$. The split selection criterion (Gain) is derived as [56]:

$$\text{Gain} = \frac{1}{2}\left(\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda} \right) - \gamma, \tag{9}$$

where $G_{\{\cdot\}} = \sum g_n$ and $H_{\{\cdot\}} = \sum h_n$ represent the sum of gradients over samples in the left/right child nodes. The primary hyperparameters to be optimized are thus η, T_{XGB}, λ, and γ. A general schematic of the stage-wise procedure is shown in Figure 7.

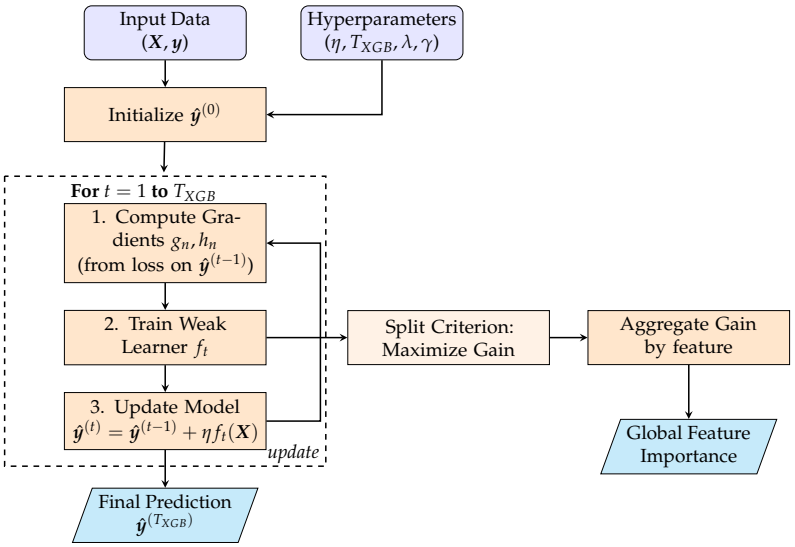

Figure 7. Stage-wise gradient boosting overview: initialization, per-iteration gradient computation, weak-learner fitting, additive model updates, and feature-wise gain aggregation.

In terms of interpretability, the mechanisms sketched above translate into well-defined global importance scores. In RF, global importance of a feature j is obtained by summing, across all trees t, the reduction in squared error produced at every split within the partition parameters $C^{(t)}$ that utilizes feature j. This process is directly tied to the training objective of minimizing $\left\| y_{\mathcal{B}^{(t)}} - \Psi^{(t)}\left(X_{\mathcal{B}^{(t)}}, C^{(t)}\right) \boldsymbol{\beta}^{(t)} \right\|_2^2$. In XGBoost, the analogous global importance for feature j is computed by accumulating the regularized split Gain dictated by the stage-

wise objective $\mathcal{J}^{(t)}$. This gain depends on the first- and second-order gradients, g_n and h_n, as well as the regularization parameters λ and γ; consequently, features repeatedly selected with high Gain receive larger global importance scores.

3.3. Deep Learning-based Tabular Data Regression with Localized Relevance Analysis

We now transition from classical estimators to deep learning architectures. In this setting, the prediction is generated by a parametric mapping defined as:

$$\hat{\boldsymbol{y}} = f(\boldsymbol{X};\boldsymbol{\Theta}) = \left(\check{f}_S \circ \check{f}_{S-1} \circ \cdots \circ \check{f}_1\right)(\boldsymbol{X}), \tag{10}$$

with $f : \mathbb{R}^{N\times P} \to \mathbb{R}^N$, $\check{f}_\ell$ denoting the s-th feature extractor, and $\boldsymbol{\Theta}$ the set of trainable parameters. This generic representation extends naturally to tabular data; in particular, TabNet realizes f as a composition that couples predictive performance with built-in explainability [29]. Its core mechanism is a sequence of S decision steps, as in Equation 10, that employs attention to select a sparse subset of features. At each step s, an attention mask $\boldsymbol{Z}^{(s)} \in \mathbb{R}^{N\times P}$ performs soft feature selection:

$$\boldsymbol{Z}^{(s)} = \text{sparsemax}(\boldsymbol{Q}^{(s-1)} \cdot \phi_s(\boldsymbol{c}^{(s-1)})). \tag{11}$$

This computation involves several components: $\boldsymbol{Q}^{(s-1)} \in \mathbb{R}^{N\times P}$ is a prior-scale matrix that tracks feature usage; $\boldsymbol{c}^{(s-1)} \in \mathbb{R}^{N\times N_a}$ is the processed feature representation from the previous step, with N_a as the attention embedding dimension; and $\phi_s : \mathbb{R}^{N\times N_a} \to \mathbb{R}^{N\times P}$ denotes a trainable mapping. The sparsemax activation is used to produce a sparse probability distribution, forcing the model to concentrate its attention on a limited subset of features [57]. The prior scale is updated recursively:

$$\boldsymbol{Q}^{(s)} = \prod_{j=1}^{s} (\nu - \boldsymbol{Z}^{(j)}), \tag{12}$$

where the scalar hyperparameter $\nu \in \mathbb{R}$ controls feature reuse. The masked features, $\boldsymbol{F}^{(s)} \in \mathbb{R}^{N\times P}$, are computed via an element-wise product, $\boldsymbol{F}^{(s)} = \boldsymbol{Z}^{(s)} \odot \boldsymbol{X}$, and are then processed by a feature transformer $\mathcal{F}$. This component employs Gated Linear Units (GLUs) as building blocks [58]:

$$\text{GLU}(\boldsymbol{h'}) = (\boldsymbol{W'}_1\boldsymbol{h'} + \boldsymbol{b}_1) \odot \sigma(\boldsymbol{W'}_2\boldsymbol{h'} + \boldsymbol{b}_2). \tag{13}$$

For an input vector $\boldsymbol{h'} \in \mathbb{R}^D$, $\boldsymbol{W'}_1, \boldsymbol{W'}_2 \in \mathbb{R}^{D\times D}$ are weight matrices, $\boldsymbol{b}_1, \boldsymbol{b}_2 \in \mathbb{R}^D$ are bias vectors, and σ is the element-wise sigmoid activation function. Residual connections are normalized by a factor of $\sqrt{0.5}$ to stabilize training. The transformer $\mathcal{F} : \mathbb{R}^{N\times P} \to (\mathbb{R}^{N\times N_d}, \mathbb{R}^{N\times N_a})$ takes the filtered features $\boldsymbol{F}^{(s)}$ and produces two outputs: an embedding for the final decision $\boldsymbol{d}^{(s)} \in \mathbb{R}^{N\times N_d}$ and a representation for the next step's attention $\boldsymbol{c}^{(s)} \in \mathbb{R}^{N\times N_a}$, where N_d is the decision embedding dimension.

For large-batch training, TabNet applies ghost batch normalization, splitting the batch into virtual mini-batches of size B_v for normalization [59]:

$$\tilde{\boldsymbol{X}} = \frac{\boldsymbol{X} - \boldsymbol{\mu}_{B_v}}{\sqrt{\boldsymbol{\sigma}^2_{B_v} + \epsilon}}, \tag{14}$$

where the vectors $\boldsymbol{\mu}_{B_v}, \boldsymbol{\sigma}^2_{B_v} \in \mathbb{R}^P$ denote the mean and variance computed over each virtual mini-batch, and ϵ is a small scalar for numerical stability. The overall decision

embedding is aggregated from all steps and mapped to the final prediction via a linear layer $\boldsymbol{W'}_{\text{final}} \in \mathbb{R}^{N_d}$:

$$\hat{\boldsymbol{y}} = \left(\sum_{s=1}^{S} \text{ReLU}(\boldsymbol{d}^{(s)}) \right) \boldsymbol{W'}_{\text{final}}. \tag{15}$$

The model is trained by minimizing a total loss $\mathcal{L}$, defined as $\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{sparse}} \mathcal{L}_{\text{sparse}}$, with the scalar $\lambda_{\text{sparse}} \in \mathbb{R}^{+}$ acting as the regularization coefficient. The task-specific loss for regression is typically the Mean Squared Error (MSE):

$$\mathcal{L}_{\text{task}} = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2, \tag{16}$$

while the sparsity regularization term encourages the model to focus on fewer features:

$$\mathcal{L}_{\text{sparse}} = -\frac{1}{N} \sum_{s=1}^{S} \sum_{n=1}^{N} \sum_{p=1}^{P} \boldsymbol{Z}_{n,p}^{(s)} \log\left(\boldsymbol{Z}_{n,p}^{(s)} + \epsilon \right). \tag{17}$$

In summary, the full TabNet processing pipeline is illustrated in Figure 8.

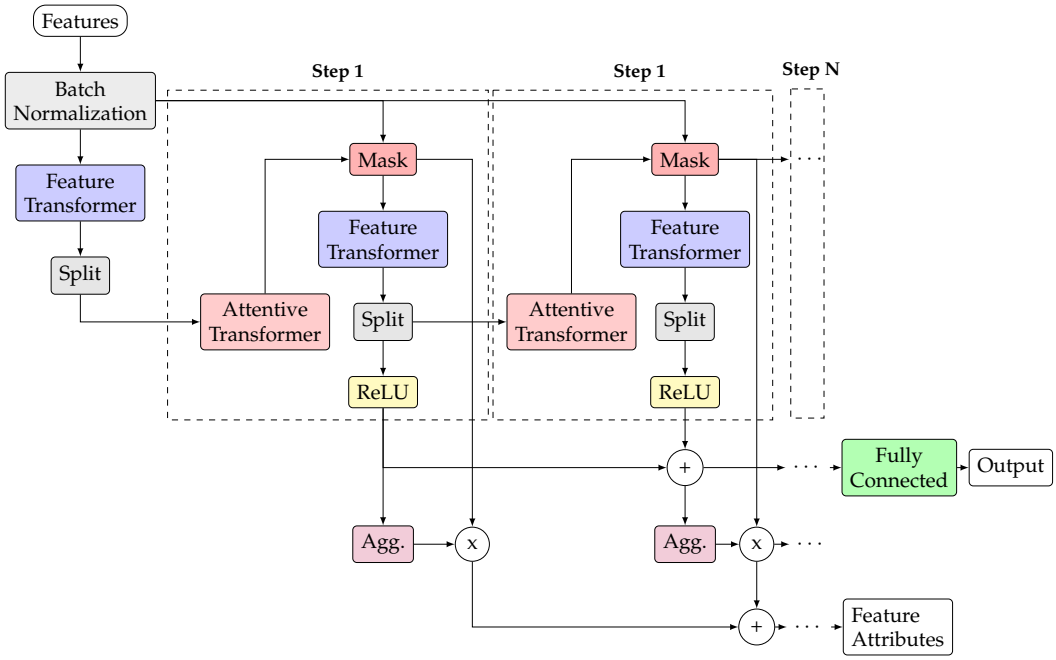

Figure 8. TabNet step-wise architecture with batch normalization, attentive masks, feature transformers, and residual aggregation; predictions are computed from aggregated features, while feature attributions derive from stepwise masks.

Next, building on the stepwise masks $\{\boldsymbol{Z}^{(s)}\}_{s=1}^{S}$ from the TabNet model, we obtain a unified feature relevance map through convex aggregation:

$$\boldsymbol{M} = \sum_{s=1}^{S} \zeta_s \boldsymbol{Z}^{(s)}; \quad \zeta_s \geq 0, \sum_{s=1}^{S} \zeta_s = 1. \tag{18}$$

The resulting matrix, $\boldsymbol{M} = \{M_{n,p} \in \mathbb{R} : n \in N, p \in P\}$, contains the aggregated relevance scores for each feature and reduces to a uniform average when $\zeta_s = \frac{1}{S}$. These scores are then mapped directly to a probability distribution over the features for each sample using a temperature-controlled softmax [60]:

$$\pi_{n,p} = \frac{\exp\left(M_{n,p}/\tau\right)}{\sum_{p'=1}^{P} \exp\left(M_{n,p'}/\tau\right)}, \qquad \sum_{p=1}^{P} \pi_{n,p} = 1. \tag{19}$$

Let $\mathbf{\Pi} = \{\pi_{n,p} \in \mathbb{R} : n \in N, p \in P\}$ be the matrix of localized relevance scores. To derive a feature importance ranking for any subset of data, we define an aggregation function $\delta_p : \mathbb{R}^{N \times P} \to \mathbb{R}^+$. Given a set of sample indices of interest, $\mathcal{D} \subseteq \{1, \ldots, N\}$, this function is defined as:

$$\delta_p(\mathbf{\Pi}, \mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \pi_{d,p}. \tag{20}$$

The resulting vector, $\boldsymbol{\delta} = \{\delta_p(\mathbf{\Pi}, \mathcal{D}) \in \mathbb{R}^+ : p \in P\}$, represents the final feature importance profile for the specified data subset. This unified formulation provides importance rankings at any desired scale. For a local analysis of a single sample n, we set $\mathcal{D} = \{n\}$, yielding the original localized profile. For a global analysis, we set $\mathcal{D} = \{1, \ldots, N\}$, yielding the dataset-level feature ranking. Furthermore, the sharpness of the underlying individual explanations can be quantified via the Shannon entropy of each relevance vector $\boldsymbol{\pi}_d = \{\pi_{d,p} \in \mathbb{R}^+ : p \in P\}$, given by:

$$H(\boldsymbol{\pi}_d) = -\sum_{p=1}^{P} \pi_{d,p} \log \pi_{d,p}, \tag{21}$$

where low entropy indicates a sharp and highly focused attribution of importance.

3.4. Fundamentals of Retrieval-Augmented Generation and Agentic Systems

Large Language Models (LLMs) exhibit two core limitations: their knowledge is static, fixed at the time of their last training, and they are prone to generating incorrect information, or "hallucinations," when operating outside their knowledge domain [61]. To mitigate these challenges and engineer more reliable, evidence-based systems, architectures have been developed to integrate external knowledge in real-time [62]. The foundational approach is Retrieval-Augmented Generation (RAG), which operates in two primary stages (Figure 9) [63]. First, during the retrieval phase, the system queries an external knowledge base to locate relevant information fragments pertinent to the query [64]. Subsequently, in the generation phase, these fragments are supplied to the LLM as context alongside the original question, thereby grounding the response in verifiable evidence and reducing hallucinations [65].

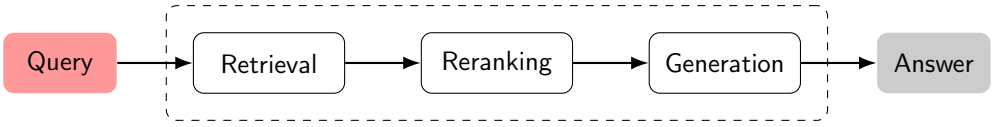

Figure 9. Linear workflow of a traditional RAG system.

Classical RAG operates in a linear, single-step fashion. While this framework is suitable for direct questions, its utility is limited when the task demands multi-step reasoning or the integration of heterogeneous sources [66]. To address these scenarios, the *Agentic RAG* paradigm has been proposed (see Figure 10) [67]. This approach redefines the LLM's role: it transitions from a context-conditioned generator to an agent capable of reasoning, planning, and acting [68]. Instead of adhering to a fixed workflow, an agentic system dynamically determines which actions to execute in order to holistically resolve complex tasks.

The transition to an agentic system is predicated on reassigning the LLM's role from a response generator to a reasoning engine [69]. The agent functions as a cognitive core, designed to decompose complex tasks into logical, executable steps [70]. When presented with a problem, it formulates a dynamic plan that determines what information is required, from which sources it should be obtained, and in what sequence it must be processed to construct a well-founded solution.

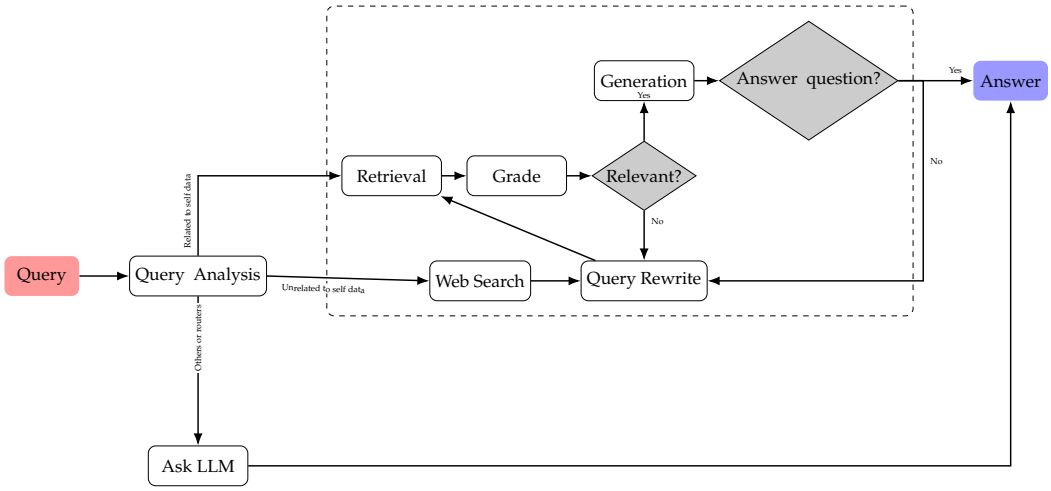

Figure 10. Cyclical and adaptive workflow of an Agentic RAG system.

To execute this plan, the agent is equipped with tools that enable it to interact with its environment and overcome the limitations of its pretrained knowledge. Beyond the textual search characteristic of classical RAG, the agent can invoke specialized functions: database connectors for SQL queries on structured data, code interpreters for quantitative analysis, or APIs for integration with external software systems. This allows it to orchestrate the retrieval and processing of heterogeneous information—both qualitative and quantitative—in a coordinated manner [71]. Lastly, the value of the agentic approach lies in its iterative operation—the reason-act-observe loop. Unlike a linear workflow, the agent executes an action, observes the outcome, and uses that evidence to inform its next step, adjusting its strategy as necessary [72]. This process is repeated to explore alternatives, corroborate findings, and accumulate evidence until sufficient inputs are gathered to synthesize a coherent final response. Then, the method generates an auditable trail of reasoning, reflected in the sequence of actions that led to the conclusion [70].

3.5. Criticality Analysis through Interpretable AI using Agentic RAG and LLM's

To leverage the comprehensive dataset, we developed an integrated diagnostic framework grounded in the Model–View–Controller (MVC) architectural pattern [73]. The system transitions from event selection to predictive analysis, culminating in an explainable, regulation-grounded recommendation for fault diagnosis. The framework comprises two main stages: an interactive analysis interface and a predictive recommendation engine.

The view and controller components provide a user-centric interface for spatiotemporal analysis, as illustrated in Figure 11. The workflow begins when the user specifies a geographic area of interest (department and municipality) and a time window (year and month) via interactive filters. In response, the system renders the corresponding MV-L2 network and lists all recorded interruption events within the selected period. The user then selects an event for detailed analysis. Upon selection, the controller invokes a downstream-tracing algorithm to identify all network assets—including poles, transformers, switches, and line segments—that are electrically connected beyond the operated protective device. This initial stage delineates a focused set of candidate components pertinent to the fault, which proceeds directly to predictive analysis.

From this focused set, the information is structured according to the granular root-cause database schema and ingested into a TabNet-based predictive model. This model has two simultaneous objectives: (i) to estimate a quality index associated with each asset, thereby quantifying their expected contribution to service degradation—after which the three assets with the largest contributions are selected as the most likely candidates

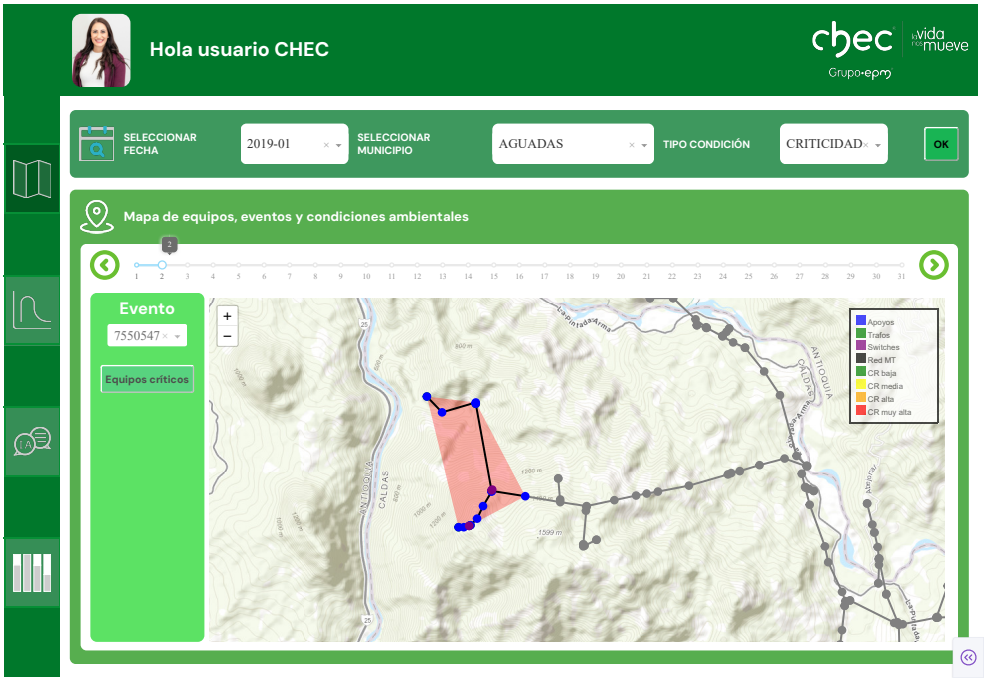

Figure 11. The user interface of the diagnostic framework. The top panel allows users to filter events by date and municipality.

responsible for the interruption; and (ii) to derive post hoc feature relevance from TabNet's masks without introducing an auxiliary interpretability loss to compute aggregate relevance. For each selected asset, an importance ranking is obtained, and the five most influential structural and exogenous variables are retained. This refined information becomes the primary input to an LLM-based recommendation agent.

The agent initiates an Agentic RAG process. Leveraging a specialized document corpus, it autonomously formulates queries over the embedded knowledge base comprising RETIE, NTC, and CHEC's internal specifications. This corpus is augmented with a set of asset-specific, structured transition documents that map structural and exogenous variables to specific sections of each unstructured source. This mapping layer enables precise retrieval and anchoring of normative evidence conditioned on the prioritized assets and variables. The workflow issues targeted queries, filters by clause and numeral identifiers, expands terminology when gaps are detected (synonyms and cross-references), and promotes only evidence corroborated across independent sources with consistent wording and scope, anchoring each conclusion to explicit citations. This ensures that the analysis is not solely driven by predictive signals but is firmly contextualized within established regulatory and engineering standards. The agent imposes scope limits by restricting conclusions to the retrieved standards and activates an insufficient-evidence mode when corroboration thresholds are not met. The output is a set of technical conclusions explicitly supported by cited clauses and the specific technical context corresponding to the high-likelihood assets and their influential variables.

To ensure full transparency and auditability, the entire decision path is synthesized into a structured and interpretable reasoning graph. This graph serves as a formal record of the diagnostic process, mapping the initial predictive outputs from the TabNet model, the retrieved regulatory evidence, and the intermediate inferential steps taken by the LLM agent. Each node represents a unit of information—such as a prioritized asset, an influential variable, or a specific regulatory clause—while edges encode the logical relations among

them. Each node and edge stores the source identifier, document version, and section anchor, providing end-to-end evidence attribution. As a final output, the system issues a coherent and traceable natural-language recommendation, accompanied by the reasoning graph and the corresponding regulatory citations.

The integrated process—combining the user interface, predictive modeling, and regulation-based reasoning—is summarized in Figure 12.

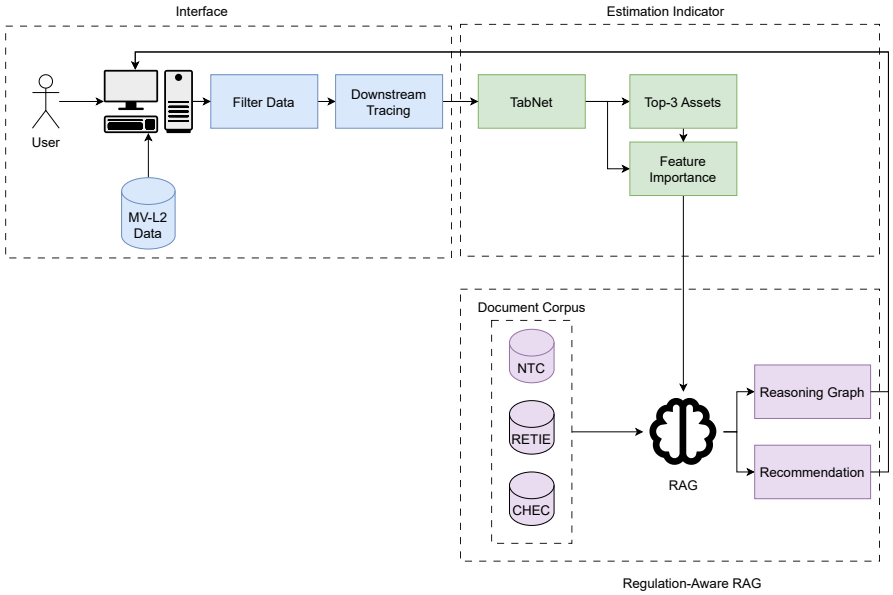

Figure 12. Architectural diagram of the integrated diagnostic framework based on interpretable AI for reliability and regulation-aware decision support.

4. Experimental Setup

4.1. Assessment and Method Comparison

The evaluation of our dual-component framework is systematically structured into two distinct parts, addressing the predictive accuracy of the failure indicator estimation and the qualitative performance of the generative recommendation system, respectively.

Assessment of failure indicator prediction to assess the efficacy of our TabNet-based prediction model and its supervised relevance analysis, its outcomes are benchmarked against a suite of well-established techniques:

– *Linear Machine Learning:* ElasticNet, which utilizes a combination of L1 and L2 regularization to improve generalization and facilitate variable selection in high-dimensional contexts [74].
– *Nonlinear Machine Learning:* RF and XGBoost are included as benchmarks. RF is known for its ability to capture intricate interactions and nonlinearities through ensemble learning, while XGBoost is regarded for its state-of-the-art performance on structured tabular data via an optimized gradient boosting framework [75,76].

The performance of these supervised models is evaluated using standard regression metrics, contrasting the reference values $\boldsymbol{y}$ with the predictions $\hat{\boldsymbol{y}}$. Let $\bar{\boldsymbol{y}} = \mu_y \mathbf{1}$ denote the mean reference vector, where $\mu_y = \frac{1}{N}\sum_{n=1}^{N} y_n$ and $\mathbf{1}$ is the all-ones vector in $\mathbb{R}^N$. These metrics are defined as follows:

$$R^2(\boldsymbol{y}, \hat{\boldsymbol{y}}) = 1 - \frac{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2}{\|\boldsymbol{y} - \bar{\boldsymbol{y}}\|_2^2}, \tag{22}$$

$$MSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_2^2, \tag{23}$$

$$MAE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|_1, \tag{24}$$

$$MAPE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{25}$$

For the second stage of our framework, this study evaluates a range of only–decoder LLMs on a specialized question–answering task designed to support CHEC's operational and normative queries [77]. The selection includes both proprietary, API-based models and open-source, locally deployable models to provide a comparison between cloud and on-premise inference capabilities [78]. The evaluated set was deliberately constructed to span diverse computational scales—ranging from lightweight models with one billion parameters to large-scale systems with tens of billions of parameters —enabling the analysis of trade-offs between inference efficiency, reasoning depth, and domain adaptation [79]. Given the computational capacity available for local deployment, the configuration emphasizes models that balance representational complexity with efficient quantized implementations, thereby enabling meaningful contrasts between more compact on-premise systems and high-capacity cloud counterparts [80]. In selecting the models, we included prominent transformers from a variety of leading developers to capture a representative snapshot of the current landscape. Table 2 summarizes the configuration of all evaluated LLMs.

Table 2. Overview of key characteristics for the LLMs selected for evaluation.

LLM	#Params	Context Length	Max Tokens	Quantization
gpt-3.5-turbo[81]	Not disclosed	16,385	16,385	Not disclosed
gpt-4o [82]	Not disclosed	128,000	128,000	Not disclosed
gemini-2.0 [83]	40B	1,048,576	8,192	Not disclosed
gemini-2.5 [84]	Not disclosed	1–2M	65,535	Not disclosed
llama-3.1-8b [85]	8B	128,000	Not specified	4 bits
llama-3.2-1b [86]	1B	128,000	8,000	4 bits
qwen-2.5-1.5b [87]	1.5B	32,768	8,192	8 bits
qwen-2.5-7b[88]	7B	131,072	8,000	16 bits
deepseek-r1-7b [89]	7B	128,000	32,768	4 bits
deepseek-r1-1.5b [90]	1.5B	128,000	32,768	4 bits

To benchmark the selected models, we constructed an expert-curated Q&A corpus comprising 53 challenges that reflect operational information-retrieval and decision-support needs in MV-L2 distribution. Tasks are organized into three groups: (i) 19 structured queries over tabular assets and event logs; (ii) 19 unstructured normative queries requiring comprehension and grounding in technical standards and internal specifications; and (iii) a recommendation task instantiated on three real-world assets, each parameterized by five critical variables, yielding 15 recommendation outputs. This taxonomy separates modality (structured vs. unstructured) and decision focus, enabling consistent comparison across models. Table 3 presents one representative example from each task category, illustrating the diversity and structure of the evaluation corpus.

To quantify the performance of the generative models, two metrics were employed. Primarily, BERTScore was utilized to assess semantic quality by computing the similarity between contextual embeddings of the generated and reference responses. To ensure linguistic consistency with the bilingual domain of the CHEC dataset, the multilingual

Table 3. Representative examples from the question-answering (Q&A) dataset

Query Type	Example Question	Reference Answer
Unstructured normative query	¿Qué tipo de aislador se recomienda en zonas con alto nivel de contaminación?	Recomendación técnica: Para instalaciones ubicadas en zonas con un nivel de contaminación alto (Grado III, que exige al menos 25 mm/kV de distancia de fuga), se recomienda utilizar aisladores poliméricos sobredimensionados o incrementar la cadena de aislamiento para asegurar una mayor distancia de fuga. Una práctica habitual es emplear aisladores de 34.5 kV en redes de 13.2 kV o adicionar más unidades en el caso de cadenas de suspensión. Los aisladores poliméricos ofrecen mejor comportamiento en ambientes con alta polución debido a su mayor resistencia a la acumulación de contaminantes y su buen desempeño mecánico. Normativas aplicables: • Según la norma IEC 60071-2, Sección de coordinación de aislamiento, Tabla 1, páginas mostradas en el contexto, el nivel de contaminación alto (Grado III) requiere 25 mm/kV de distancia de fuga mínima. • De acuerdo con el RETIE 2013, Artículo 3.17.2, se exige que los aisladores cumplan con la compatibilidad de dimensiones y accesorios, y que cuenten con su respectivo certificado de conformidad. • La Publicación IEC 720 describe los requisitos de resistencia mecánica y comportamiento dieléctrico para aisladores tipo Line Post, validando su aplicación ante altas solicitaciones en entornos contaminados.
Structured query	¿Cuántas interrupciones hubo entre el año 2019 y el año 2023?	Entre el año 2019 y el año 2023 se registró un total de 148,041 interrupciones, según los registros históricos de la base de datos de eventos de red.
Recommendation query	transformador_h6-rh — Humedad Relativa: 81%	Recomendación técnica integral conforme a RETIE, NTC 2050, IEC 60076 y lineamientos de mantenimiento CHEC. En ambientes con humedad relativa superior al 80%, se recomienda instalar deshumidificadores o deshidratadores de aire (gel de sílice), utilizar envolventes con protección IP55 o superior, aplicar recubrimientos anticorrosivos, incorporar sistemas de calefacción interna o ventilación forzada con filtros antihumedad, y verificar periódicamente la hermeticidad del tanque y las conexiones. Las pruebas de aislamiento y estanqueidad deben realizarse conforme a IEC 60076-1, IEEE C57.152 y ASTM D877.

case-sensitive BERT model was adopted [91]. Let the reference response be denoted by the token sequence $\boldsymbol{x}' = \{x'_{n'}\}_{n'=1}^{N'}$ and the candidate response by $\hat{\boldsymbol{x}}' = \{\hat{x}'_{n'}\}_{n'=1}^{N'}$, where N' represents the aligned length of both sequences. Furthermore, let $\mathcal{V}$ be the WordPiece subword vocabulary of the tokenizer; consequently, for all n', it holds that $x'_{n'}, \hat{x}'_{n'} \in \mathcal{V}$. A contextual embedding mapping is defined as $E : \mathcal{V} \to \mathbb{R}^d$. Assuming the embeddings are pre-normalized to a unit norm, the cosine similarity is equivalent to their dot product. It is from this property that BERTScore is decomposed into three components—Precision, Recall, and F_1. In the specific context of regulatory compliance, these metrics provide distinct diagnostic insights: Precision quantifies the model's ability to avoid hallucinations (i.e., minimizing the fabrication of non-existent regulations), while Recall assesses the completeness of the answer (i.e., ensuring no critical normative details are omitted). Finally, the F_1 score offers a holistic measure of semantic alignment. These components are calculated from the cosine similarities between the vector representations of both sequences as:

$$P_{\text{BERT}} = \frac{1}{N'} \sum_{j=1}^{N'} \max_{1 \le i \le N'} E(x_i)^T E(\hat{x}_j), \tag{26}$$

$$R\text{BERT} = \frac{1}{N'} \sum_{i=1}^{N'} \max_{1 \le j \le N'} E(x_i)^T E(\hat{x}_j), \tag{27}$$

$$\text{F1} = 2\frac{P\text{BERT} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \tag{28}$$

Complementing the assessment of semantic quality, the second metric, inference time, was used to measure computational efficiency. This is defined as the average time required to generate a complete answer and was evaluated exclusively on locally deployed models to ensure a fair comparison of computational overhead, independent of network latency.

4.2. Training and Implementation Details

As a preliminary quality-control step, records with durations exceeding 100 hours were discarded to reduce the influence of extreme outliers during model fitting. From an initial set of 314 candidate columns, we excluded the continuity index SAIFI from the predictor space, yielding a modeling matrix with 312 predictors ($\boldsymbol{X}$). Missing numerical entries were imputed using a distribution-aware sentinel defined as $-10.0 \times \max(\text{column})$, which preserves scale while making imputed values explicitly distinguishable during learning. Categorical variables were label-encoded using scikit-learn v1.6.1. The targets were normalized to a fixed range with a MinMaxScaler to standardize the optimization objective across models. To ensure robust estimation and evaluation, we adopted a dual validation strategy. First, to explicitly evaluate model generalization under temporal drift and evolving environmental conditions, we implemented a time-aware rolling window cross-validation scheme. Starting from January 1, 2019, this protocol employed a moving 12-month training window to predict the subsequent 6-month testing horizon, shifting the window forward in 6-month increments throughout the study period. This approach allows for the assessment of predictive stability against seasonal shifts and asset aging. Second, to provide a standard aggregate performance benchmark, we utilized a randomized two-stage split: first, an 80/20 train–test partition; second, an 80/20 split of the training fold to obtain a validation subset. Both partitions used stratified sampling over target quartiles to preserve outcome distributions across folds.

All predictive models were tuned via Bayesian optimization with a Gaussian-process surrogate using Optuna v3.5.0, minimizing $1 - R^2$ to align the search with maximization of R^2. Each study executed 20 trials per model. The search spaces were specified as follows:

- *ElasticNet:* The maximum number of iterations was set as an integer value within the range [500, 3000], while the l_1-ratio was defined as a continuous value over [0.05, 0.95]. The regularization coefficient α and the stopping criterion tolerance were drawn from a log-uniform distribution over the ranges $[10^{-4}, 10^{1}]$ and $[10^{-6}, 10^{-3}]$, respectively.
- *Random Forest:* The following hyperparameters were configured with integer values: the number of estimators in [1, 100], the maximum tree depth in [2, 24], the minimum samples per leaf in [1, 10], and the minimum samples required for a split in [2, 20]. Additionally, the fraction of features considered at each split was set as a continuous value over the interval [0.4, 1.0].
- *XGBoost:* The maximum depth and the number of boosting rounds were set as integer values within the ranges [2, 24] and [1, 100], respectively. The subsample ratio and the per-tree column subsampling ratio were defined as continuous values within [0.6, 1.0] and [0.5, 1.0]. Finally, the learning rate η, the ℓ_1 penalty, and the ℓ_2 penalty were drawn from a log-uniform distribution over $[10^{-3}, 0.3]$, $[10^{-6}, 1.0]$ and $[10^{-6}, 10.0]$.
- *TabNet:* Architectural hyperparameters for feature dimensionality (n_d), attention output dimensionality (n_a), and the number of steps were set as integer values within the ranges [8, 128], [8, 128], and [2, 10], respectively. Regularization parameters (the γ coefficient and the sparsity coefficient λ_{sparse}) and optimizer settings (learning rate and weight decay) were drawn from log-uniform distributions over the ranges $[10^{-6}, 2]$, $[10^{-6}, 0.9]$, $[10^{-3}, 10^{-1}]$, and $[10^{-4}, 10^{-1}]$, respectively. Categorical hyperparameters were selected from fixed sets: the masking function from {entmax, sparsemax}; batch size from {1024, 2048, 4096}; virtual batch size from {512, 1024, 2048}; and the op-

timizer from {Adam, AdamW, SGD, RMSprop}. To enforce non-negativity on the SAIDI/SAIFI predictions, a ReLU activation function was applied to the final output layer. During the TabNet search, each configuration was trained for up to 40 epochs with an early-stopping patience of 40. Following model selection, the best-performing configuration was retrained on the pooled training and validation data; specifically for TabNet, this final training phase ran for 200 epochs with a patience of 70. The test performance for all models was subsequently evaluated on the hold-out set.

For the RAG-based generative agent, the evaluation methodology was specifically designed to ensure reproducibility and consistent behavior across all tested systems. To this end, a deterministic output is enforced by setting the temperature parameter to 0, while other generative hyperparameters, such as top_p, top_k, and any repetition penalties, remain at their default values as specified by their respective APIs.

Furthermore, a standardized zero-shot prompt template is employed for all queries. Context is injected using the stuff chain type, which concatenates the five most relevant document chunks retrieved from the vector database and inserts them directly into the prompt. Crucially, to maintain consistent grounding granularity across the regulatory corpus, a page-level chunking strategy was implemented: each document was segmented into one chunk per page, with a fixed overlap of 200 tokens between adjacent segments. The retrieval process is underpinned by vector embeddings generated using OpenAI's text-embedding-ada-002 model, with all vectors stored and queried from a persistent Chroma vector database [92].

The agent's operational workflow unfolds in a structured sequence. Upon receiving a user query, a primary dispatching agent, powered by gpt-3.5-turbo, first analyzes the input and selects the most appropriate tool from a predefined set based on its semantic description. Upon invocation, the selected tool executes the RAG pipeline: it queries its dedicated, domain-specific vector store to retrieve the five most relevant document chunks. These chunks are subsequently compiled into a context that is passed to the designated generative model under evaluation, which then synthesizes the final textual response. This entire sequence is performed for each question in the evaluation corpus to generate the final results.

Experiments were executed in two complementary environments. The predictive pipeline ran on Google Colab with an NVIDIA (Santa Clara, CA, USA) A100 (40.0GB VRAM) and 83.5GB RAM. The generative evaluation was conducted on a local workstation running Ubuntu22.04, equipped with an Intel Core i9-11900 CPU, 64GB of RAM, and an NVIDIA (Santa Clara, CA, USA) RTX 3070 Ti GPU (8GB VRAM). All experiments used Python 3.12 with a global random seed of 42, NumPy v2.0.2, and PyTorch v2.8.0. For deterministic reproducibility, we enabled cuDNN v91002 deterministic kernels where applicable and disabled non-deterministic algorithms in PyTorch. Core libraries for the predictive pipeline included cuML v25.06.00, cuPy v13.3.0, XGBoost v3.1.1, and pytorch-tabnet v4.1.0. The generative stack was orchestrated using the LangChain v0.3.3 framework and its associated libraries, including langchain-openai v0.2.2, langchain-google-genai v2.0.0, and chromadb v0.5.12. Open-source models locally executed via the Ollama v0.5.3 runtime. Source code and datasets are available at https://github.com/UN-GCPDS/CRITAIR (accessed on October 30, 2025).

5. Results and Discussion

5.1. Predictive Performance for Reliability Indicator Estimation

The predictive capabilities of the proposed framework were evaluated across three hierarchical granularities—global, municipal, and feeder-level—to quantify both aggre-

gate performance and stability under conditions representative of operational network management.

Prior to numerical benchmarking, we examine the decomposition of the SAIFI signal (Figure 13, left) to formally ground the evaluation methodology. The presence of a stable seasonal component provides direct justification for selecting a six-month rolling window, ensuring that the validation protocol captures semi-annual operational periodicity. Moreover, the trend component reveals a structural regime shift accompanied by increased volatility from 2023 onward. To explicitly accommodate this non-stationary and the anticipated drift in both asset performance and meteorological conditions, the study prioritizes a time-aware evaluation strategy, complemented by a standard stratified randomized split for comparative reference.

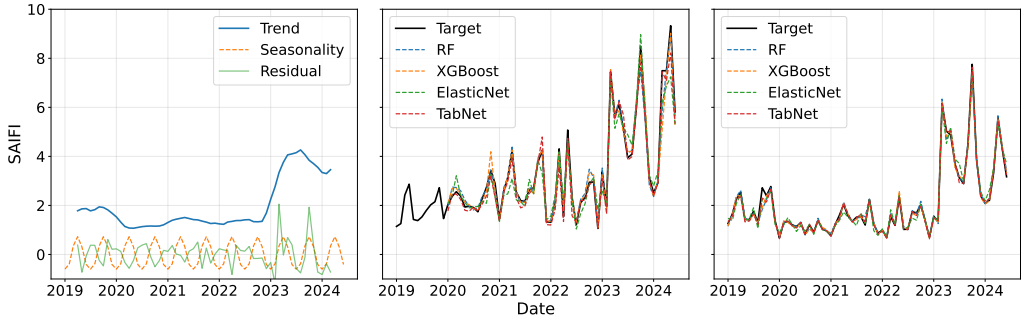

Figure 13. Time-series analysis and comparison of SAIFI forecasts against observed values. **Left:** Decomposition of the historical SAIFI signal into trend, seasonality, and residual components. **Middle:** Model forecasts versus observed targets using the time-aware rolling window validation scheme to assess performance under temporal drift. **Right:** Model forecasts versus observed targets using the standard stratified randomized split.

At the global resolution—under the time-aware rolling-window configuration (Table 4, top)—TabNet exhibits strong resilience to temporal drift, yielding the highest variance explanation ($R^2 = 0.83$) and the lowest absolute error (MAE $= 3.5 \times 10^{-4}$). Although Random Forest remains competitive in terms of relative percentage error (MAPE $= 5.9 \times 10^{1}$%), TabNet maintains superior control over absolute deviation. A complementary pattern emerges in the randomized-split scenario (Table 4, bottom), where the relaxation of temporal constraints enables TabNet to reach its peak performance ($R^2 = 0.93$, MSE $= 1.5 \times 10^{-5}$), outperforming XGBoost ($R^2 = 0.86$). Across both validation regimes, the results highlight the limitations of linear baselines such as ElasticNet ($R^2 \approx 0.63$˜0.71) in modeling the nonlinear structure of SAIFI dynamics. A qualitative comparison in Figure 13 further reinforces these findings: under both evaluation schemes, TabNet's forecasts closely follow observed behavior, particularly during abrupt excursions linked to elevated network stress conditions, whereas alternative models demonstrate delayed or smoothed response.

This aggregate performance, specifically under the randomized split strategy, is corroborated at finer resolutions. When disaggregated to the five municipalities with the highest SAIFI contribution (see Table 5), TabNet consistently outperforms or matches the benchmarks. For instance, in "La Dorada" and "Manizales," it secures superior R^2 values and minimal errors, underscoring that its high accuracy is not merely an artifact of aggregation but is sustained in high-priority operational zones. This robustness extends to the most granular scale—the distribution feeder level (Table 6), where the model accounts for nearly all the variance in critical circuits such as "ROS23L15" ($R^2 \approx 1.0$). This level of precision validates its use for prioritizing maintenance and planning localized capital investments.

The stability of the models was further confirmed through Bayesian hyperparameter optimization. The optimization landscapes in Figure 14 reveal that the selected configura-

Table 4. Comparative evaluation of predictive models for SAIFI estimation: Time-Aware Rolling Window vs. Standard Randomized Split.

Validation Method	Model	R^2	MSE	MAE	MAPE [%]
Time-Aware Split	ElasticNet	6.3×10^{-1}	3.0×10^{-6}	7.8×10^{-4}	1.4×10^{2}
	RandomForest	7.6×10^{-1}	$\mathbf{2.0 \times 10^{-6}}$	3.8×10^{-4}	$\mathbf{5.9 \times 10^{1}}$
	XGBoost	8.1×10^{-1}	$\mathbf{2.0 \times 10^{-6}}$	3.5×10^{-4}	6.9×10^{1}
	TabNet	$\mathbf{8.3 \times 10^{-1}}$	$\mathbf{2.0 \times 10^{-6}}$	$\mathbf{3.5 \times 10^{-4}}$	8.4×10^{1}
Randomized Split	ElasticNet	7.1×10^{-1}	6.6×10^{-5}	3.4×10^{-3}	1.4×10^{2}
	RandomForest	7.9×10^{-1}	4.7×10^{-5}	8.1×10^{-4}	$\mathbf{3.9 \times 10^{1}}$
	XGBoost	8.6×10^{-1}	3.0×10^{-5}	7.6×10^{-4}	5.2×10^{1}
	TabNet	$\mathbf{9.3 \times 10^{-1}}$	$\mathbf{1.5 \times 10^{-5}}$	$\mathbf{6.8 \times 10^{-4}}$	6.4×10^{1}

Table 5. Disaggregated predictive performance across the five municipalities contributing most significantly to SAIFI.

Municipality	Model	R^2	MSE	MAE	MAPE [%]
DOSQUEBRADAS	RandomForest	9.6×10^{-1}	4.4×10^{-5}	$\mathbf{2.4 \times 10^{-3}}$	$\mathbf{5.1 \times 10^{1}}$
	XGBoost	9.6×10^{-1}	5.3×10^{-5}	2.5×10^{-3}	5.6×10^{1}
	ElasticNet	5.6×10^{-1}	5.4×10^{-4}	1.2×10^{-2}	1.2×10^{2}
	TabNet	$\mathbf{9.6 \times 10^{-1}}$	$\mathbf{4.3 \times 10^{-5}}$	2.5×10^{-3}	6.5×10^{1}
MANIZALES	RandomForest	4.5×10^{-1}	3.6×10^{-4}	1.5×10^{-3}	$\mathbf{4.8 \times 10^{1}}$
	XGBoost	6.7×10^{-1}	2.1×10^{-4}	1.4×10^{-3}	6.2×10^{1}
	ElasticNet	7.6×10^{-1}	1.5×10^{-4}	4.8×10^{-3}	1.3×10^{2}
	TabNet	$\mathbf{8.5 \times 10^{-1}}$	$\mathbf{1.0 \times 10^{-4}}$	$\mathbf{1.3 \times 10^{-3}}$	7.4×10^{1}
LA DORADA	RandomForest	8.9×10^{-1}	2.5×10^{-5}	1.1×10^{-3}	$\mathbf{5.4 \times 10^{1}}$
	XGBoost	9.2×10^{-1}	1.6×10^{-5}	9.5×10^{-4}	6.7×10^{1}
	ElasticNet	5.2×10^{-1}	1.1×10^{-4}	4.7×10^{-3}	1.5×10^{2}
	TabNet	$\mathbf{9.5 \times 10^{-1}}$	$\mathbf{1.1 \times 10^{-5}}$	$\mathbf{9.0 \times 10^{-4}}$	7.9×10^{1}
CHINCHINÁ	RandomForest	8.4×10^{-1}	4.6×10^{-5}	1.5×10^{-3}	$\mathbf{4.0 \times 10^{1}}$
	XGBoost	7.6×10^{-1}	7.1×10^{-5}	1.5×10^{-3}	5.3×10^{1}
	ElasticNet	4.5×10^{-1}	1.6×10^{-4}	7.5×10^{-3}	1.4×10^{2}
	TabNet	$\mathbf{9.1 \times 10^{-1}}$	$\mathbf{2.5 \times 10^{-5}}$	$\mathbf{1.3 \times 10^{-3}}$	5.9×10^{1}
VILLAMARÍA	RandomForest	9.1×10^{-1}	1.7×10^{-5}	1.3×10^{-3}	$\mathbf{5.2 \times 10^{1}}$
	XGBoost	9.4×10^{-1}	$\mathbf{1.0 \times 10^{-5}}$	$\mathbf{1.0 \times 10^{-3}}$	6.7×10^{1}
	ElasticNet	5.5×10^{-1}	7.9×10^{-5}	5.3×10^{-3}	1.4×10^{2}
	TabNet	$\mathbf{9.4 \times 10^{-1}}$	1.0×10^{-5}	1.0×10^{-3}	7.1×10^{1}

tions (marked with 'X') occupy broad, high-performance regions. This suggests that the reported performance is robust and not contingent on hypersensitive parameter tuning.

Table 6. Feeder-level predictive performance for the five distribution circuits with the highest SAIFI.

Feeder	Model	R^2	MSE	MAE	MAPE [%]
ROS23L15	RandomForest	9.9×10^{-1}	3.1×10^{-5}	2.3×10^{-3}	$\mathbf{5.3 \times 10^{1}}$
	XGBoost	9.9×10^{-1}	2.8×10^{-5}	2.0×10^{-3}	6.3×10^{1}
	ElasticNet	5.1×10^{-1}	1.5×10^{-3}	1.9×10^{-2}	1.2×10^{2}
	TabNet	$\mathbf{1.0 \times 10^{0}}$	$\mathbf{1.9 \times 10^{-5}}$	$\mathbf{2.0 \times 10^{-3}}$	7.3×10^{1}
BQE23L12	RandomForest	$\mathbf{9.9 \times 10^{-1}}$	$\mathbf{8.0 \times 10^{-6}}$	$\mathbf{1.2 \times 10^{-3}}$	$\mathbf{3.6 \times 10^{1}}$
	XGBoost	9.5×10^{-1}	4.5×10^{-5}	1.9×10^{-3}	4.2×10^{1}
	ElasticNet	5.5×10^{-1}	3.7×10^{-4}	1.2×10^{-2}	1.2×10^{2}
	TabNet	9.8×10^{-1}	1.4×10^{-5}	1.4×10^{-3}	5.4×10^{1}
ROS23L16	RandomForest	9.8×10^{-1}	2.7×10^{-5}	$\mathbf{2.1 \times 10^{-3}}$	$\mathbf{4.3 \times 10^{1}}$
	XGBoost	9.6×10^{-1}	6.4×10^{-5}	2.9×10^{-3}	4.9×10^{1}
	ElasticNet	4.8×10^{-1}	8.4×10^{-4}	1.5×10^{-2}	1.1×10^{2}
	TabNet	$\mathbf{9.7 \times 10^{-1}}$	$\mathbf{4.6 \times 10^{-5}}$	3.0×10^{-3}	6.2×10^{1}
ROS23L14	RandomForest	9.8×10^{-1}	1.9×10^{-5}	1.9×10^{-3}	$\mathbf{4.4 \times 10^{1}}$
	XGBoost	9.7×10^{-1}	3.8×10^{-5}	2.5×10^{-3}	5.0×10^{1}
	ElasticNet	4.5×10^{-1}	6.7×10^{-4}	1.4×10^{-2}	1.0×10^{2}
	TabNet	$\mathbf{9.9 \times 10^{-1}}$	$\mathbf{1.1 \times 10^{-5}}$	$\mathbf{1.9 \times 10^{-3}}$	5.6×10^{1}
DOR23L14	RandomForest	9.9×10^{-1}	5.0×10^{-6}	1.1×10^{-3}	$\mathbf{5.6 \times 10^{1}}$
	XGBoost	9.9×10^{-1}	7.0×10^{-6}	$\mathbf{9.4 \times 10^{-4}}$	6.2×10^{1}
	ElasticNet	4.4×10^{-1}	4.2×10^{-4}	1.0×10^{-2}	1.5×10^{2}
	TabNet	$\mathbf{9.9 \times 10^{-1}}$	$\mathbf{5.0 \times 10^{-6}}$	1.1×10^{-3}	7.2×10^{1}

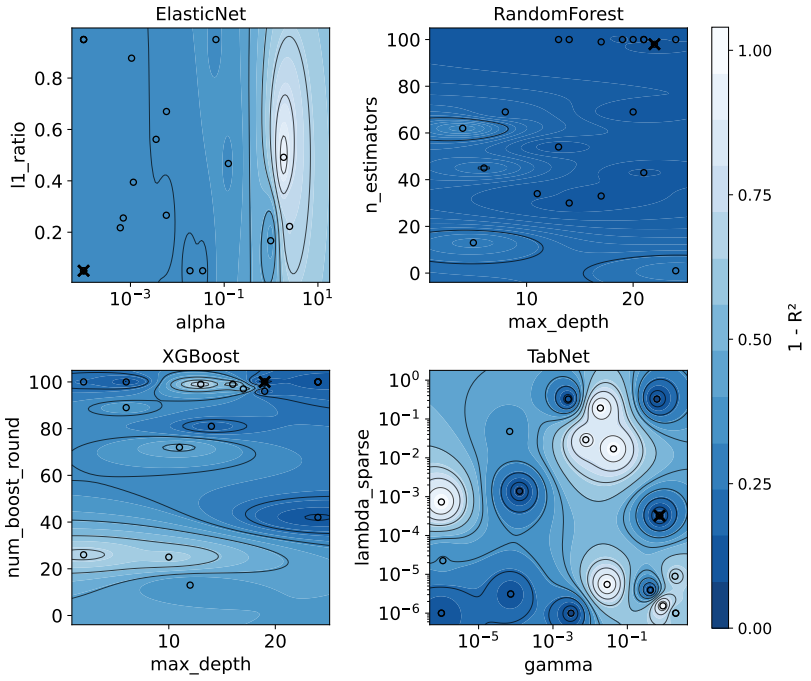

Figure 14. Hyperparameter optimization landscapes for each predictive model. The contours illustrate the optimization loss $(1 - R^2)$ in relation to two key hyperparameters. Circles correspond to the hyperparameter configurations evaluated during the Bayesian search, while the 'X' marks the best-performing selection.

5.2. Global and Instance-Level Feature Attribution Analysis

Beyond predictive accuracy, a central aim of CRITAIR is to elucidate the factors contributing to network interruptions. This inquiry is structured at two scales: global, to identify systemic trends, and local, to diagnose specific events.

The global feature-importance analysis (Figure 15) reveals a consensus among the evaluated models. They converge in identifying load density (`CNT_TRAFOS_AFEC`) and

various meteorological conditions (e.g., `h7-pres`, `h1-slp`) as determinant factors. Although informative, this high-level perspective inherently obscures the unique characteristics of individual interruption events.

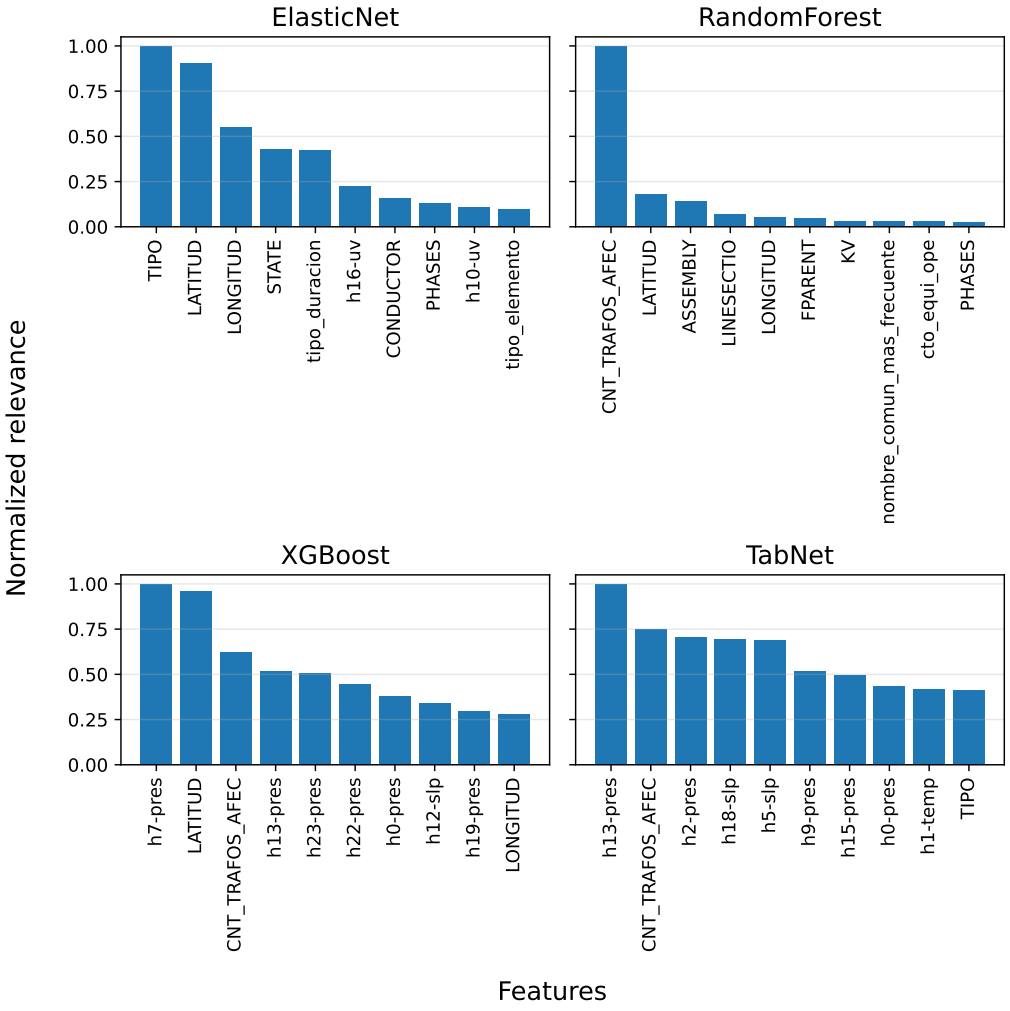

Figure 15. Global feature importance rankings derived from the training data for each model. The plots show the normalized relevance of the top 10 most influential features.

To transcend this limitation, CRITAIR leverages TabNet's architecture, whose sequential attention mechanism assigns distinct feature importance values for each prediction. Figure 16 depicts these instance-wise attributions across the test set, where each row corresponds to an event and color intensity denotes the contribution of each feature. This granular perspective facilitates a transition from aggregate analysis to specific diagnostics.

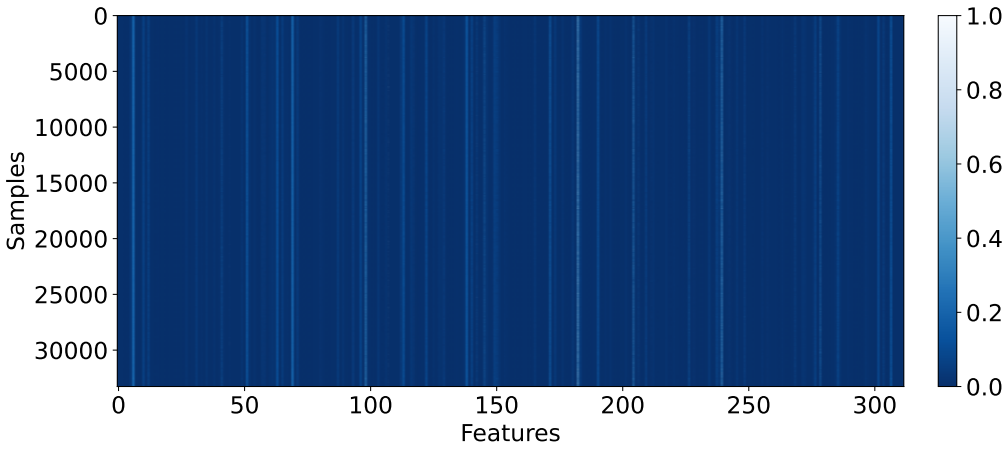

Figure 16. Visualization of TabNet's instance-wise feature importance (test set). Each row corresponds to a sample and each column to a feature. The color intensity represents the relevance assigned by the model's internal attention mechanism to a specific feature for a given sample.

The utility of this capability is exemplified in Figure 17, which contrasts the most influential variables in two high-impact scenarios. For the municipality with the highest SAIFI contribution (left panel), the prevailing contributors are meteorological, implying that interruptions are largely driven by environmental conditions affecting a high-density network. Conversely, for the highest-impact feeder (right panel), structural attributes such as circuit length (`LENGTH`) and conductor gauge (`CALIBRECONDUCTOR`) assume primary importance.

To explicitly validate the reliability of these attention-based attributions, we benchmarked TabNet's explanations against Shapley Additive exPlanations (SHAP). As depicted in the bottom panels of Figure 17 and the intersection diagrams in Figure 18, both methods consistently identify core structural drivers such as `CNT_TRAFOS_AFEC` and `TIPO`, confirming the model's grounding in physical network characteristics. However, a divergence in sensitivity is observed: while SHAP tends to distribute importance heavily across static infrastructure variables (e.g., `LINESECTIO`, `KV`), TabNet's sparse attention mechanism exhibits a sharper sensitivity to dynamic meteorological fluctuations (e.g., `h*-pres`, `h*-slp`). This differentiation suggests that while SHAP effectively highlights systemic vulnerabilities, TabNet captures the transient environmental context triggering specific failure events, a critical feature for real-time operational diagnostics.

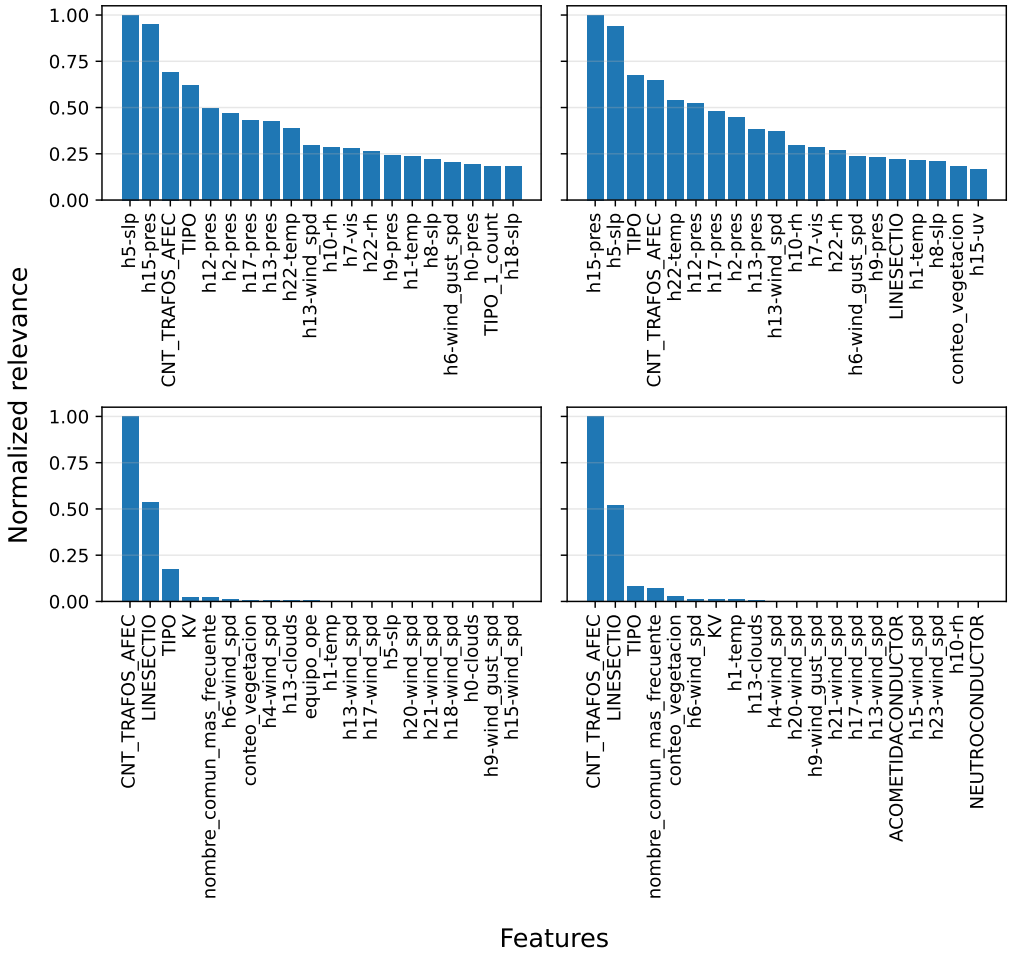

Figure 17. Comparative instance-level feature importance for two high-impact scenarios. **Left:** Top features for the municipality with the highest aggregate SAIFI. **Right:** Top features for the highest-impact distribution feeder. **Top:** Native attention-based attributions derived from TabNet's internal masks. **Bottom:** Corresponding SHAP values included to validate the reliability of the attention mechanisms against a model-agnostic benchmark.

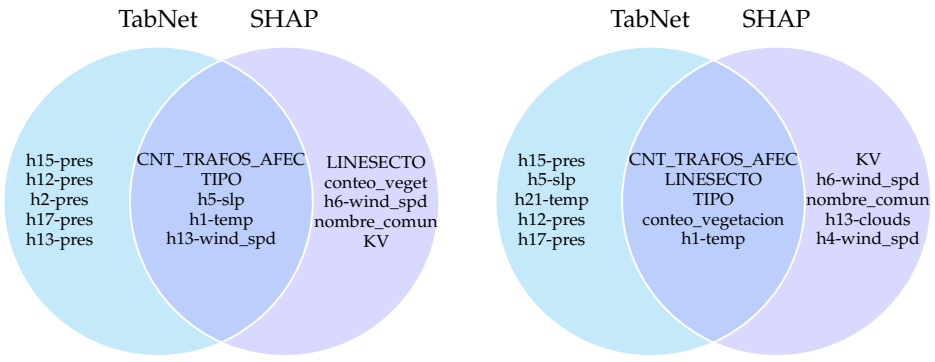

Figure 18. Venn diagrams analyzing feature convergence between attention-based masks (TabNet) and SHAP. The sets comprise the most influential variables for each method. **Left:** Municipality with the highest aggregate SAIFI. **Right:** Highest-impact distribution feeder.

5.3. Performance of the Regulation-Aware Agentic RAG System

The framework's final component is a reasoning engine that translates predictive outputs into actionable recommendations grounded in technical regulations. The selection

of a Large Language Model (LLM) for this engine must present an optimal balance between semantic quality and computational efficiency to be viable in an operational support environment. The comparative assessment (Figure 19) reveals that models like `Llama 3.2:1B` and `gpt-3.5-turbo` achieve this balance. Although larger API-based models attain slightly higher semantic quality, their increased latency renders them less practical for direct integration into real-time operational workflows, thereby validating the utility of lightweight models for local deployment.

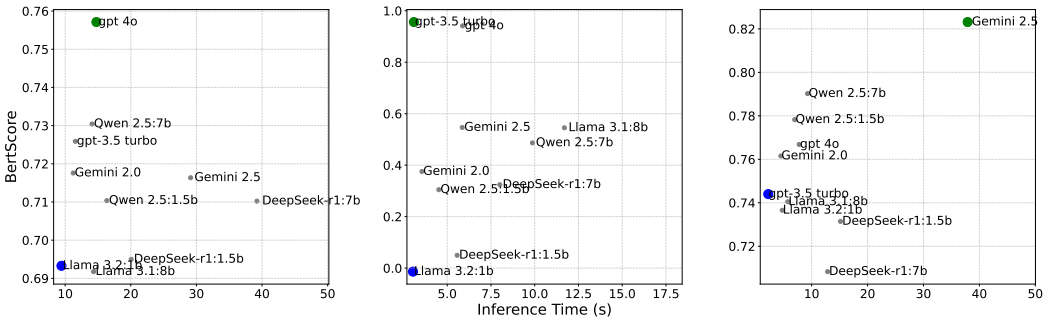

Figure 19. Performance trade-off analysis correlating semantic quality (F1 BERTScore) with inference time across Agentic RAG tasks. Gray markers represent model instances, while green and blue points indicate maximum accuracy and the best efficiency–performance balance, respectively. **Left:** Unstructured data processing. **Middle:** Structured data interpretation. **Right:** Recommendation synthesis.

To scrutinize regulatory adherence beyond simple similarity, we extended the evaluation within the unstructured data domain by computing Precision-BERTScore and Recall-BERTScore alongside the standard F1 metric. As illustrated in Figure 20, these complementary metrics enable a granular assessment of failure modes: specifically, whether a model tends to hallucinatory generation (low precision) or information omission (low recall). The violin plots reveal that while the evaluated models generally maintain a consistent semantic density (mostly within the 0.6–0.8 range), their reliability profiles vary significantly regarding normative grounding.

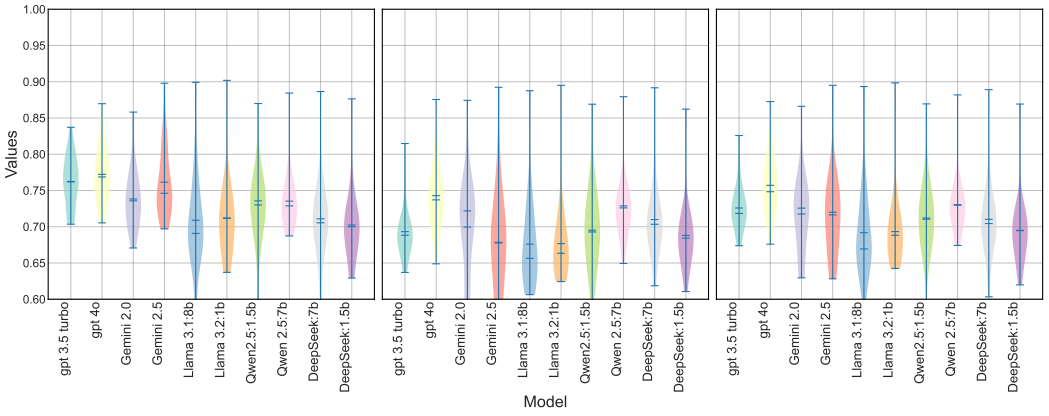

Figure 20. Distribution of BERTScore metrics across LLMs for unstructured data. **Left:** Precision-BERTScore distribution. **Middle:** Recall-BERTScore distribution. **Right:** F1-Score distribution.

Building on this distributional analysis, three distinct behavioral patterns emerge. First, GPT-4o stands out as the most reliable benchmark, achieving the highest median F1-Score (0.7571). This performance reflects a balanced capability to preserve normative content (recall of 0.7429) while effectively minimizing the fabrication of non-existent regulations (precision of 0.7723). In contrast, Llama-3.1:8B exhibits the lowest recall (0.6759),

indicating a systematic tendency to omit regulatory details present in the ground truth. Such omissions are particularly problematic in technical domains where the completeness of safety protocols is non-negotiable. Conversely, DeepSeek-r1:1.5b records the lowest precision (0.7024), suggesting a higher frequency of introducing content unsupported by the retrieved context. This behavior points to reduced controllability or higher generative drift, which can undermine trust in automated recommendations.

Overall, despite these localized differences, the performance band remains relatively narrow (F1-Scores between 0.69 and 0.76). This stability indicates that the retrieval pipeline feeding contextual information is robust; the observed discrepancies thus stem primarily from each model's intrinsic generative tendencies rather than RAG failures. This underscores the necessity of selecting models based not just on aggregate F1 scores, but on their specific precision-recall profile suited to the safety constraints of power systems.

The end-to-end workflow is demonstrated in a practical use case (Figure 21). Upon selecting an interruption event via the user interface, the TabNet model assesses the involved assets and visually highlights those with the highest estimated SAIFI contribution. This data-driven prioritization then informs the Agentic RAG system, which generates a specific diagnostic recommendation.

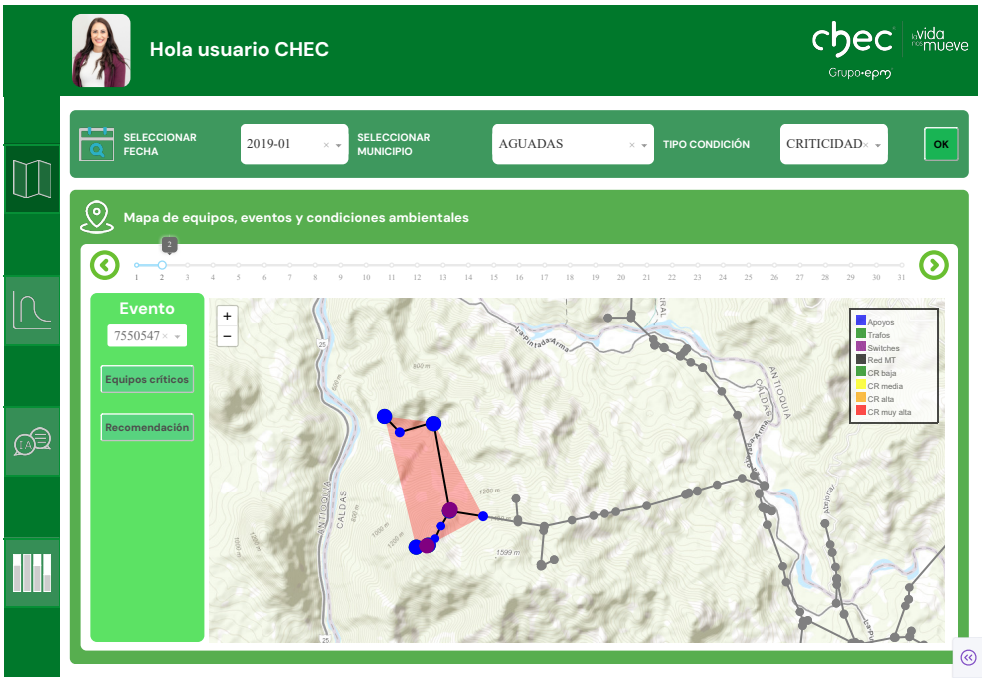

Figure 21. The CRITAIR system's user interface for a selected failure event. The map visualizes critical assets, with icon size proportional to their estimated SAIFI contribution as determined by the TabNet model.

A representative recommendation exemplifies how the system bridges predictive insights with regulatory standards. For instance, an ambient temperature of 14.1°C associated with sectionalizer `N43189` is contextualized against operational ranges defined in `RETIE` and `IEC 62271-1`. Analogously, recorded precipitation prompts a recommendation for a specific IP protection rating, citing `IEC 60529`. This process yields a concrete, verifiable directive that links a field condition to a technical requirement.

The system's integrity and auditability are anchored by an interpretable reasoning graph (Figure 22). This graph acts as a transparent, auditable trail documenting each diagnostic step: from the prioritized asset and its critical variables to the retrieved regulatory

clauses and the final recommendation. Each node and link are verifiable, facilitating subsequent regulatory audits or technical reviews. In this manner, CRITAIR completes the diagnostic cycle: commencing with quantitative risk estimation, advancing to the elucidation of probable causes, and culminating in an operational recommendation that is both auditable and compliant with governing regulations.

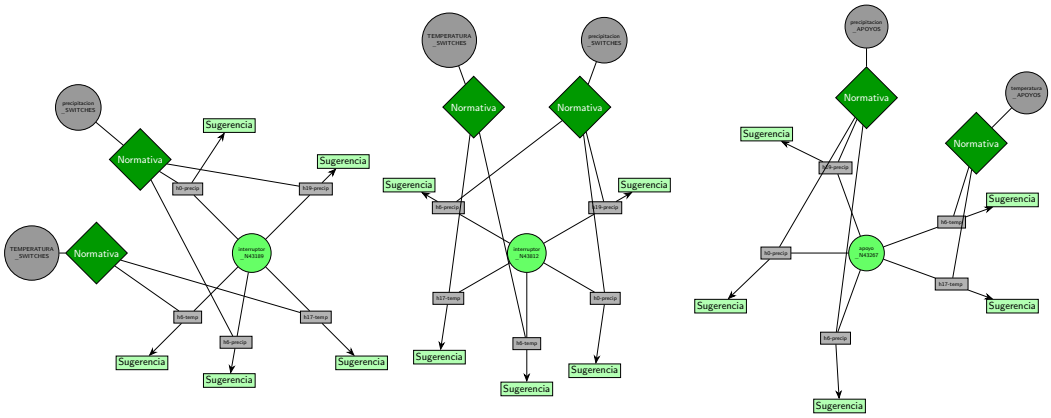

Figure 22. The interpretable reasoning graph providing an auditable trail for a specific asset diagnosis. The graph explicitly maps the predictive model's outputs (Critical Variables) to retrieved documentary evidence (Normativa).

5.4. Limitations

Although the CRITAIR framework represents a significant advancement in integrating predictive analytics and regulatory reasoning, several inherent limitations must be acknowledged, which in turn open future research avenues.

First, the performance of both the predictive model (TabNet) and the reasoning system (Agentic RAG) is fundamentally contingent upon the quality and completeness of the input data [93]. Despite comprehensive data enrichment, the absence or imprecision of asset records, unrecorded climatic events, or missing construction metadata can introduce biases, thereby affecting both the precision of SAIDI/SAIFI predictions and the relevance of the normative recommendations.

Second, the failure and severity prediction model — based on TabNet — was trained under a single, shared hyperparameter configuration across all evaluated settings [94]. This design choice promotes reproducibility and facilitates direct comparison against classical linear and nonlinear regressors (ElasticNet, Random Forest, XGBoost), but it restricts domain-specific optimization at the level of circuit topology, climatic region, or operational period. An adaptive hyperparameter search tailored to each zone or temporal window could in principle improve SAIDI/SAIFI estimation and increase the stability of the attention masks. However, such specialization would come at the cost of higher computational complexity and an increased risk of localized overfitting.

Finally, the system's evaluation was conducted on the operational environment and data from a single distribution network. While this ensures contextual relevance, the framework's generalizability to networks with different topologies, voltage levels, asset densities, and climatic profiles has not been tested [95]. Transferring the model to new operational contexts would likely necessitate significant hyperparameter retuning for the predictive model and adaptation of the agent's document corpus, posing a challenge for its immediate deployment in operational contexts beyond the one evaluated.

6. Conclusions

This paper has introduced CRITAIR, a hybrid and interpretable framework designed to support decision-making in the reliability management of medium-voltage (MV-L2)

distribution networks by aligning predictive analytics with regulatory governance requirements. CRITAIR integrates three key components: a TabNet-based predictive module for SAIDI/SAIFI estimation, an Agentic Retrieval-Augmented Generation (RAG) layer for normative grounding, and interpretable reasoning graphs to ensure end-to-end auditability.

The predictive module has demonstrated competitive performance against robust baselines such as Random Forest and XGBoost, achieving high accuracy in estimating reliability indicators. Crucially, through its sequential and sparse attention mechanism, it provides both global and local feature attributions, enabling the identification of the structural and meteorological factors that contribute most to interruptions without sacrificing transparency. Our Agentic RAG reasoning module has proven its capacity to effectively connect predictive insights with regulatory evidence extracted from technical documents like RETIE and NTC 2050. The generated recommendations are not only coherent and verifiable, as evidenced by high semantic alignment scores (BERTScore), but also interpretable by domain experts. The final transformation of the decision pathway into an explicit reasoning graph ensures complete traceability, an indispensable requirement in highly regulated environments. Collectively, CRITAIR bridges the existing gap between predictive analytics, which often operate as "black boxes," and the imperative for transparent and auditable governance in the power sector. By offering an integrated solution that is predictively accurate, explainable-by-design, and regulation-aware, this framework represents a valuable tool for the digital transformation of electric distribution utilities.

Future work will focus on expanding the framework to include resilience analysis by incorporating variables related to high-impact and low-probability events [96]. Because CRITAIR was trained and evaluated solely on data from CHEC, future research should examine its applicability across utilities with differing network topologies, climatic conditions, vegetation profiles, and regulatory frameworks. Evaluating the framework on multi-utility datasets will help assess model transferability and identify the domain-adaptation strategies needed for broader, regulation-aware deployment. Furthermore, we plan to enrich the analytical framework by integrating economic variables, such as operational (OPEX) and capital (CAPEX) expenditures [97]. This extension would enable CRITAIR not only to diagnose faults and recommend technical actions but also to assess their economic viability and prioritize interventions based on their impact on budgets and long-term asset management planning. Additionally, the integration of more advanced multi-agent architectures will be explored to collaboratively resolve more complex queries [98]. Finally, the implementation of continuous learning mechanisms will be investigated to allow the system to dynamically adapt to network changes and regulatory updates [99].

Author Contributions: Conceptualization, D.A.P.-R.,S.P.-Q., J.C.Á.-B.,A.M.Á.-M., and G.C.-D.; data curation, J.C.Á.-B., D.A.P.-R.; methodology, D.A.P.-R.,S.P.-Q., A.M.Á.-M.,J.C.Á.-B., and G.C.-D.; project administration, A.M.Á.-M.; supervision, A.M.Á.-M. and G.C.-D.; resources, D.A.P.-R.,S.P.-Q. and A.M.Á.-M. All authors have read and agreed to the published version of this manuscript.

Funding: This study was funded under grants provided for the project: Asesoría para implementar dashboard inteligente para el diagnóstico de redes eléctricas de nivel de tensión 2, a partir del análisis de criticidad dado por variables exógenas y endógenas, y generación de recomendaciones mediante técnicas de lenguaje natural -Contrato CRW254513 de 2024, funded by CHEC-Grupo EPM. Also, A.M. Alvarez-Meza and G. Castellanos-Dominguez thanks to the project: 'Sistema de visión artificial para el monitoreo y seguimiento de efectos analgésicos y anestésicos administrados vía neuroaxial epidural en población obstétrica durante labores de parto para el fortalecimiento de servicios de salud materna del Hospital Universitario de Caldas—SES HUC', Hermes 57661, funded by Universidad Nacional de Colombia.

Data Availability Statement: Data available upon reasonable request via email.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Krstivojević, J.; Stojković Terzić, J. Enhancing Reliability Performance in Distribution Networks Using Monte Carlo Simulation for Optimal Investment Option Selection. *Applied Sciences* **2025**, *15*. https://doi.org/10.3390/app15084209.
2. Seppälä, J.; Järventausta, P. Analyzing Supply Reliability Incentive in Pricing Regulation of Electricity Distribution Operators. *Energies* **2024**, *17*. https://doi.org/10.3390/en17061451.
3. Han, D.; Cho, I. Interactive visualization for smart power grid efficiency and outage exploration. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData). IEEE, 2023, pp. 656–661.
4. U.S. Energy Information Administration. U.S. electricity customers averaged five and one-half hours of power interruptions in 2022, 2024. Explains use of SAIDI/SAIFI and Major Event Days in U.S. reporting.
5. Weiss, M.; et al. Impact of Regulation on the Quality of Electric Power Distribution Services in Latin America and the Caribbean. Technical report, Inter-American Development Bank, 2021.
6. North American Electric Reliability Corporation. 2024 State of Reliability Overview. Technical report, 2024. Includes resilience framing for extreme events.
7. Comisión de Regulación de Energía y Gas, CREG. Circular CREG 053 de 2024: Metas de calidad media (SAIDI/SAIFI) para Operadores de Red, 2024.
8. XM Compañía de Expertos en Mercados. Publicación de indicadores de calidad (Resolución CREG 015 de 2018), 2025.
9. Ministerio de Minas y Energía de Colombia. Resolución 40117 de 2024: Modificación del Reglamento Técnico de Instalaciones Eléctricas (RETIE), 2024.
10. ICONTEC. Código Eléctrico Colombiano – NTC 2050 (versión vigente 2024), 2024. Referencia informativa; verificar edición oficial de ICONTEC.
11. Central Hidroeléctrica de Caldas S.A. E.S.P. (CHEC). Informe de ejecución 2024 — Plan de Inversión CHEC 2023–2027 (Actividad Distribución). Technical report, 2024.
12. Central Hidroeléctrica de Caldas S.A. E.S.P. (CHEC). Informe de ejecución del Plan de Inversiones 2023 — Distribución. Technical report, 2024.
13. Troncia, M.; Ruggeri, S.; Soma, G.G.; Pilo, F.; Ávila, J.P.C.; Muntoni, D.; Gianinoni, I.M. Strategic decision-making support for distribution system planning with flexibility alternatives. *Sustainable Energy, Grids and Networks* **2023**, *35*, 101138. https://doi.org/https://doi.org/10.1016/j.segan.2023.101138.
14. Ghasemkhani, B.; Kut, R.A.; Yilmaz, R.; Birant, D.; Arıkök, Y.A.; Güzelyol, T.E.; Kut, T. Machine Learning Model Development to Predict Power Outage Duration (POD): A Case Study for Electric Utilities. *Sensors* **2024**, *24*. https://doi.org/10.3390/s24134313.
15. Saeed, W.; Omlin, C. Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems* **2023**, *263*, 110273. https://doi.org/https://doi.org/10.1016/j.knosys.2023.110273.
16. Zhan, J.; Wu, C.; Yang, C.; Miao, Q.; Ma, X. HFN: Heterogeneous feature network for multivariate time series anomaly detection. *Information Sciences* **2024**, *670*, 120626.
17. Shadi, M.R.; Mirshekali, H.; Shaker, H.R. Explainable artificial intelligence for energy systems maintenance: A review on concepts, current techniques, challenges, and prospects. *Renewable and Sustainable Energy Reviews* **2025**, *216*, 115668.
18. Ghasemkhani, B.; Kut, R.A.; Yilmaz, R.; Birant, D.; Arıkök, Y.A.; Güzelyol, T.E.; Kut, T. Machine Learning Model Development to Predict Power Outage Duration (POD): A Case Study for Electric Utilities. *Sensors (Basel, Switzerland)* **2024**, *24*, 4313.
19. Willems, N.; Kar, B.; Levinson, S.; Turner, B.; Brewer, J.; Prica, M. Probabilistic Restoration Modeling of Wide-Area Power Outage. *IEEE Access* **2024**.
20. Alsaigh, R.; Mehmood, R.; Katib, I. AI explainability and governance in smart energy systems: a review. *Frontiers in Energy Research* **2023**, *11*, 1071291.
21. Wang, D.; Maharjan, S.; Zheng, J.; Liu, L.; Wang, Z. Data-driven quantification and visualization of resilience metrics of power distribution system. *arXiv preprint arXiv:2508.12408* **2025**.

22. Lin, J.; Xie, R.; Lin, H.; Guo, X.; Mao, Y.; Fang, Z. A Study on the Key Factors Influencing Power Grid Outage Restoration Times: A Case Study of the Jiexi Area. *Processes* **2025**, *13*. https://doi.org/10.3390/pr13092708.
23. Aldhubaib, H.A.; Hassan Ahmed, M.; Salama, M.M. A weather-based power distribution system reliability assessment. *Alexandria Engineering Journal* **2023**, *78*, 256–264. https://doi.org/https://doi.org/10.1016/j.aej.2023.07.033.
24. Zhou, Z.; Li, Y.; Guo, Z.; Yan, Z.; Chow, M.Y. A White-Box Deep-Learning Method for Electrical Energy System Modeling Based on Kolmogorov-Arnold Network. *arXiv preprint arXiv:2409.08044* **2024**.
25. Kostopoulos, G.; Davrazos, G.; Kotsiantis, S. Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review. *Electronics* **2024**, *13*. https://doi.org/10.3390/electronics13142842.
26. Shadi, M.R.; Mirshekali, H.; Shaker, H.R. Explainable artificial intelligence for energy systems maintenance: A review on concepts, current techniques, challenges, and prospects. *Renewable and Sustainable Energy Reviews* **2025**, *216*, 115668. https://doi.org/https://doi.org/10.1016/j.rser.2025.115668.
27. Chatterjee, J.; Dethlefs, N. XAI4Wind: A multimodal knowledge graph database for explainable decision support in operations & maintenance of wind turbines. *arXiv preprint arXiv:2012.10489* **2020**.
28. Shalev-Shwartz, S.; Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*; Cambridge University Press, 2014.
29. Arik, S.Ö.; Pfister, T. Tabnet: Attentive interpretable tabular learning. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2021, Vol. 35, pp. 6679–6687.
30. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **2019**, *1*, 206–215.
31. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.S.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.t. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the EMNLP (1), 2020, pp. 6769–6781.
32. Trangcasanchai, S. Improving Question Answering Systems with Retrieval Augmented Generation. PhD thesis, University of Helsinki, 2024.
33. Bockling, S.; et al. Walk the chain: Multi-agent reasoning for retrieval-augmented generation. In Proceedings of the ICLR, 2025.
34. Alotaibi, I.; Abido, M.A.; Khalid, M.; Savkin, A.V. A comprehensive review of recent advances in smart grids: A sustainable future with renewable energy resources. *Energies* **2020**, *13*, 6269.
35. Murphy, K.P. *Probabilistic machine learning: an introduction*; MIT press, 2022.
36. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* **2018**.
37. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692* **2019**.
38. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models are Few-Shot Learners. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020.
39. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* **2023**.
40. Bai, Y.; Zhao, Y.; Wu, H.; Lin, Z.; Sun, C.; Li, Y.; Xu, H.; Duan, N.; Zhou, M.; et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.05675* **2025**.
41. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* **2020**, *21*, 1–67.
42. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Genera-

tion, Translation, and Comprehension. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 7871–7880.
43. Qi, S.; Gui, L.; He, Y.; Yuan, Z. A Survey of Automatic Hallucination Evaluation on Natural Language Generation. *arXiv preprint arXiv:2404.12041* **2024**.
44. Xi, Z.; Chen, W.; Guo, X.; He, W.; Ding, Y.; Hong, B.; Zhang, M.; Wang, J.; Jin, S.; Zhou, E.; et al. The rise and potential of large language model based agents: A survey. arXiv 2023. *arXiv preprint arXiv:2309.07864* **2025**, *10*.
45. Chen, X.; Wang, Y.; Liu, H. Application of Knowledge Graph Technology in Fault Diagnosis of Power Systems. *Frontiers in Energy Research* **2022**, *10*, 988280. https://doi.org/10.3389/fenrg.2022.988280.
46. Li, J.; Zhang, L.; Zhou, P. Knowledge Graph Construction for Fault Diagnosis in Power Systems. *Electronics* **2023**, *12*, 4808. https://doi.org/10.3390/electronics12234808.
47. Chen, Q.; Li, Q.; Wu, J.; Mao, C.; Peng, G.; Wang, D. Application of knowledge graph in power system fault diagnosis and disposal: A critical review and perspectives. *Frontiers in Energy Research* **2022**, *10*, 988280.
48. Liu, R.; Fu, R.; Xu, K.; Shi, X.; Ren, X. A review of knowledge graph-based reasoning technology in the operation of power systems. *Applied Sciences* **2023**, *13*, 4357.
49. Team, N.R. GraphRAG: Enhancing Retrieval-Augmented Generation with Knowledge Graphs. https://neo4j.com/blog/developer/graphrag-and-agentic-architecture-with-neoconverse/, 2024.
50. Liu, C.; et al. KG-SMILE: Knowledge graph-supported interpretable recommendations. *Expert Systems with Applications* **2025**, *240*, 122556.
51. Ranstam, J.; Cook, J.A. LASSO regression. *Journal of British Surgery* **2018**, *105*, 1348–1348.
52. Nachouki, M.; Mohamed, E.A.; Mehdi, R.; Abou Naaj, M. Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. *Trends in Neuroscience and Education* **2023**, *33*, 100214.
53. Du, K.L.; Zhang, R.; Jiang, B.; Zeng, J.; Lu, J. Foundations and innovations in data fusion and ensemble learning for effective consensus. *Mathematics* **2025**, *13*, 587.
54. Kumar, A.; Sinha, S.; Saurav, S. Random forest, CART, and MLR-based predictive model for unconfined compressive strength of cement reinforced clayey soil: A comparative analysis. *Asian Journal of Civil Engineering* **2024**, *25*, 2307–2323.
55. Uyar, S.G.K.; Ozbay, B.K.; Dal, B. Interpretable building energy performance prediction using XGBoost Quantile Regression. *Energy and Buildings* **2025**, *340*, 115815.
56. Wiens, M.; Verone-Boyle, A.; Henscheid, N.; Podichetty, J.T.; Burton, J. A tutorial and use case example of the eXtreme gradient boosting (XGBoost) artificial intelligence algorithm for drug development applications. *Clinical and Translational Science* **2025**, *18*, e70172.
57. Martins, A.; Astudillo, R. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Proceedings of the International conference on machine learning. PMLR, 2016, pp. 1614–1623.
58. Dauphin, Y.N.; Fan, A.; Auli, M.; Grangier, D. Language modeling with gated convolutional networks. In Proceedings of the International conference on machine learning. PMLR, 2017, pp. 933–941.
59. Dimitriou, N.; Arandjelovic, O. A new look at ghost normalization. *arXiv preprint arXiv:2007.08554* **2020**.
60. Xuan, H.; Yang, B.; Li, X. Exploring the impact of temperature scaling in softmax for classification and adversarial robustness. *arXiv preprint arXiv:2502.20604* **2025**.
61. Khanda, R. Agentic ai-driven technical troubleshooting for enterprise systems: A novel weighted retrieval-augmented generation paradigm. *arXiv preprint arXiv:2412.12006* **2024**.
62. Low, Y.S.; Jackson, M.L.; Hyde, R.J.; Brown, R.E.; Sanghavi, N.M.; Baldwin, J.D.; Pike, C.W.; Muralidharan, J.; Hui, G.; Alexander, N.; et al. Answering real-world clinical questions using large language model, retrieval-augmented generation, and agentic systems. *Digital Health* **2025**, *11*, 20552076251348850.
63. Zhao, P.; Zhang, H.; Yu, Q.; Wang, Z.; Geng, Y.; Fu, F.; Yang, L.; Zhang, W.; Jiang, J.; Cui, B. Retrieval-augmented generation for ai-generated content: A survey. *arXiv preprint arXiv:2402.19473* **2024**.

64. Singh, A.; Ehtesham, A.; Kumar, S.; Khoei, T.T. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136* **2025**.
65. Pandey, V. Agentic AI with retrieval-augmented generation for automated compliance assistance in finance. *International Journal of Science and Research Archive* **2025**, *15*, 1620–1631. https://doi.org/10.30574/ijsra.2025.15.2.1522.
66. Liang, J.; Su, G.; Lin, H.; Wu, Y.; Zhao, R.; Li, Z. Reasoning RAG via System 1 or System 2: A Survey on Reasoning Agentic Retrieval-Augmented Generation for Industry Challenges. *arXiv preprint arXiv:2506.10408* **2025**.
67. Kukreja, S.; Kumar, T.; Bharate, V.; Gadwe, S.; Dasgupta, A.; Guha, D. Performance Enhancement of Agentic Retrieval Augmented Generation Using Relevance Generative Answering. In Proceedings of the 2025 5th International Conference on Artificial Intelligence and Education (ICAIE). IEEE, 2025, pp. 465–469.
68. Maragheh, R.Y.; Vadla, P.; Gupta, P.; Zhao, K.; Inan, A.; Yao, K.; Xu, J.; Kanumala, P.; Cho, J.; Kumar, S. ARAG: Agentic Retrieval Augmented Generation for Personalized Recommendation. *arXiv preprint arXiv:2506.21931* **2025**.
69. Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q.V.; Zhou, D.; et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **2022**, *35*, 24824–24837.
70. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. React: Synergizing reasoning and acting in language models. In Proceedings of the International Conference on Learning Representations (ICLR), 2023.
71. Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N.V.; Wiest, O.; Zhang, X. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* **2024**.
72. Lee, M.C.; Zhu, Q.; Mavromatis, C.; Han, Z.; Adeshina, S.; Ioannidis, V.N.; Rangwala, H.; Faloutsos, C. HybGrag: Hybrid retrieval-augmented generation on textual and relational knowledge bases. *arXiv preprint arXiv:2412.16311* **2024**.
73. Necula, S. Exploring the model-view-controller (mvc) architecture: A broad analysis of market and technological applications **2024**.
74. Elkhidir, E.; Patel, T.; Rotimi, J.O.B. Predictive modelling for residential construction demands using ElasticNet Regression. *Buildings* **2025**, *15*, 1649.
75. Wekalao, J.; Njoroge, S.M.; Elamri, O. Enhanced malaria detection using a hybrid borophene-based terahertz biosensor with random forest regression analysis. *Brazilian Journal of Physics* **2025**, *55*, 1–18.
76. Qi, Z.; Feng, Y.; Wang, S.; Li, C. Enhancing hydropower generation Predictions: A comprehensive study of XGBoost and Support Vector Regression models with advanced optimization techniques. *Ain Shams Engineering Journal* **2025**, *16*, 103206.
77. Roberts, J. How powerful are decoder-only transformer neural models? In Proceedings of the 2024 International Joint Conference on Neural Networks (IJCNN). IEEE, 2024, pp. 1–8.
78. Machado, J. Toward a Public and Secure Generative AI: A Comparative Analysis of Open and Closed LLMs. *arXiv preprint arXiv:2505.10603* **2025**.
79. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* **2020**.
80. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems* **2023**, *36*, 10088–10115.
81. Zhao, Z.R.; Chou, P.C.; Mir, T.H. A Comparative Study of GPT3. 5 Fine Tuning and Rule-Based Approaches. In Proceedings of the Large Language Models for Automatic Deidentification of Electronic Health Record Notes: International Workshop, IW-DMRN 2024, Kaohsiung, Taiwan, January 15, 2024, Revised Selected Papers. Springer Nature, 2025, Vol. 2148, p. 30.
82. Aryal, S.; Agyemang-Prempeh, J. Howard university-ai4pc at semeval-2025 task 2: Improving machine translation with context-aware entity-only pre-translations with gpt4o. In Proceedings of the Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025, pp. 1885–1889.

83. Balestri, R. Gender and content bias in Large Language Models: a case study on Google Gemini 2.0 Flash Experimental. *Frontiers in Artificial Intelligence* **2025**, *8*, 1558696.
84. Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* **2025**.
85. Kassianik, P.; Saglam, B.; Chen, A.; Nelson, B.; Vellore, A.; Aufiero, M.; Burch, F.; Kedia, D.; Zohary, A.; Weerawardhena, S.; et al. Llama-3.1-foundationai-securityllm-base-8b technical report. *arXiv preprint arXiv:2504.21039* **2025**.
86. Azaiz, I.; Kiesler, N.; Strickroth, S.; Zhang, A. Open, Small, Rigmarole–Evaluating Llama 3.2 3B's Feedback for Programming Exercises. *arXiv preprint arXiv:2504.01054* **2025**.
87. Yang, W.; Yue, X.; Chaudhary, V.; Han, X. Speculative thinking: Enhancing small-model reasoning with large model guidance at inference time. *arXiv preprint arXiv:2504.12329* **2025**.
88. Wu, Y.; Mei, J.; Yan, M.; Li, C.; Lai, S.; Ren, Y.; Wang, Z.; Zhang, J.; Wu, M.; Jin, Q.; et al. Writingbench: A comprehensive benchmark for generative writing. *arXiv preprint arXiv:2503.05244* **2025**.
89. Aksyonov, K.A.; Sun, L.; Kalinin, I.A.; Aksyonova, O.P.; Aksyonova, E.K. Deploying a Local Language Learning Assistant Using a Small Large Language Model. In Proceedings of the 2025 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT). IEEE, 2025, pp. 372–375.
90. Sonawane, V.; Sambare, G.B.; Ambala, S.; Kadam, G. Implementation of an Interactive Query System Using Nomic Text Embed, DeepSeek R1 1.5 B, and Cosine Similarity rankers. In Proceedings of the 2025 International Conference on Computing Technologies (ICOCT). IEEE, 2025, pp. 1–6.
91. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the International Conference on Learning Representations (ICLR), 2020. Available at: https://arxiv.org/abs/1904.09675.
92. Goel, R. Using text embedding models as text classifiers with medical data. *arXiv preprint arXiv:2402.16886* **2024**.
93. Hector, I.; Panjanathan, R. Predictive maintenance in Industry 4.0: a survey of planning models and machine learning techniques. *PeerJ Computer Science* **2024**, *10*, e2016.
94. Baratchi, M.; Wang, C.; Limmer, S.; Van Rijn, J.N.; Hoos, H.; Bäck, T.; Olhofer, M. Automated machine learning: past, present and future. *Artificial intelligence review* **2024**, *57*, 122.
95. Wang, Y.; Zhao, H.; Lin, H.; Xu, E.; He, L.; Shao, H. A Generalizable Physics-Enhanced State Space Model for Long-Term Dynamics Forecasting in Complex Environments. *arXiv preprint arXiv:2507.10792* **2025**.
96. Shen, J.; Bao, X.; Chen, X.; Wu, X.; Qiu, T.; Cui, H. Seismic resilience assessment method for tunnels based on cloud model considering multiple damage evaluation indices. *Tunnelling and Underground Space Technology* **2025**, *157*, 106360.
97. Bovera, F.; Schiavo, L.L.; Vailati, R. Combining Forward-Looking Expenditure Targets and Fixed OPEX-CAPEX Shares for a Future-Proof Infrastructure Regulation: the ROSS Approach in Italy. *Current Sustainable/Renewable Energy Reports* **2024**, *11*, 105–115.
98. Icarte-Ahumada, G.; He, Z.; Godoy, V.; García, F.; Oyarzún, M. A Multi-Agent System for Parking Allocation: An Approach to Allocate Parking Spaces. *Electronics* **2025**, *14*, 840.
99. Findik, Y.; Hasenfus, H.; Azadeh, R. Collaborative Adaptation for Recovery from Unforeseen Malfunctions in Discrete and Continuous MARL Domains. In Proceedings of the 2024 IEEE 63rd Conference on Decision and Control (CDC). IEEE, 2024, pp. 394–400.

Appendix H Prompt Templates

This appendix presents the verbatim prompt templates injected into the LLM during the evaluation to ensure full reproducibility of the generative components.

Appendix H.1 Structured Query Prompt

Task: Operational questions based on tabular data (DataFrames).

```
Este DataFrame contiene informacion acerca de interrupciones o eventos presentadas en redes
    electricas de media tension, mas especificamente en tres tipos de equipos:
    Transformadores, interruptores y tramos de linea.

Las columnas incluyen:
- Evento: Id de la interrupcion o el evento.
- equipo_ope: Codigo del equipo en el que ocurrio la interrupcion.
- tipo_equi_ope: Indica si la interrupcion ocurrio sobre un Transformador, interruptor o
    tramo de linea.
- cto_equi_ope: Codigo del circuito.
- tipo_elemento: Capacidad en kV (33, 13.2, TFD, TFP).
- inicio: Fecha y hora del inicio.
- fin: Fecha y hora de la finalizacion.
- duracion_h: Duracion en horas.
- tipo_duracion: Categoria (> 3 min y <= 3 min).
- causa: Causa del evento.
- CNT_TRAFOS_AFEC: Cantidad de transformadores afectados.
- cnt_usus: Cantidad de usuarios afectados.
- SAIDI: Promedio de duracion por usuario.
- SAIFI: Promedio de interrupciones por usuario.
- PHASES: Numero de fases (3., 1., 2.).
- FPARENT: Codigo del circuito padre.
- FECHA, LONGITUD, LATITUD, DEP, MUN: Datos espacio-temporales.

A continuacion, se muestran las primeras 5 filas del DataFrame:

{head_df}

De acuerdo a esto responde a las preguntas formuladas por el usuario:
Human: {human_input}
```

Appendix H.2 Unstructured Normative Query Prompt

Task: Regulatory compliance questions based on RAG context.

```
Se te proporcionara una serie de textos que contienen instrucciones sobre como resolver
     preguntas acerca de normativas en redes electricas de nivel de tension 2. Segun estos
     textos, responde a la pregunta de la manera mas completa posible.

Dado el siguiente contexto, responde a las preguntas hechas por el usuario.

IMPORTANTE: Estructura tu respuesta de la siguiente manera:
1. Primero proporciona la recomendacion tecnica o respuesta directa a la pregunta
2. Luego indica claramente de acuerdo a que normativa(s) se basa esta recomendacion

CRITICO - Referencias normativas:
- Especifica SIEMPRE el nombre COMPLETO de la normativa.
- Cita el ARTICULO o SECCION especifica.
- Si aplica, menciona el APARTADO o LITERAL concreto.
- Incluye el NUMERO de pagina o tabla si esta disponible.
- Si hay multiples normativas aplicables, citalas TODAS.

Contexto:
{context}

Human: {human_input}
Chatbot (RESPUESTA FORMAL):
```

Appendix H.3 Recommendation Task Prompt

Task: Expert technical recommendations based on specific variable validation.

```
Eres un experto tecnico en infraestructura electrica. Tu funcion es dar recomendaciones y
     pautas normativas basadas en el contexto que se te proporciona.

De acuerdo al valor de la variable que menciona el usuario en su pregunta, sigue estos
     pasos:
1. Identifica la variable y el valor que proporciona el usuario.
2. Consulta el contexto normativo proporcionado (normas minimas o rangos).
3. Compara el valor dado con las normas del contexto.
   - Si el valor NO cumple con la norma, debes decirlo claramente, explicar por que no
     cumple y recomendar la accion necesaria.
   - Si el valor SI cumple con la norma, debes confirmarlo y brindar informacion adicional.
4. Presenta la respuesta de forma clara y directa.

Usa el contexto y el historial de la conversacion para responder a las preguntas del
     usuario:

{context}

{chat_history}
Human: {human_input}
Chatbot (RESPUESTA RECOMENDACION):
```

Appendix I Model Hyperparameter Configuration

To ensure the reproducibility of the predictive stability analysis (Section 5), Table A7 details the final hyperparameter sets for each model. These values were obtained through an automated tuning process maximizing the validation metric on the time-aware split.

Table A7. Optimized hyperparameters for the predictive models (ElasticNet, Random Forest, XGBoost, and TabNet).

Model	Parameter	Value
ElasticNet	Alpha	1.0×10^{-4}
	L1 Ratio	0.05
	Max Iterations	3000
	Tolerance	1.46×10^{-4}
Random Forest	N Estimators	98
	Max Depth	22
	Max Features	0.43 (Fraction)
	Min Samples Leaf	4
	Min Samples Split	9
XGBoost	Max Depth	19
	Learning Rate (eta)	0.3
	Subsample	0.6
	Colsample By Tree	1.0
	Reg Lambda (L2)	10.0
	Reg Alpha (L1)	1.0×10^{-6}
	Num Boost Round	100
TabNet	N_d (Prediction Layer)	75
	N_a (Attention Layer)	27
	Steps	9
	Gamma	0.734
	Lambda Sparse	3.2×10^{-4}
	Mask Type	Sparsemax
	Learning Rate	0.071
	Batch Size	2048
	Virtual Batch Size	2048
	Optimizer	Adam
	Momentum	0.381
	Weight Decay	1.2×10^{-3}