



# CRITAIR: A Hybrid Methodology for Criticality Analysis and Intelligent Recommendations in Electric Distribution Networks

# CRITAIR: Una Metodología Híbrida para el Análisis de Criticidad y Recomendaciones Inteligentes en Redes de Distribución Eléctrica

Santiago Pineda Quintero

Universidad Nacional de Colombia  
Facultad de ingeniería eléctrica, electrónica y computación  
Departamento de ingeniería y arquitectura  
Manizales, Colombia  
2025

# CRITAIR: A Hybrid Methodology for Criticality Analysis and Intelligent Recommendations in Electric Distribution Networks

## CRITAIR: Una Metodología Híbrida para el Análisis de Criticidad y Recomendaciones Inteligentes en Redes de Distribución Eléctrica

Santiago Pineda Quintero

Tesis presentada como requisito parcial para optar por el título de:  
**Magister en Ingeniería - Automatización Industrial**

Director(a):

Prof. Dr. Andrés Marino Álvarez Meza

Codirector(a):

Prof. Dr. César Germán Castellanos Domínguez

Línea de investigación:

Inteligencia Artificial

Grupo de investigación:

Grupo de control y procesamiento digital de señales (GCPDS)

Universidad Nacional de Colombia

Facultad de ingeniería eléctrica, electrónica y computación

Departamento de ingeniería y arquitectura

2025

"Logic will get you from A to B. Imagination will take you everywhere."

*Albert Einstein*

"Believe you can, and you are halfway there."

*Theodore Roosevelt*

# Declaración

Me permito afirmar que he realizado ésta tesis de manera autónoma y con la única ayuda de los medios permitidos y no diferentes a los mencionados el presente texto. Todos los pasajes que se han tomado de manera textual o figurativa de textos publicados y no publicados, los he reconocido en el presente trabajo. Ninguna parte del presente trabajo se ha empleado en ningún otro tipo de tesis.

Manizales., 09-15-2025

---

Santiago Pineda Quintero

# Acknowledgements

Firstly, I extend my gratitude to my supervisor, Andrés Marino Álvarez, and my co-supervisor, Germán Castellanos Domínguez. Their invaluable guidance and support have been indispensable in successfully completing this research. Their contributions have influenced this study and left an indelible impact on my academic journey.

I wish to express my deepest gratitude to my dear parents, whose unwavering support, wise counsel during challenging times, and unconditional love have been the foundation of my strength and perseverance.

I am particularly grateful to all members of the Signal Processing and Recognition Group (SPRG) at the National University of Colombia in Manizales. A special note of thanks goes to Juan Sebastian Méndez, Juan Carlos Aguirre, and Mateo Tobón, whose camaraderie and academic discussions have significantly enriched my experience over these years.

Finally, I sincerely thank the company CHEC (Central Hidroeléctrica de Caldas, Grupo EPM) for their valuable collaboration throughout this research. In particular, I acknowledge their willingness to provide access to internal datasets on medium-voltage distribution networks, which were fundamental for the validation of the proposed methodology. Their openness to university–industry cooperation not only enabled the successful completion of this thesis but also highlighted the importance of bridging academic research with real-world operational challenges.

# List of Symbols and Abbreviations

## Number Sets

$ \cdot $	Set cardinality
$diag(\cdot)$	Diagonal operator
$\mathcal{X}$	Functional
$\Gamma_{x,x'}$	Cross-spectral density
$\Gamma_{x,x}$	Power spectral density
$\delta_K(\cdot, \cdot)$	Kronecker delta function
$\mathbf{X}$	Matrix
$\mathbb{N}$	The set of natural numbers
$\mathbb{E}\{\cdot\}$	Expectation operator
$\mathbb{R}$	The set of real numbers
$Tr(\cdot)$	Trace operator
$\mathbf{x}$	Vector
$x$	Scalar

## Symbols

$\mu V$	Microvolts
$Acc$	Accuracy performance
$Hz$	Hertz
$mm$	Millimeters
$mV$	Milivolts
$N_C$	Number of channels/electrodes
$N_f$	Number of filters
$N_p$	Number of features
$N_t$	Number of time samples
$s$	seconds

## Abbreviations

AI	Artificial Intelligence
CHEC	Central Hidroeléctrica de Caldas
CREG	Comisión de Regulación de Energía y Gas
CRITAIR	Criticality Analysis and Intelligent Recommendations
DL	Deep learning
DSP&CG	Digital Signal Processing and Control Group
EPC	Electric Power Companies
LLM	Large Language Model
ML	Machine Learning
MV-L2	Medium-Voltage Level 2
NLP	Natural Language Processing
NTC	Normas Técnicas Colombianas
RAG	Retriever Augmented Generation
RETIE	Reglamento Técnico de Instalaciones Eléctricas
SAIDI	System Average Interruption Duration Index
SAIFI	System Average Interruption Frequency Index

# Resumen

Los sistemas eléctricos modernos enfrentan niveles crecientes de complejidad y demanda, lo que hace prioritario comprender las causas raíz de fallas e interrupciones. Esta comprensión es fundamental para optimizar indicadores de confiabilidad como el System Average Interruption Duration Index (SAIDI) y el System Average Interruption Frequency Index (SAIFI), mejorando así la calidad del servicio y la experiencia del usuario final.

Sin embargo, en las empresas de energía eléctrica (EPCs), las interrupciones inesperadas en redes de media tensión nivel 2 (MV-L2) continúan deteriorando estos indicadores, afectando directamente la percepción del servicio. Esta problemática se debe en gran medida a la ausencia de metodologías sistemáticas que permitan identificar con precisión las variables internas y externas que influyen en dichos indicadores, limitando la gestión proactiva de activos y la prevención de fallas.

Este trabajo identifica dos desafíos principales. Primero, la falta de modelos analíticos capaces no solo de predecir métricas de confiabilidad, sino también de explicar las causas subyacentes de las interrupciones. Segundo, la dificultad de transformar grandes volúmenes de datos históricos y normativos en recomendaciones claras y accionables, debido a la ausencia de sistemas que integren conocimiento experto con analítica de datos de forma interpretable.

Para abordar estos desafíos, se propone CRITAIR (Criticality Analysis and Intelligent Recommendations), una metodología híbrida e interpretable de dos etapas. En la primera etapa, se entrena un modelo TabNet utilizando datos históricos de interrupciones enriquecidos con variables meteorológicas y metadatos constructivos, permitiendo estimar el indicador SAIDI y detectar variables influyentes a nivel global y local. En la segunda etapa, estas variables se integran en un sistema Agentic RAG (Retrieval-Augmented Generation), el cual combina recuperación semántica y generación de texto mediante modelos de lenguaje, permitiendo generar recomendaciones contextualizadas basadas en datos estructurados y documentos normativos. Adicionalmente, el sistema produce grafos de razonamiento interpretables que explican el proceso de toma de decisiones.

Los resultados muestran que el modelo TabNet alcanzó un coeficiente de determinación  $R^2 = 0.88$  para SAIDI, identificando como variables más relevantes la precipitación, ráfagas de viento, nubosidad, corriente mínima y calibre del conductor, explicando el 67.3% de la variabilidad. El sistema Agentic RAG alcanzó un BERTScore de 0.956 en consultas tabulares, 0.984 en interpretación normativa y 0.743 en generación de recomendaciones. Además, el sistema genera grafos interpretables que permiten validar las decisiones del modelo. Los resultados fueron validados con datos reales de CHEC, demostrando su aplicabilidad en contextos operativos reales.

**Palabras clave:** TabNet, Generación Aumentada por Recuperación, Modelos de Lenguaje de Gran Escala, BertScore, SAIDI, SAIFI, RETIE, NTC, Media Tensión Nivel 2.

# Abstract

Modern power systems face increasing levels of complexity and demand, making it a priority to understand the root causes of faults and outages. Such understanding is essential for optimizing reliability indicators such as System Average Interruption Duration Index (SAIDI) and System Average Interruption Frequency Index (SAIFI), thereby enhancing service quality and the overall user experience.

However, in Electric Power Companies (EPCs), unexpected outages in Medium-Voltage Level 2 (MV-L2) networks continue to degrade key reliability indicators such as SAIDI and SAIFI, ultimately impacting the end user’s perception of service quality. This ongoing issue is largely due to the lack of a systematic methodology to accurately identify the internal and external variables that most influence these indicators, limiting proactive asset management and hindering the prevention of future failures.

This work identifies two key problems that hinder data-driven decision-making aimed at improving reliability in MV-L2 networks. First, there is a lack of analytical models capable of both predicting reliability metrics and explaining the underlying causes of service interruptions. Second, organizations struggle to translate large volumes of historical and regulatory data into clear, actionable recommendations due to the absence of systems that integrate domain knowledge with data-driven insights in an interpretable manner.

To address these challenges, we propose CRITAIR (Criticality Analysis and Intelligent Recommendations), a two-stage hybrid and interpretable methodology designed to identify, explain, and recommend actions aimed at improving reliability in medium-voltage networks. In the first stage, a TabNet model is trained using historical outage records enriched with meteorological variables and construction-related metadata, enabling accurate estimation of SAIDI while identifying influential variables at both global and local levels. In the second stage, the extracted feature importance is integrated into an Agentic RAG (Retrieval-Augmented Generation) system, which combines semantic retrieval with text generation using large language models to generate contextualized recommendations based on both structured and unstructured data. Additionally, the system produces interpretable reasoning graphs that explain the decision-making process of the intelligent agent.

The results show that the TabNet model achieved a coefficient of determination of  $R^2 = 0.88$  for the SAIDI indicator, identifying precipitation, wind gusts, cloud cover, minimum current, and conductor gauge as the most relevant variables, explaining 67.3% of the observed variability. The Agentic RAG system achieved a BERTScore of 0.956 for tabular queries, 0.984 for regulatory interpretation, and 0.743 for recommendation generation. Furthermore, the system generates interpretable reasoning graphs that enhance transparency and trust. These results were validated using real-world data from CHEC, demonstrating the applicability of the proposed methodology in operational environments.

**Keywords:** TabNet, Retrieval-Augmented Generation, Large Language Models, BertScore, SAIDI, SAIFI, RETIE, NTC, Medium-Voltage Level 2.

# List of Figures

<b>1-1</b>	Schematic representation of criticality in MV-L2 networks. . . . .	2
<b>1-2</b>	Illustration of the two key challenges identified: Lack of analytical models with predictive and explanatory capabilities and Absence of integrated, interpretable decision support systems . .	4
<b>1-3</b>	Lack of analytical models with predictive and explanatory capabilities graphical scheme . . .	5
<b>1-4</b>	Absence of integrated, interpretable decision support systems graphical scheme . . . . .	6
<b>1-5</b>	Comparative spider chart of major families of regression models for MV-L2 reliability prediction: Linear classical models, non-linear classical ML models , deep neural networks, and attention-based tabular models such as TabNet . . . . .	8
<b>1-6</b>	Comparative spider chart of NLP-based decision-support system families: NLP Classic Tasks, Standard RAG Architectures and Agentic RAG Architectures . . . . .	10
<b>3-1</b>	Schematic display of the main thesis contributions, including the predictive and interpretable TabNet architecture for SAIDI estimation and feature attribution, the integration of an Agentic RAG system that combines structured outage data with unstructured regulatory knowledge to generate recommendations, and the development of interpretable reasoning graphs that provide both quantitative and qualitative transparency of the intelligent agent’s decision-making process. . . . .	13
<b>4-1</b>	Schematic representation of a linear modeling workflow: data input, coefficient estimation, and feature relevance extraction. . . . .	27
<b>4-2</b>	Conceptual pipeline for Random Forest regression: bagging, tree training, ensemble averaging, and importance derivation. . . . .	28
<b>4-3</b>	TabNet step-wise architecture: each decision step applies attentive feature selection, transformation, and residual aggregation before producing the final prediction. . . . .	33
<b>5-1</b>	Pipeline for TabNet-based criticality analysis. Preprocessing integrates data imputation, encoding, and normalization; training optimizes TabNet hyperparameters with Bayesian search; interpretability is provided through attention masks and sparsity regularization. . . .	49
<b>5-2</b>	Convergence of TabNet regression loss during training. . . . .	50
<b>5-3</b>	Feature importance distribution across risk levels derived from TabNet attention masks. . . .	51
<b>5-4</b>	Correlation heatmap across environmental and structural variables, showing clusters of interdependent drivers influencing SAIDI. . . . .	52
<b>5-5</b>	Focused correlation analysis between SAIDI and top explanatory variables across different risk categories. . . . .	52
<b>5-6</b>	Criticality analysis by municipality showing the top 20 most relevant variables explaining SAIDI. 53	
<b>5-7</b>	Criticality analysis by circuit, highlighting environmental and structural interactions. . . . .	54
<b>6-1</b>	General architecture of the Agentic RAG-based chatbot for structured and unstructured queries.	56

<b>6-2</b>	Architecture of the recommendation system combining TabNet outputs with RAG. . . . .	57
<b>6-3</b>	Performance of ten LLMs in structured outage queries (BERTScore vs inference time). . . . .	59
<b>6-4</b>	Performance of six LLMs executed locally in structured outage queries. . . . .	60
<b>6-5</b>	Performance of ten LLMs in unstructured normative queries. . . . .	61
<b>6-6</b>	Performance of six LLMs executed locally in unstructured normative queries. . . . .	62
<b>6-7</b>	Performance of ten LLMs in recommendation queries. . . . .	63
<b>6-8</b>	Performance of six LLMs executed locally in recommendation queries. . . . .	63
<b>7-1</b>	Graph illustrating the relationship between the most influential variables across the top three critical assets. Color intensity denotes normalized relevance. . . . .	66
<b>7-2</b>	Reasoning graph showing how the LLM processes critical variables to generate technical recommendations. . . . .	66
<b>7-3</b>	Retrieval of relevant regulatory sources used by the LLM to contextualize the variable Pole Length. . . . .	67
<b>7-4</b>	Interpretation of specific compliance criteria for the variable Pole Length. . . . .	67
<b>7-5</b>	Graph showing asset contribution to SAIDI for the event. Larger circles indicate higher impact. . . . .	68

# List of Tables

<b>4-1</b>	Schema of EVENTOS . . . . .	18
<b>4-2</b>	Schema of APOYOS . . . . .	18
<b>4-3</b>	Schema of REDMT . . . . .	19
<b>4-4</b>	Schema of SWITCHES . . . . .	19
<b>4-5</b>	Schema of TRAFOS . . . . .	19
<b>4-6</b>	Meteorological variables and descriptions . . . . .	20
<b>4-7</b>	Summary of structured and unstructured documents used by the recommendation system . . . . .	21
<b>4-8</b>	Example questions by Q&A bitácora category . . . . .	22
<b>4-9</b>	Configuration of evaluated LLMs . . . . .	23
<b>4-10</b>	Table of Variables for Switches . . . . .	24
<b>4-11</b>	Table of Variables for Network Sections . . . . .	25
<b>4-12</b>	Table of Variables for Transformers . . . . .	25
<b>4-13</b>	Table of Variables for Utility Poles . . . . .	26
<b>5-1</b>	Performance comparison between TabNet and baseline regressors. TabNet achieved the best $R^2$ , with the added value of interpretability. . . . .	51
<b>6-1</b>	Pipeline for structured outage queries. . . . .	56
<b>6-2</b>	Pipeline for normative queries. . . . .	57
<b>6-3</b>	Pipeline for criticality-based recommendations. . . . .	57

# Content

<b>Acknowledgements</b>	<b>ii</b>
<b>Symbols and Abbreviations</b>	<b>iv</b>
Nomenclature . . . . .	iv
List of Abbreviations . . . . .	iv
<b>Resumen</b>	<b>v</b>
<b>Abstract</b>	<b>vi</b>
<b>List of Figures</b>	<b>viii</b>
<b>List of Tables</b>	<b>ix</b>
<b>Content</b>	<b>xii</b>
<b>1 Preliminaries</b>	<b>1</b>
1.1 Motivation . . . . .	1
1.2 Problem Statement . . . . .	4
1.2.1 Lack of data-driven models with predictive and explanatory capabilities . . . . .	4
1.2.2 Absence of integrated, interpretable decision support systems . . . . .	5
1.2.3 Research Question . . . . .	6
1.3 State of the Art . . . . .	7
1.3.1 Predictive and Interpretable Regression Models for MV-L2 Reliability . . . . .	7
1.3.2 Only-Decoder LLM Architectures as Generative Reasoning Engines . . . . .	8
1.3.3 Interpretable NLP-Based Decision Support Systems . . . . .	9
1.3.4 Knowledge Graphs for Interpretability in Intelligent Agent Systems . . . . .	10
1.3.5 Summary . . . . .	11
<b>2 Aims</b>	<b>12</b>
2.1 General Objective . . . . .	12
2.2 Specific Objectives . . . . .	12
<b>3 Outline and Contributions</b>	<b>13</b>
3.1 Predictive and Interpretable Modeling of SAIDI with TabNet . . . . .	14
3.2 Agentic RAG for Regulation-Aware Recommendations . . . . .	14

3.3	Interpretable Reasoning Graphs for Trust and Auditability . . . . .	14
<b>4</b>	<b>Materials and Methods</b>	<b>15</b>
4.1	Reliability Indicators in Medium-Voltage Distribution Networks . . . . .	15
4.1.1	Definition of SAIDI and SAIFI . . . . .	15
4.1.2	International Standards and Normative Context . . . . .	16
4.1.3	Operational Relevance in Medium-Voltage Networks . . . . .	16
4.1.4	Limitations of Current Use . . . . .	16
4.2	Datasets . . . . .	17
4.2.1	Structured Dataset for Reliability Prediction in MV-L2 Networks . . . . .	17
4.2.2	Unstructured Dataset for Regulation-Aware Recommendations . . . . .	20
4.2.3	Confidentiality Statement on the Internal Database Use . . . . .	26
4.3	Mathematical Background . . . . .	26
4.3.1	Classical Regressors . . . . .	26
4.3.2	TabNet Architecture . . . . .	29
4.3.3	Loss Functions for Regression . . . . .	33
4.3.4	Performance Metrics for Regression . . . . .	35
4.3.5	Transformer Architecture . . . . .	36
4.3.6	Training Objectives for LLMs . . . . .	39
4.3.7	Loss Functions for LLMs . . . . .	40
4.3.8	Performance Metrics for LLMs . . . . .	42
4.3.9	Vector Embeddings and Semantic Spaces . . . . .	43
4.3.10	Similarity Search Metrics . . . . .	43
4.3.11	Retrieval Performance Metrics . . . . .	44
4.3.12	Combined Loss in RAG . . . . .	46
<b>5</b>	<b>TabNet for Reliability and Criticality Analysis</b>	<b>48</b>
5.1	Methodology . . . . .	48
5.2	Experimental Framework . . . . .	49
5.3	Results and Discussion . . . . .	50
5.3.1	Convergence Behavior . . . . .	50
5.3.2	Comparison with Baseline Regressors . . . . .	51
5.3.3	Feature Importance and Interpretability . . . . .	51
5.3.4	Criticality Diagnosis by Municipality and Circuit . . . . .	53
5.4	Summary . . . . .	54
<b>6</b>	<b>Agentic RAG-based Conversational System for Asset Queries and Recommendations</b>	<b>55</b>
6.1	System Architecture . . . . .	55
6.1.1	Structured Queries: Interruption Databases . . . . .	56
6.1.2	Unstructured Queries: Regulatory Documents . . . . .	56
6.1.3	Criticality-Based Recommendations . . . . .	57
6.2	Experimental Framework . . . . .	58
6.3	Results and Discussion . . . . .	59
6.3.1	Structured Data Queries . . . . .	59
6.3.2	Unstructured Normative Queries . . . . .	61

6.3.3	Recommendation Queries . . . . .	62
6.3.4	Discussion . . . . .	64
6.4	Summary . . . . .	64
<b>7</b>	<b>Interpretable Reasoning Graphs for Trust and Auditability</b>	<b>65</b>
7.1	End-to-End Methodology and Reasoning Flow . . . . .	65
7.2	Results and Expert Validation . . . . .	68
7.3	Summary . . . . .	68
<b>8</b>	<b>Final Remarks</b>	<b>70</b>
8.1	Conclusions . . . . .	70
8.2	Future Work . . . . .	71
	<b>References</b>	<b>73</b>

# 1 Preliminaries

## 1.1 Motivation

Modern electric power distribution systems, especially at the medium-voltage level 2, play a crucial role in delivering reliable energy to both urban and rural areas. These systems are increasingly exposed to internal and external stressors—such as aging infrastructure, climate variability, and increasing load demand—that jeopardize their operational integrity and service continuity. In Electric Power Companies, unexpected outages at this network level have a direct impact on key performance metrics like the System Average Interruption Duration Index (SAIDI) and the System Average Interruption Frequency Index (SAIFI), which serve as global indicators of service reliability and customer satisfaction [Krstivojević & Stojković Terzić, 2025]. These metrics are also critical in the regulatory evaluation of service providers and influence investment decisions, operational planning, and customer trust. Addressing interruptions in MV-L2 networks has therefore become a strategic priority for EPCs aiming to ensure resilience, comply with regulations, and improve public perception of service quality [Seppälä & Järventausta, 2024].

In this context, Central Hidroeléctrica de Caldas (CHEC), a utility company belonging to the Grupo Empresas Públicas de Medellín (EPM), operates and maintains medium-voltage distribution networks in the Eje Cafetero region from Colombia <sup>1</sup> With a mix of urban, peri-urban, and rural circuits subject to diverse environmental and topographic conditions, CHEC faces significant operational challenges in maintaining network reliability. Moreover, the company is subject to regulatory pressure from national entities such as the CREG (Comisión de Regulación de Energía y Gas) and the Superintendencia de Servicios Públicos, which evaluate performance based on indicators like SAIDI and SAIFI [Delavechia et al., 2023]. These factors make the proactive identification of critical network components and the justification of operational decisions a high priority for the company [Dehghanian et al., 2011].

As part of its digital transformation strategy, CHEC is actively seeking intelligent, explainable, and cost-effective solutions that help bridge the gap between raw data and actionable decision-making [Troncia et al., 2023]. Traditional methods based on manual inspection or static reports often fall short in identifying patterns across vast amounts of heterogeneous data [Zhan et al., 2024]. Therefore, methodologies capable of combining predictive modeling with regulatory context and operational reasoning are of special interest to support technical and planning teams.

---

<sup>1</sup>“Somos CHEC”, página institucional de CHEC, [https://www.chec.com.co/Home/Institucional/Quienes-somos/Somos-CHEC?utm\\_source](https://www.chec.com.co/Home/Institucional/Quienes-somos/Somos-CHEC?utm_source). “Estructura corporativa del Grupo EPM”, sitio oficial de EPM, [https://www.epm.com.co/institucional/sobre-epm/gobierno-corporativo/estructura-grupo-epm/?utm\\_source](https://www.epm.com.co/institucional/sobre-epm/gobierno-corporativo/estructura-grupo-epm/?utm_source). Consultadas el 15 de septiembre de 2025.

Given the increasing complexity of these networks and the diverse factors influencing their performance, the ability to proactively predict, explain, and mitigate service interruptions has become a cornerstone of modern asset management and operational planning [Mortensen, 2024]. As illustrated in Figure 1-1, the schematic representation highlights the notion of criticality in MV-L2 networks, emphasizing the need for systematic approaches. While utilities often collect large volumes of historical fault and environmental data, the lack of a systematic, interpretable, and predictive methodology continues to hinder the transformation of this data into actionable knowledge [dechgummarn et al., 2023].

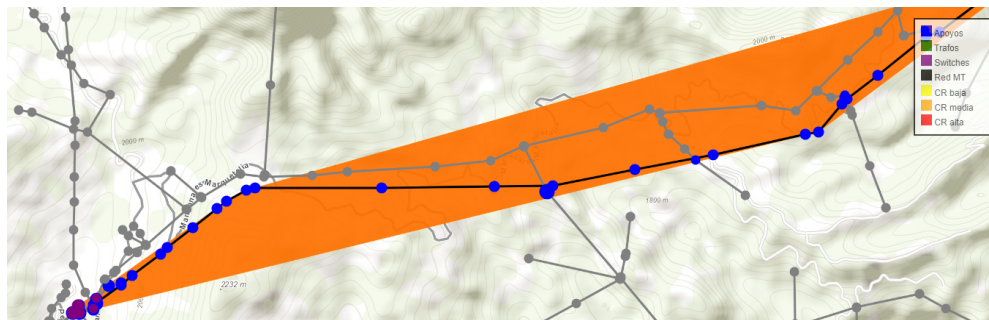


Figure 1-1: Schematic representation of criticality in MV-L2 networks.

Artificial Intelligence (AI) and Machine Learning (ML) have recently emerged as powerful tools to transform operational data into high-value insights across multiple utility domains, including outage prediction, failure classification, and asset prioritization [Ghasemkhani et al., 2024]. In the context of MV-L2 networks, these technologies enable the identification of latent relationships between operational failures and diverse influencing factors, such as meteorological conditions, equipment specifications, and historical maintenance patterns. Despite the availability of raw data, EPCs often lack robust AI-driven systems that integrate both predictive and prescriptive capabilities while remaining interpretable to field engineers and regulatory bodies [Saeed & Omlin, 2023]. Moreover, regulations such as RETIE (Reglamento Técnico de Instalaciones Eléctricas) or NTC (Normas Técnicas Colombianas) demand traceable and justifiable reasoning for maintenance and investment decisions, further reinforcing the need for explainable AI (XAI) in this sector [U.S. Commercial Service, 2021].

A critical limitation of most current AI applications is their "black box" nature, which limits adoption in high-stakes engineering environments where transparency, auditability, and regulatory compliance are non-negotiable [Teixeira et al., 2025]. Interpretability in AI systems is essential for trust, accountability, and adoption by utility companies, and it is increasingly mandated in regulatory frameworks governing infrastructure resilience and energy system transparency [Jørgensen & Ma, 2025].

This thesis is the direct result of an applied extension project developed in collaboration between the Universidad Nacional de Colombia, Manizales campus, and CHEC (Central Hidroeléctrica de Caldas), a utility company belonging to the Grupo EPM. The project, entitled "Advisory for the implementation of an intelligent dashboard for the diagnosis of medium-voltage level 2 networks, based on criticality analysis from endogenous and exogenous variables, and the generation of recommendations through natural language techniques," was designed to bridge the gap between academic research and industrial needs in the Colombian power sector. Its central aim was to deliver practical tools that enhance the monitoring, diagnosis, and decision-making processes of MV-L2 networks through data-driven methodologies.

Within the scope of this project, the research team sought to consolidate and analyze heterogeneous data sources, including lightning measurements, construction characteristics of circuits, historical records of service interruptions, and meteorological variables. These datasets were integrated into a unified platform to enable both descriptive and predictive analyses of network failures and asset criticality. By correlating operational conditions with outage patterns, the project sought to establish a systematic basis for preventive maintenance, investment prioritization, and risk management. This initiative responded not only to CHEC's operational

needs but also to the regulatory environment in which Colombian utilities operate, where reliability indicators such as SAIDI and SAIFI are continuously monitored by national entities.

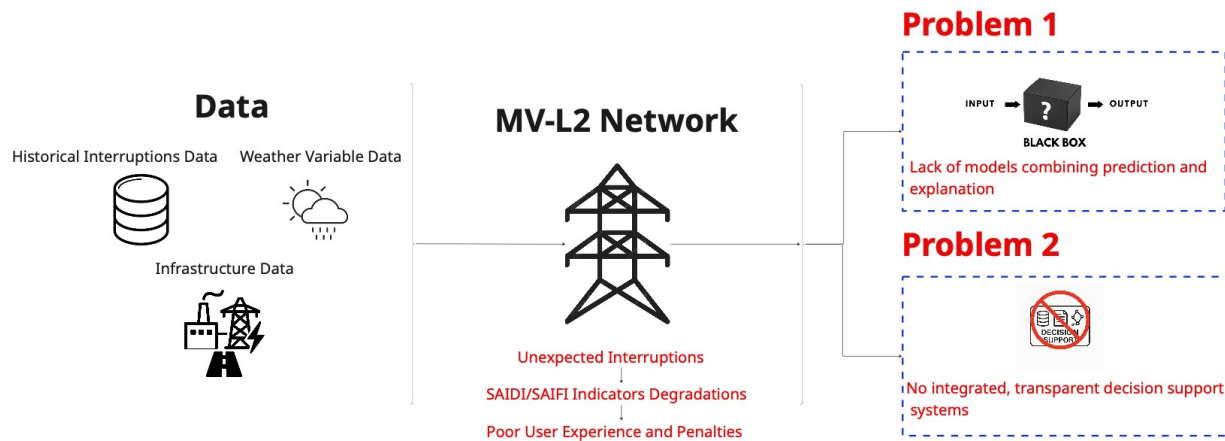
A distinctive feature of the project was the incorporation of Natural Language Processing (NLP) techniques to generate user-friendly decision support. Specifically, a chatbot-based assistant was conceived to provide engineers and planners with rapid access to key information and context-sensitive recommendations. Through NLP, the assistant was designed to query both structured data and unstructured regulatory documents, offering practical insights aligned with technical and compliance requirements. Additionally, a smart dashboard was developed to visualize the results of descriptive and predictive analyses, offering interactive and cloud-enabled functionalities for real-time monitoring and collaborative evaluation.

This joint initiative between Universidad Nacional and CHEC thus provided the academic and industrial foundation for the present thesis. By leveraging real operational challenges identified in CHEC's MV-L2 networks, the project framed the research questions and motivated the development of CRITAIR, a hybrid methodology that integrates interpretable predictive modeling with recommendation systems grounded in regulatory and operational knowledge. As such, the thesis not only contributes to the scientific community but also directly addresses pressing industry needs, reinforcing the role of university-industry collaboration in advancing reliable and explainable solutions for modern power systems.

## 1.2 Problem Statement

In Medium-Voltage Level 2 (MV-L2) networks operated by Electric Power Companies (EPCs), unexpected outages continue to degrade key reliability indicators such as the System Average Interruption Duration Index (SAIDI) and the System Average Interruption Frequency Index (SAIFI) [Ghasemkhani *et al.*, 2024]. This persistent degradation ultimately impacts end-user perception of service quality and exposes utilities to regulatory penalties and reputational risks [Zhu *et al.*, 2021]. A central reason for this issue is the lack of a systematic methodology that can accurately identify the internal and external variables most influencing these indicators, thereby limiting proactive asset management and hindering the prevention of future failures [Wang *et al.*, 2025]. This general problem is driven primarily by two causes, as illustrated in Figure 1-2:

First, the lack of models that can both predict reliability metrics and explain the underlying causes of service interruptions, particularly when incorporating external factors such as weather conditions or construction metadata [Wang *et al.*, 2025]; and second, the absence of integrated, interpretable decision-support systems that combine data insights from heterogeneous sources with domain knowledge, enabling the generation of clear, actionable, and trustworthy recommendations [Chatterjee & Dethlefs, 2020].



**Figure 1-2:** Illustration of the two key challenges identified: Lack of analytical models with predictive and explanatory capabilities and Absence of integrated, interpretable decision support systems

### 1.2.1 Lack of data-driven models with predictive and explanatory capabilities

In the operational context of utilities such as CHEC, the ability to forecast reliability indicators like SAIDI and SAIFI is indispensable for planning, maintenance, and regulatory compliance. However, most analytical tools currently deployed in the sector remain descriptive or narrowly focused on accuracy, often disregarding interpretability. Traditional regression or classification approaches tend to emphasize prediction without explaining the underlying drivers of interruptions, leaving decision-makers with limited understanding of why certain circuits or assets exhibit high vulnerability [Lin *et al.*, 2025]. This gap is particularly critical in MV-L2 networks, where the diversity of assets, geographies, and environmental stressors makes one-size-fits-all models insufficient.

Moreover, while utilities collect a wealth of operational and environmental information—including outage logs, equipment characteristics, and meteorological variables—few models adequately integrate these heterogeneous datasets. For example, CHEC operates in the Eje Cafetero region, where weather variability (e.g., intense rainfall, wind gusts, and cloud cover) has a tangible effect on outage frequency and duration

[Aldhubaib *et al.*, 2023b]. Failing to incorporate such exogenous factors reduces the explanatory power of models and diminishes their usefulness in preventive planning. Without understanding the interplay between external conditions and asset failures, EPCs risk underestimating vulnerabilities and reacting only after service quality has already been compromised.

Finally, the lack of interpretability in many AI systems exacerbates the problem. As illustrated in Figure 1-3, deep learning models, though accurate, are often treated as “black boxes” whose internal reasoning is opaque to field engineers and regulators [Zhou *et al.*, 2024]. In contexts where regulatory frameworks such as RETIE or NTC demand traceability and justifiable evidence for investment and maintenance decisions, black-box outputs cannot be directly adopted. Engineers and planners require models that not only predict but also highlight the most influential factors—such as conductor gauge, insulation age, or precipitation intensity—both globally and locally. In this sense, the absence of interpretable analytical models hinders CHEC’s ability to prioritize interventions, justify budgets, and comply with national oversight requirements.

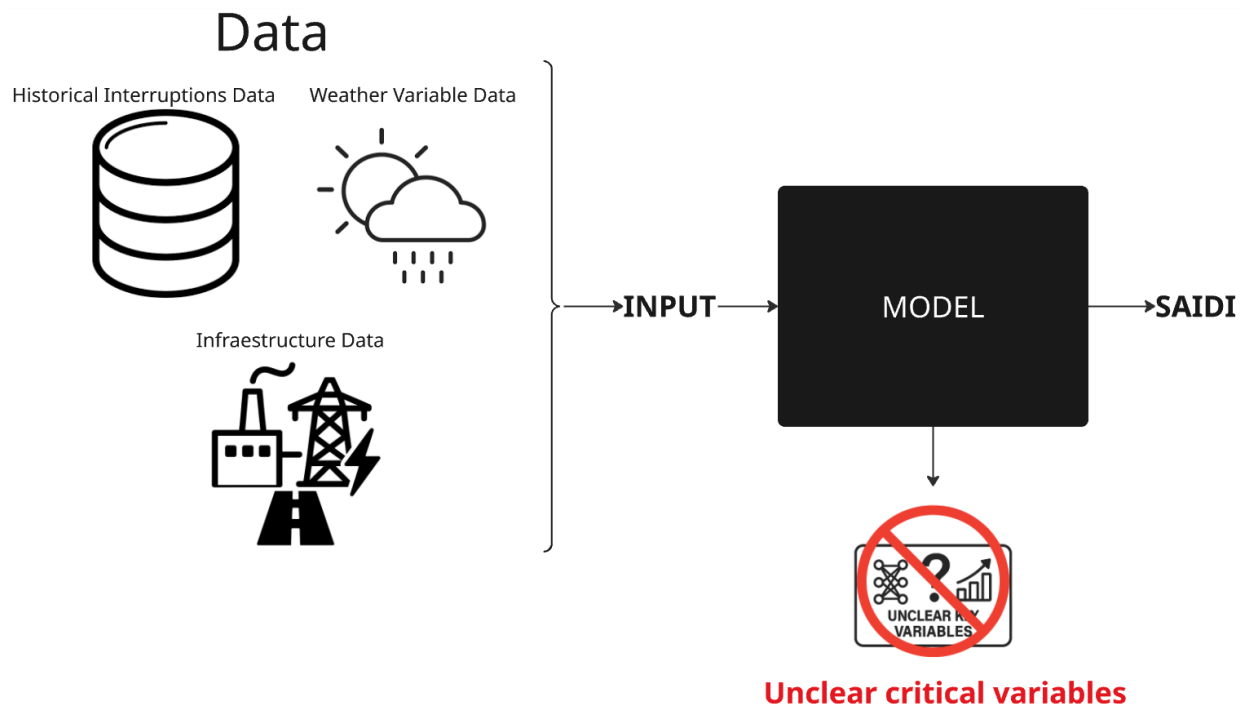


Figure 1-3: Lack of analytical models with predictive and explanatory capabilities graphical scheme

### 1.2.2 Absence of integrated, interpretable decision support systems

Even when predictive models are available, utilities like CHEC often struggle to bridge the gap between technical insights and actionable recommendations. This is because existing decision-making processes remain fragmented: on one side, engineers rely on statistical analyses or predictive models of outage data; on the other, they must separately interpret regulatory documents such as RETIE or NTC 2050, along with internal manuals and inspection reports [Kostopoulos *et al.*, 2024]. The lack of an integrated framework forces decision-makers to manually synthesize heterogeneous information, which is time-consuming, error-prone, and inconsistent across teams. As a result, recommendations for network reinforcement or preventive actions may fail to align with both predictive evidence and regulatory constraints.

Another challenge is the limited transparency of many AI-based decision support tools. When EPCs adopt commercial platforms or proprietary systems, these often generate recommendations without showing the

reasoning behind them [Shadi *et al.*, 2025]. In high-stakes sectors such as energy distribution, decision-makers require interpretable systems that justify why a specific asset or region is critical, what regulatory clauses are implicated, and how each recommendation links to historical and environmental data. Without such transparency, recommendations lack credibility, reducing trust among engineers, managers, and regulators. For CHEC, which is subject to evaluations by national authorities like CREG and the Superintendencia de Servicios Públicos, the absence of transparent, auditable reasoning chains creates additional compliance and reputational risks.

Lastly, the siloed treatment of structured and unstructured data prevents utilities from fully exploiting their knowledge base. As illustrated in Figure 1-4, CHEC's operational history includes not only databases of outage events but also vast textual resources such as technical regulations, maintenance logs, and field inspection notes. Current practices rarely integrate these sources into a unified decision-support pipeline, meaning valuable regulatory or operational knowledge remains disconnected from predictive insights. The consequence is a reactive rather than proactive planning process, where decisions are made without full consideration of regulatory obligations or latent risk patterns. Thus, the absence of integrated, interpretable decision support systems severely limits the transformation of raw data and documentation into trustworthy, actionable knowledge.

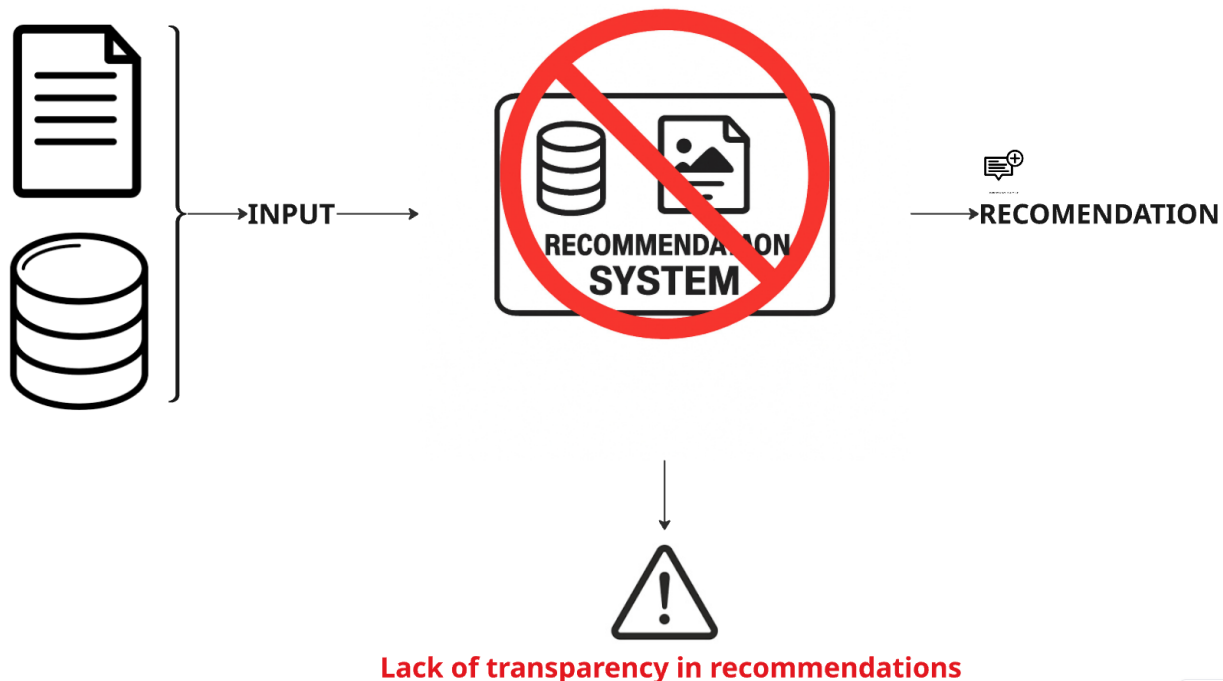


Figure 1-4: Absence of integrated, interpretable decision support systems graphical scheme

### 1.2.3 Research Question

How can we design and implement a systematic, interpretable AI-based methodology that integrates predictive modeling of MV-L2 reliability indicators with regulatory and operational context, in order to identify the most influential internal and external variables affecting SAIDI and SAIFI and generate transparent, actionable recommendations for Electric Power Companies?

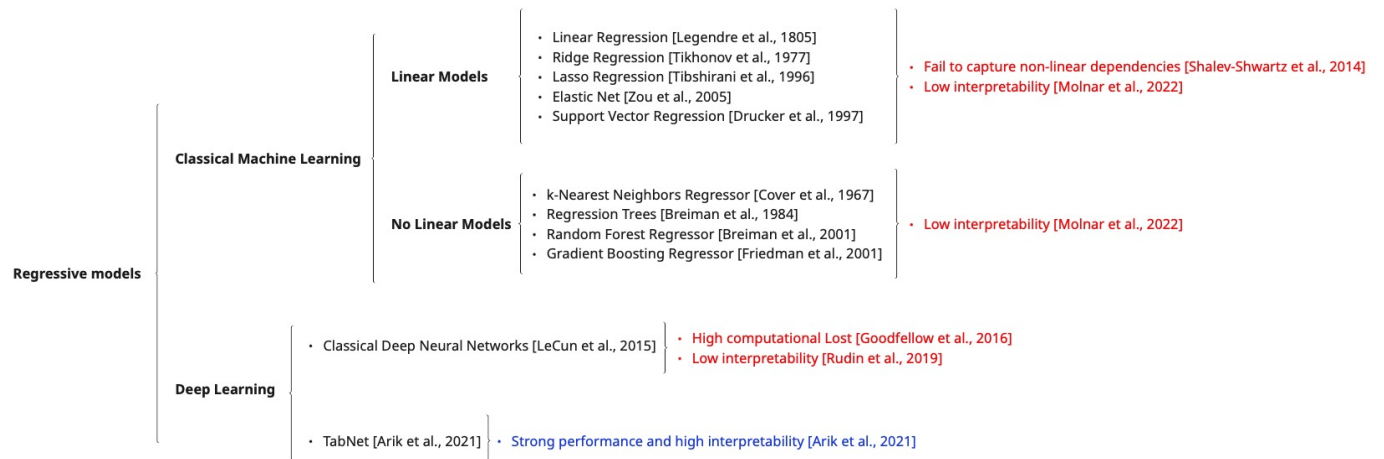
## 1.3 State of the Art

This section presents an overview of the state-of-the-art approaches addressing the two key challenges identified in Section 1.2. First, we review the evolution of regression-based and interpretable models for predicting reliability indicators such as SAIDI and SAIFI, with a particular focus on those capable of integrating exogenous variables. Then, we explore the progression of natural language processing (NLP)-driven decision support systems, ranging from simple question-answering (QA) models to advanced architectures like Agentic and Multi-Agent Retrieval-Augmented Generation (RAG). Each part highlights recent trends, major advantages, and practical limitations.

### 1.3.1 Predictive and Interpretable Regression Models for MV-L2 Reliability

In medium-voltage level 2 (MV-L2) networks, predicting reliability indicators such as SAIDI and SAIFI has traditionally relied on basic statistical regressors. Linear models are widely used due to their ease of implementation and low computational cost. However, their inability to capture non-linear relationships or incorporate external variables—such as weather conditions—limits their effectiveness [Shalev-Shwartz & Ben-David, 2014]. To address this, more flexible models like Random Forests and XGBoost have been applied. These models handle complex interactions and achieve higher accuracy, but are often considered black boxes; their internal logic is difficult to interpret, and feature importance scores are insufficient to support regulated decision-making processes [Rudin, 2019].

More recently, deep learning architectures such as deep neural networks (DNNs) have demonstrated strong performance in predicting reliability indices, particularly when trained on large historical datasets that include exogenous variables. However, their lack of transparency makes them unsuitable for high-stakes and regulated contexts [Rudin, 2019]. This has led to the emergence of TabNet, a deep learning architecture designed specifically for tabular data. TabNet uses sparse attention and sequential feature selection to identify the most relevant variables for each instance. It provides both global and local interpretability—crucial for utilities operating under regulatory oversight. Moreover, TabNet naturally incorporates exogenous factors such as precipitation, wind gusts, and conductor gauge, and can quantify their relative importance in reliability estimation tasks [Arik & Pfister, 2021].



**Figure 1-5:** Comparative spider chart of major families of regression models for MV-L2 reliability prediction: Linear classical models, non-linear classical ML models, deep neural networks, and attention-based tabular models such as TabNet

In summary, while simple models like linear regression are easy to implement, they fall short in modeling complexity and external dependencies. Ensemble models such as Random Forest Regressor offer better accuracy but lack transparency. TabNet represents a modern, balanced alternative—accurate, interpretable, and exogenously aware—making it highly suitable for reliability analysis in MV-L2 networks. Its limitations include higher computational cost and the need for well-preprocessed data.

### 1.3.2 Only-Decoder LLM Architectures as Generative Reasoning Engines

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 for selecting models suitable for explainable, regulation-aware reliability systems.

Only-encoder models, such as BERT, RoBERTa, and DistilBERT, rely on bidirectional transformers that contextualize input sequences without generating text [Devlin et al., 2018, Liu et al., 2019]. They excel in extractive and discriminative tasks, including text classification, entity recognition, and span-based question answering. Their deep bidirectional attention enables fine-grained understanding of input context. However, their lack of generative capability prevents their direct application to tasks requiring the production of coherent textual explanations, summaries, or recommendations—core requirements in 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 [Brown et al., 2020, Touvron et al., 2023, Bai et al., 2025]. 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.

Encoder–decoder models combine both paradigms, using a dedicated encoder to process the input and a decoder to generate outputs [Raffel *et al.*, 2020, Lewis *et al.*, 2019]. 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.

In this study, all evaluated models belong to the only-decoder architecture family. This design choice aligns with the operational requirements of the CRITAIR framework, which prioritizes real-time reasoning, long-context understanding, and generative explainability. Only-decoder LLMs, when grounded through Retrieval-Augmented Generation (RAG) and Agentic reasoning mechanisms, can integrate analytical signals from structured reliability datasets with normative knowledge sources (e.g., RETIE and NTC 2050) to produce coherent, regulation-aware textual outputs. Their generative flexibility enables CRITAIR to operate as a conversational decision-support system capable of producing traceable and interpretable justifications under regulatory scrutiny.

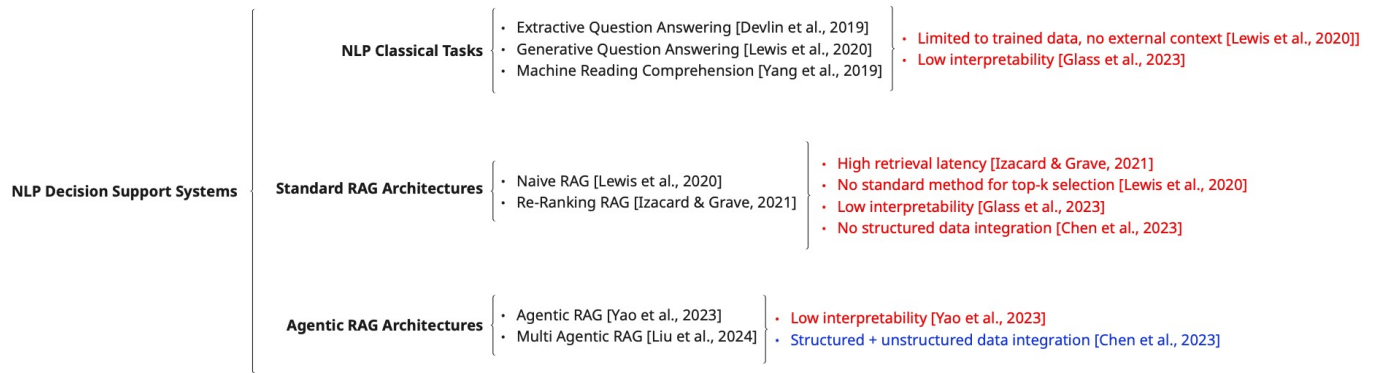
By leveraging only-decoder LLMs, CRITAIR achieves an effective balance between generative expressiveness and regulatory grounding. These models constitute the linguistic and cognitive core that enables higher-level reasoning layers—such as RAG, Agentic RAG, and GraphRAG—to fuse structured reliability data with textual regulatory evidence. The adoption of only-decoder architectures therefore establishes the computational backbone of the CRITAIR framework, supporting context-aware reasoning, narrative synthesis, and multi-source evidence integration. This convergence between interpretable predictive modeling and regulation-grounded generative reasoning defines the next research frontier: decision-support systems that extend beyond prediction into explainable, audit-ready recommendations.

### 1.3.3 Interpretable NLP-Based Decision Support Systems

The second challenge relates to the absence of systems that can transform predictive outputs into understandable, traceable, and regulation-aware recommendations. Recent developments in this space have progressed from basic QA systems to advanced retrieval-augmented reasoning pipelines capable of integrating structured and unstructured information in real time [Löwenmark *et al.*, 2025].

Early approaches used LLM-based QA systems, allowing users to query technical or regulatory documents such as RETIE or NTC 2050. These systems are easy to deploy and useful for straightforward queries, but they suffer from hallucinations, lack of traceability, and limited contextual understanding [Karpukhin *et al.*, 2020]. To improve upon this, Classic RAG architectures emerged. These combine semantic document retrieval with grounded text generation, reducing hallucinations and enhancing factual consistency. While these systems significantly improve over vanilla QA, they still rely on single-step query-response logic and are limited in handling structured (e.g., tabular) data or reasoning over time [Trangcasanchai, 2024].

A major breakthrough is the development of Agentic RAG architectures, where autonomous agents can plan, decompose, and execute multi-step reasoning using internal and external tools. These agents can access structured databases, regulatory documents, and field logs simultaneously, enabling contextual and auditable recommendations. Variants such as Modular RAG, which separate retrieval, parsing, and reasoning modules, and Multi-Agent RAG, where agents interact in parallel or hierarchical settings, further extend the system’s capabilities [Böckling *et al.*, 2025]. While these architectures offer the highest levels of interpretability and integration, they require more advanced orchestration, dedicated infrastructure, and custom evaluation protocols [Chen *et al.*, 2025].



**Figure 1-6:** Comparative spider chart of NLP-based decision-support system families: NLP Classic Tasks, Standard RAG Architectures and Agentic RAG Architectures

In conclusion, simple QA systems are useful but insufficient for mission-critical recommendations. Classic RAG architectures introduce document grounding, but fall short in adaptability and structured integration. Agentic and Multi-Agent RAG systems represent the current frontier in interpretable decision support for utilities—capable of reasoning across diverse data sources and aligning with regulatory demands. Their implementation, however, remains technically demanding.

### 1.3.4 Knowledge Graphs for Interpretability in Intelligent Agent Systems

Beyond regression models and Retrieval-Augmented Generation (RAG) systems, recent research highlights the role of *Knowledge Graphs* (KGs) as a means of enhancing interpretability and reasoning in intelligent agents. KGs explicitly represent entities, their attributes, and relationships, enabling structured reasoning that complements the statistical predictions of machine learning models. This representation makes it possible to integrate heterogeneous knowledge—ranging from outage logs and equipment metadata to regulatory clauses—into a unified semantic framework that agents can query and explain in auditable ways.

In the energy sector, KGs have been increasingly applied for **fault diagnosis and asset management**. Chen et al. [Chen et al., 2022] demonstrate how KGs can represent the lifecycle of electrical equipment and encode causal relations among operating conditions, environmental stressors, and failure events, thereby facilitating knowledge-driven operation and maintenance strategies that are inherently interpretable. Similarly, Li et al. [Li et al., 2023] propose a domain-specific KG for power system failures, incorporating equipment entities, semantic concept graphs, business logic graphs, and historical failure cases, which collectively support transparent reasoning over operational and regulatory contexts.

From the perspective of **explainable reasoning**, Zhang et al. [Zhang et al., 2023] introduce a “rule-enhanced cognitive graph” that embeds logical rules into the KG structure, allowing agents not only to retrieve information but also to infer and justify causal relationships in power grid operation. More generally, Tang et al. [Tang et al., 2019] emphasize that KGs are effective for integrating fragmented operational data, inspection records, and technical manuals into a relational structure, thus reducing information silos and improving knowledge reuse.

In the field of **NLP-driven decision support**, several works combine KGs with RAG architectures. Recent frameworks such as *GraphRAG* extend standard RAG by embedding structured knowledge graphs alongside vector indices, enabling agents to ground their outputs in explicit relational structures rather than in isolated document fragments [Team, 2024]. Other approaches, such as KG-SMILE [Bouadi et al., 2025], further enhance transparency by attributing which entities and relations in the KG influenced the generated recommendations, thus providing interpretable evidence for end-users.

The main advantage of KGs in this context lies in their ability to provide **traceable, regulation-aware explanations**. For instance, if precipitation and conductor gauge are identified as key risk drivers, the KG can simultaneously retrieve related RETIE or NTC clauses and connect them to historical outage cases, offering a reasoning chain that is both data-driven and regulation-compliant. Nevertheless, challenges remain, including the need to construct robust ontologies, ensure continuous updates with new failure cases and regulatory changes, and maintain scalability in multi-hop reasoning tasks.

For the purposes of the **CRITAIR methodology**, incorporating a Knowledge Graph layer would directly reinforce the interpretability objective. The KG can serve as the backbone of the reasoning graphs already envisioned in the system, explicitly mapping critical variables, related regulatory fragments, and historical precedents. This integration not only strengthens the transparency and trustworthiness of recommendations but also ensures compliance with regulatory frameworks by making the decision-making process auditable and reproducible.

### 1.3.5 Summary

In summary, addressing reliability in MV-L2 networks requires a dual strategy. First, models like TabNet offer interpretable, high-accuracy predictions with native support for exogenous variables—bridging the gap between performance and transparency. Second, Agentic RAG systems provide contextualized, auditable decision support by combining structured and unstructured knowledge sources, including regulatory frameworks. While both technologies show great promise, they require careful deployment, domain adaptation, and rigorous evaluation to ensure trustworthy, regulatory-compliant results. Together, they form the basis for the proposed CRITAIR methodology.

## 2 Aims

The analysis above leads us to the following general and specific objectives:

### 2.1 General Objective

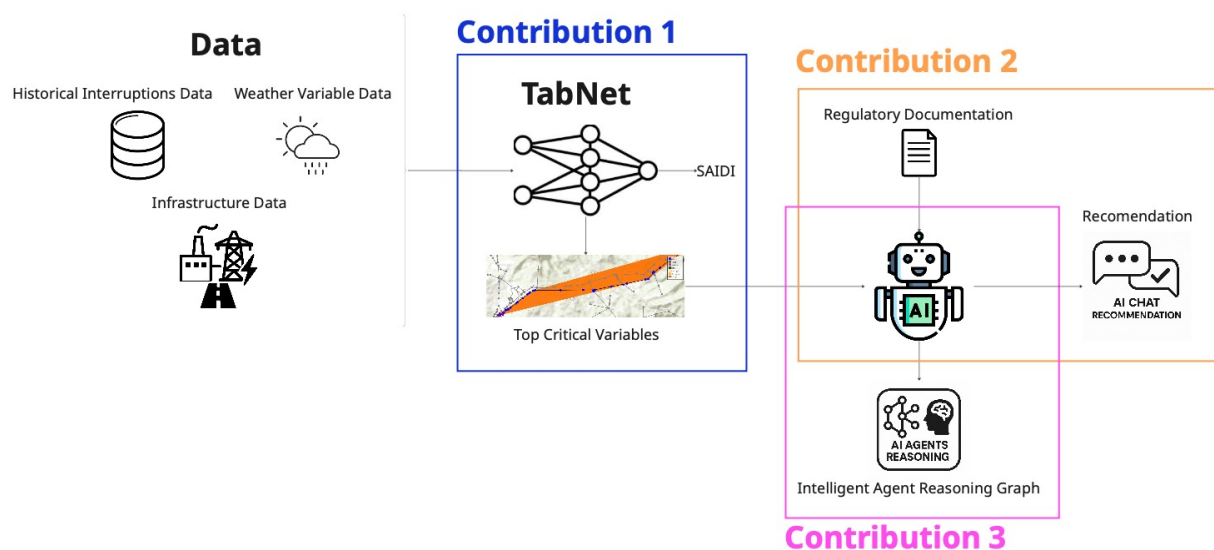
To design and implement an interpretable AI-based methodology that integrates predictive modeling with regulatory and operational knowledge, in order to identify the internal and external factors influencing SAIDI and SAIFI in MV-L2 networks, and to generate transparent, actionable recommendations for Electric Power Companies (EPCs). The proposed methodology will be validated through regression metrics (e.g.,  $R^2$  score) for SAIDI prediction, feature attribution analysis for explanation, NLP evaluation scores (e.g., BERTScore) for recommendation generation and expert-based assessment conducted by CHEC engineers in real operational scenarios.

### 2.2 Specific Objectives

- 1 To develop a predictive and explanatory model using TabNet, capable of estimating SAIDI values based on historical outage data enriched with exogenous variables (e.g., meteorological and construction metadata), and to evaluate its performance through regression accuracy ( $R^2$ ) and global/local feature importance rankings.
- 2 To implement an Agentic RAG system that integrates structured data and unstructured regulatory documents, enabling the generation of domain-aware recommendations, and to assess its effectiveness using NLP evaluation metrics (e.g., BERTScore, which measures semantic similarity between a target text and an LLM-generated sentence in a high-dimensional embedding space) across tasks such as tabular reasoning, regulation interpretation, and recommendation synthesis.
- 3 To enhance the transparency and trustworthiness of the system by generating interpretable reasoning graphs that visualize the decision-making process of the AI agent, and to validate their utility through expert-based qualitative evaluation in real-world operational scenarios within CHEC.

## 3 Outline and Contributions

The following section outlines the key contributions of this thesis, which are visually summarized in Figure 3-1. Broadly, this work proposes a novel, interpretable AI-based methodology designed to improve the reliability of Medium-Voltage Level 2 (MV-L2) power networks. This is achieved through the integration of predictive modeling, regulatory reasoning, and recommendation systems validated in real-world scenarios with expert feedback from CHEC engineers.



**Figure 3-1:** Schematic display of the main thesis contributions, including the predictive and interpretable TabNet architecture for SAIDI estimation and feature attribution, the integration of an Agentic RAG system that combines structured outage data with unstructured regulatory knowledge to generate recommendations, and the development of interpretable reasoning graphs that provide both quantitative and qualitative transparency of the intelligent agent’s decision-making process.

First, the thesis introduces a predictive and interpretable modeling pipeline based on TabNet, capable of estimating SAIDI indicators using enriched historical outage records that include meteorological and construction-related metadata. This stage enables both regression and explanatory tasks by identifying the most influential internal and external variables affecting reliability metrics. Second, the methodology incorporates a Retrieval-Augmented Generation (RAG) system with agentic behavior to bridge structured operational data and unstructured regulatory documents. This system allows decision-makers to query the model and receive contextualized, regulation-aware recommendations about asset criticality, risk mitigation, and maintenance actions. Lastly, the thesis proposes a visualization layer in the form of interpretable reasoning graphs. These graphs reveal the full decision-making process of the AI agent, showing which features were prioritized, which regulatory fragments were retrieved, and how the final recommendation was

formulated. This is key to ensuring trust, explainability, and regulatory compliance in real-world energy operations.

## 3.1 Predictive and Interpretable Modeling of SAIDI with TabNet

As discussed in the background and problem analysis, a major challenge in MV-L2 reliability management is the absence of analytical models that are both predictive and explainable. This thesis addresses this gap by implementing a TabNet model, chosen for its native interpretability through sparse attention mechanisms and sequential feature selection [Arik & Pfister, 2021]. The model is trained on historical outage data enriched with meteorological variables (e.g., rainfall, wind gusts, cloud cover) and construction metadata (e.g., conductor gauge, insulation type). The model is configured in both regression and autoencoder modes, allowing for accurate estimation of SAIDI while uncovering global and local feature importance. The resulting explanations provide utility engineers with concrete evidence on which variables are driving the most critical interruptions across different geographical and operational contexts.

## 3.2 Agentic RAG for Regulation-Aware Recommendations

A second major contribution of this work is the design and implementation of an Agentic Retrieval-Augmented Generation (RAG) system that supports intelligent recommendations by combining structured operational data and unstructured regulatory documentation. This system uses semantic retrieval to access relevant fragments from technical regulations (such as RETIE and NTC 2050) and historical reports, and combines them with the output of the TabNet model to generate recommendations that are both context-aware and regulation-compliant. The agent is capable of answering questions such as “Why is this circuit critical?”, “Which regulation applies?”, and “What maintenance actions are recommended?” — all with traceable logic and citation of sources. The RAG system is validated through NLP evaluation metrics such as BERTScore and through qualitative feedback from domain experts at CHEC, who reviewed the generated recommendations in operational scenarios [Singh *et al.*, 2025].

## 3.3 Interpretable Reasoning Graphs for Trust and Auditability

The final contribution focuses on interpretability and trust. The thesis introduces reasoning graphs that visualize the full decision-making path taken by the AI system. These graphs include nodes representing critical variables, retrieved regulatory evidence, and inference steps made by the language model leading to a recommendation. This design directly addresses the “black box” issue of many AI models, and aligns with explainability standards increasingly required by regulatory agencies and internal utility auditing processes [Dorji *et al.*, 2025]. By making the reasoning process explicit, the system ensures that operational decisions can be justified, audited, and trusted — both internally and externally. This contribution was validated through expert-based evaluations at CHEC, where engineers confirmed the utility of the graphs for transparent, explainable decision support in the planning and maintenance of MV-L2 networks.

## 4 Materials and Methods

### 4.1 Reliability Indicators in Medium-Voltage Distribution Networks

The evaluation of service reliability in electrical distribution systems is commonly based on standardized indices that quantify both the frequency and duration of interruptions experienced by end users. Among these, the **System Average Interruption Duration Index (SAIDI)** and the **System Average Interruption Frequency Index (SAIFI)** are the two most widely adopted worldwide. These indices are not only technical measures but also carry regulatory significance, as they are directly linked to service quality assessment, penalty mechanisms, and incentive schemes imposed by supervisory agencies. In the Colombian context, regulatory entities such as the Comisión de Regulación de Energía y Gas (CREG) and the Superintendencia de Servicios Públicos Domiciliarios (SSPD) use these indicators to evaluate utility performance and enforce compliance with the *Régimen de Calidad del Servicio* [cre, 2008, cre, 2018, ret, 2022].

#### 4.1.1 Definition of SAIDI and SAIFI

*SAIDI.* The System Average Interruption Duration Index (SAIDI) represents the average total duration of sustained interruptions (typically longer than 3 minutes) that a customer experiences during a predefined period, usually one year [iee, 2012a, Billinton & Allan, 1996a]. It is formally defined as:

$$SAIDI = \frac{\sum_{i=1}^N U_i \cdot N_i}{N_T},$$

where  $U_i$  is the interruption duration for event  $i$  (in hours),  $N_i$  is the number of customers affected by interruption  $i$ , and  $N_T$  is the total number of served customers. SAIDI is expressed in hours per customer per year, and lower values indicate better performance [U.S. Energy Information Administration, 2023a].

*SAIFI.* The System Average Interruption Frequency Index (SAIFI) quantifies the average number of sustained interruptions that a customer experiences over the same period. It is defined as [iee, 2012a, Billinton & Allan, 1996b]:

$$SAIFI = \frac{\sum_{i=1}^N N_i}{N_T}, \tag{4-1}$$

where  $N_i$  and  $N_T$  retain their previous definitions. SAIFI is expressed in interruptions per customer per year

[*U.S. Energy Information Administration, 2023b*].

Together, SAIDI and SAIFI provide complementary perspectives: the former captures the cumulative outage time, while the latter measures outage recurrence [*Billinton & Allan, 1996b, iee, 2012a*].

### 4.1.2 International Standards and Normative Context

The calculation of SAIDI and SAIFI follows guidelines established by the **IEEE Standard 1366-2012** [iee, 2012b], which defines sustained interruptions, exclusion criteria (e.g., force majeure events), and normalization procedures. These indices are widely recognized by regulatory agencies, system operators, and utilities as benchmarks for reliability assessment.

In **Colombia**, the CREG has adopted SAIDI and SAIFI as central metrics for evaluating the quality of distribution service. Through resolutions such as **CREG 097 of 2008** and subsequent updates (e.g., CREG 015 of 2018), utilities are required to report monthly and annual values of these indices. Compliance is enforced by the SSPD, which applies penalties or incentives depending on whether utilities exceed or fall below reference thresholds [cre, 2008, cre, 2018].

Additionally, the **RETIE** (*Reglamento Técnico de Instalaciones Eléctricas*) establishes safety and reliability criteria that indirectly affect outage performance. For example, RETIE specifies design and maintenance requirements for medium-voltage components (transformers, poles, switches, conductors), whose failure rates directly impact SAIDI and SAIFI levels [ret, 2022]. Therefore, these indices serve as both technical and regulatory connectors between asset condition, operational performance, and user satisfaction.

### 4.1.3 Operational Relevance in Medium-Voltage Networks

Medium-voltage level 2 (MV-L2) networks, which operate at nominal voltages up to 33 kV, are particularly critical for reliability assessment. Outages at this level affect large numbers of customers simultaneously and thus dominate the contribution to SAIDI and SAIFI. In regions such as the Colombian Eje Cafetero, environmental stressors like heavy rainfall, wind gusts, and lightning strikes exacerbate vulnerability, making accurate estimation and reduction of SAIDI/SAIFI an operational priority [*Aldhubaib et al., 2023a, Zhu et al., 2021*].

For Electric Power Companies (EPCs) such as CHEC, maintaining SAIDI and SAIFI within regulatory limits is essential not only for avoiding penalties but also for preserving reputation and customer trust. These indices provide actionable feedback loops: high SAIFI values may suggest recurrent issues with specific feeders or components, while elevated SAIDI values highlight weaknesses in restoration procedures or asset redundancy. Thus, predictive and explanatory modeling of SAIDI and SAIFI enables utilities to move from reactive correction to proactive reliability management.

### 4.1.4 Limitations of Current Use

While SAIDI and SAIFI are invaluable for benchmarking, they have limitations. Both are aggregate indicators that do not account for spatial or temporal heterogeneity, nor do they distinguish between critical and non-critical customers. Additionally, their explanatory capacity is limited: knowing that SAIDI is high does not reveal whether the cause is environmental, technical, or operational. For this reason, recent research advocates for combining SAIDI/SAIFI with predictive modeling, exogenous variable integration, and

interpretable decision-support systems [Wang et al., 2025, Chatterjee & Dethlefs, 2020]. These limitations motivate the development of methodologies such as CRITAIR, which aim not only to predict reliability indices but also to explain their underlying drivers and suggest actionable mitigation strategies.

## 4.2 Datasets

This section describes the two datasets used in this thesis: (i) a structured dataset composed of historical outage records, MV-L2 asset metadata, and meteorological variables, which was used to train and evaluate the predictive model; and (ii) an unstructured dataset composed of regulatory texts and field documentation, which supported the evaluation of the interpretability and regulatory alignment of the recommendation system. Both datasets were integrated into a unified decision-support framework through a hybrid AI pipeline.

### 4.2.1 Structured Dataset for Reliability Prediction in MV-L2 Networks

To train and evaluate the predictive and interpretable model developed in this thesis, we constructed a structured dataset derived from historical MV-L2 network records, including outage events, asset metadata, and exogenous weather information. This dataset was obtained through a rigorous ETL (Extract, Transform, Load) pipeline designed to consolidate heterogeneous sources, standardize formats, enrich technical attributes, and ensure temporal alignment across tables.

The structured data component is composed of the following main sources:

- **Outage Events:** Cleaned and preprocessed version of the `EVENTOS.csv` file, including unplanned interruptions from 2019 to June 2024. Filtering was applied to retain only interruptions at voltage levels  $\leq 33\text{kV}$  and durations  $\leq 100$  hours. Categorical codes were mapped to descriptive labels for interpretability.
- **Asset Metadata:** Equipment-specific datasets describing characteristics of poles (`APOYOS.csv`), line segments (`REDMT.csv`), switches (`SWITCHES.csv`), and transformers (`TRAFOS.csv`). Each file was homogenized, enriched, and filtered to retain only operational and connected components.
- **Meteorological Variables:** For each outage event, 24-hour antecedent weather data was collected via API and merged, allowing the model to capture environmental dependencies on network reliability.

All data were handled in `.csv` format using Python (Pandas, NumPy). Processing and training were performed on a local workstation (32GB RAM, Intel Core i9 CPU, 2TB SSD) and a dedicated GPU server with an NVIDIA RTX A6000 GPU. The final dataset supported both regression tasks and global/local feature attribution analysis.

## Summary of Structured Tables

Table 4-1: Schema of EVENTOS

Column	Data Type
evento	int64
equipo_ope	object
tipo_equi_ope	object
clo_equi_ope	object
tipo_elemento	object
inicio	datetime64[ns]
fin	datetime64[ns]
duracion_h	float64
tipo_duracion	object
causa	object
CNT_TRAFOS_AFEC	int64
cnt_usus	int64
SAIDI	float64
SAIFI	float64

Table 4-2: Schema of APOYOS

Column	Data Type
CODE	object
ASSEMBLY	object
XPOS	float64
YPOS	float64
TOWNER	object
FECHA	datetime64[ns]
FECHA_C	period[M]
TIPO	object
CLASE	object
MATERIAL	object
LONG_APOYO	float64
VIEN_PRIM	float64
VIEN_SEC	float64
TIERRA_PIE	object

Table 4-3: Schema of REDMT

Column	Data Type
CODE	object
PHASES	int64
FPARENT	object
ELNODE1	object
ELNODE2	object
CONDUCTOR	object
NEUTRAL	object
LENGTH	float64
KVNOM	float64
TOWNER	object
FECHA	datetime64[ns]
CALIBRECONDUCTOR	object
CAPACITY	float64
RESISTANCE	float64
LATITUD, LONGITUD	float64
DEP, MUN	object

Table 4-4: Schema of SWITCHES

Column	Data Type
CODE	object
PHASES	int64
ASSEMBLY	object
STATE	object
FECHA	datetime64[ns]
CALIBRECONDUCTOR	object
CAPACITY	float64
RESISTANCE	float64
LATITUD, LONGITUD	float64

Table 4-5: Schema of TRAFOS

Column	Data Type
CODE	object
PHASES	int64
OWNER1	object
TRFTYPE	object
TIPO_SUB	object
IMPEDANCE	float64
FECHA_ACT	datetime64[ns]
KVA	float64
KV1	float64
LATITUD, LONGITUD	float64
DEP, MUN	object

## Meteorological Variables

Each outage event was enriched with 24-hour antecedent weather data obtained through an external API. A total of 11 meteorological variables were incorporated into the structured dataset:

**Table 4-6:** Meteorological variables and descriptions

Variable	Description
precip	Precipitation (mm) – associated with grounding failures and insulation moisture
pres	Local atmospheric pressure – influences thermal dissipation and dielectric strength
rh	Relative humidity (%) – relevant to corrosion and partial discharges
slp	Sea level pressure (hPa) – contextual supplement to local pressure
solar_rad	Solar radiation ( $W/m^2$ ) – affects degradation of outdoor components
temp	Ambient temperature ( $^{\circ}C$ ) – impacts transformer and conductor performance
uv	UV index – accelerates degradation of polymeric materials
vis	Visibility (km) – supports maintenance planning and field logistics
wind_gust_spd	Wind gust speed (m/s) – linked to mechanical stress on structures
wind_spd	Average wind speed (m/s) – impacts line vibration and conductor stability
wind_dir	Wind direction (degrees) – useful for directional exposure analysis

### 4.2.2 Unstructured Dataset for Regulation-Aware Recommendations

In order to develop the regulatory-aware recommendation component of the CRITAIR methodology, a comprehensive corpus of unstructured domain-specific documents was constructed, vectorized, and evaluated under a rigorous multi-agent benchmarking framework. This dataset provides the semantic foundation upon which the intelligent agent makes regulation-compliant decisions and suggestions.

#### Document Corpus: Regulatory and Operational Sources

To support the intelligent recommendation system within the CRITAIR methodology, a heterogeneous document corpus was compiled. This corpus includes both structured operational records and unstructured regulatory documents, ensuring coverage of technical, normative, and contextual knowledge relevant to medium-voltage grid assets.

These documents were stored in .pdf, .docx, and .txt formats and loaded using a recursive character-based text splitter with a maximum token limit adapted to the embedding model. Metadata tags were added to each chunk for component type, standard source, and reliability focus.

All documents were embedded using `text-embedding-3-small` from OpenAI and stored in a ChromaDB vector database. This allowed for high-quality semantic search and RAG-based retrieval across both regulation and operations-related documents.

A comprehensive summary of these documents and their regulatory or operational role is shown in Table 4-7.

**Table 4-7:** Summary of structured and unstructured documents used by the recommendation system

Document	Description
Switch and Transformer Interruption Tables	Historical records of outage events, including date, time, event type, duration, and affected equipment. Data is stored in structured formats to facilitate querying and analysis using tools such as <b>pandas</b> .
RETIE Regulations	Technical regulation for electrical installations in Colombia. Stored in unstructured formats (PDFs) and indexed in a vector database to support efficient retrieval using NLP techniques.
RETIE General Provisions	Define safety requirements to protect people, animals, and the environment from electrical risks, as well as standards for electrical equipment and products.
APOYOS_LONG_APOYO	Guidelines on the required length and dimensions of utility poles, including adjustment recommendations in case of noncompliance.
CARGA_TRABAJO_APOYOS	Normative parameters on permissible load capacities under various operational conditions.
NORMATIVA_APOYOS_TIERRA_PIE	Specifications to ensure proper and safe grounding in utility poles.
VIENTO_APOYOS	Standards on structural wind resistance for pole-mounted structures.
RAYOS_APOYOS	Protection requirements against lightning and electrical surges in poles.
PRECIPITACION_APOYOS	Guidelines to mitigate the effects of rain and humidity on poles.
RADIACION_APOYOS	Norms on solar radiation protection for pole materials.
TEMPERATURA_APOYOS	Permissible temperature ranges to ensure safe pole operation.
KVA_TRAFOS	Standards on nominal capacity requirements for distribution transformers.
KV1_TRAFOS	Specifications for permissible primary voltage levels in transformers.
IMPEDANCIA_TRAFOS	Normative limits for transformer impedance values.
TEMPERATURA_TRANSFORMADOR	Standards on operational temperature ranges for transformers.
HUMEDAD_TRANSFORMADORES	Guidelines for acceptable humidity levels within transformers.
ACEITE_TRANSFORMADORES	Standards on dielectric oil quality and required characteristics.
MATERIAL_REDMT	Permissible materials for medium-voltage network construction.
GUARDACONDUCTOR_REDMT	Specifications for using guard wires in overhead networks.
VIENTO_REDMT	Technical requirements for network wind resistance.
TEMPERATURA_REDMT	Acceptable operational temperature range for medium-voltage network components.
PRECIPITACION_REDMT	Standards for protecting the network from rain and humidity.
KVNOM_REDMT	Limits on nominal voltage levels allowed in the network.
KV_SWITCHES	Standards for nominal voltage ratings in switches.
PHASES_SWITCHES	Norms defining the required number of phases in medium-voltage switches.
STATE_SWITCHES	Operational state requirements and compliance rules for switches.
VIENTO_SWITCHES	Resistance standards for switches under wind conditions.
TEMPERATURA_SWITCHES	Permissible temperature range for safe switch operation.
PRECIPITACION_SWITCHES	Guidelines for protection against rain and humidity.

Document	Description
HUMEDAD_SWITCHES	Specifications ensuring safe operation under allowed humidity levels.
RAYOS_SWITCHES	Requirements for switch protection against lightning and power surges.

### Expert-Guided Question-Answering Bitácoras for Model Validation

To evaluate the practical value and explainability of the recommendation system, we constructed a set of *Expert-Guided Q&A Bitácoras*—domain-grounded question-answer logs curated in collaboration with electrical engineers at CHEC. These bitácoras serve as the primary validation benchmark to assess the system’s capacity to retrieve accurate information from heterogeneous sources and generate actionable recommendations.

Each question in the bitácoras was designed to challenge a specific reasoning capability of the system, either related to structured data querying, unstructured regulation interpretation, or multi-source recommendation generation. For every question, two sets of answers were recorded:

- **Target Answer:** A reference response manually written and validated by a domain expert at CHEC.
- **Model Answers:** Independent responses automatically generated by each LLM under evaluation, given the same input query and contextual settings.

This parallel evaluation structure enabled a comparative assessment of factual consistency, normative alignment, completeness, and explainability of the generated answers. The questions were grouped into three types, summarized in Table 4-8.

Table 4-8: Example questions by Q&A bitácora category

Bitácora Type	Example Questions
<b>Structured Data Queries</b>	What was the cause of the outage on transformer T-0254 on April 3, 2023? Which zone had the highest number of switch interruptions in the last 12 months? List all medium-voltage circuits with over 10 events exceeding 30 minutes.
<b>Unstructured Regulation Interpretation</b>	What is the maximum operating temperature allowed for medium-voltage transformers? Does RETIE require mandatory grounding for all pole-mounted switches? Which RETIE sections regulate the wind load resistance of overhead conductors?
<b>Recommendations Based on Critical Variables</b>	What operational measures should be taken for transformer T-0875 given its recent temperature and oil deterioration readings? Recommend preventive actions for pole P-4221 based on its exposure to precipitation and lack of grounding. Suggest top 3 switch replacements based on failure risk indicators and regulatory non-compliance.

## LLM Configuration and Evaluation Protocol

To evaluate the adaptability of different Large Language Models (LLMs) to the hybrid retrieval and recommendation tasks defined in this work, we conducted a comparative benchmarking across ten state-of-the-art models from OpenAI, Google, Meta, Alibaba, and DeepSeek. These models were selected to represent diverse families of architectures, pretraining sizes, and context handling capabilities.

Each model was executed under identical prompt settings, using the same set of expert-guided questions introduced in Section 4-8, with full access to the same structured and unstructured knowledge base. Their configurations are summarized in Table 4-9.

Table 4-9: Configuration of evaluated LLMs

LLM	Number of Parameters	Context Length	Max Output Tokens	Quantization
gpt-3.5-turbo	Not disclosed	16,385	16,385	Not disclosed
gpt-4o	Not disclosed	128,000	128,000	Not disclosed
gemini-2.0	40B	1,048,576	8,192	Not disclosed
gemini-2.5	Not disclosed	1–2M	65,535	Not disclosed
llama-3.1-8b	8B	128,000	Not specified	4 bits
llama-3.2-1b	1B	128,000	8,000	4 bits
qwen-2.5-1.5b	1.5B	32,768	8,192	8 bits
qwen-2.5-7b	7B	131,072	8,000	16 bits
deepseek-r1-7b	7B	128,000	32,768	4 bits
deepseek-r1-1.5b	1.5B	128,000	32,768	4 bits

## Intermediate Mapping Tables for Agent Reasoning

To guide the decision-making process of the Agentic RAG pipeline, we constructed a set of intermediate mapping tables that define the critical variables, associated normative requirements, reference documents, and technical recommendations for each major network component type. These tables serve as reasoning anchors for the agent, enabling it to contextualize each user query (structured or unstructured) and retrieve the most appropriate regulatory fragment while also suggesting feasible actions.

**Table 4-10:** Table of Variables for Switches

Variable	Descripción	Normativa	Documento	Sugerencia
KV	Nivel de tensión nominal (kV).	- RETIE Artículo 3.17.17 - NTC 2050 - NTC 3285 - Normativa CHEC - IEC 62271-1 - IEC 62271-100	KV_SWITCHES	Verificar compatibilidad con la capacidad nominal del sistema eléctrico.
PHASES	Número de fases.	- RETIE (Artículo 3.17.17) - NTC 2050 (Sección 230-205) - NTC 2133 - IEC 62271 - normativa de CHEC - IEEE C37.20	PHASES_SWITCHES	Confirmar que se ajusta al diseño del sistema de distribución.
STATE	Estado operativo del switch (abierto/cerrado).	- RETIE (Artículo 3.17.17) - NTC 2050 (Sección 230-205) - IEC 62271-103 - normativa CHEC - IEEE C37.20	STATE_SWITCHES	Garantizar que el estado corresponde al diseño del sistema y las condiciones de operación.
VELOCIDAD_VIENTO	Velocidad del viento en la región.	- RETIE (Artículo 3.17.18) - NTC 2076 - IEC 62271-1 - directrices de CHEC	VIENTO_SWITCHES	Validar que los switches soporten la carga eólica máxima en condiciones extremas.
TEMPERATURA_AMBIENTE	Variación de temperatura ambiente.	- RETIE Artículo 3.17.17 - IEC 62271-1 - NTC 2050 - CHEC - Adaptaciones Locales - ASTM G155	TEMPERATURA_SWITCHES	Verificar el rango de operación térmica del switch y evitar fallas en altas o bajas temperaturas.
PRECIPITACIÓN	Intensidad y frecuencia de lluvias.	- RETIE Artículo 3.17.17 - IEC 60529 - NTC 2050 - CHEC	precipitacion_SWITCHES	Garantizar el grado de protección IP adecuado frente a ingreso de agua en condiciones adversas.
HUMEDAD_RELATIVA	Nivel de humedad en el entorno.	- RETIE Artículo 3.17.17 - IEC 62271-1 - NTC 2050 - CHEC - IEC 60068-2-78 - IEC 62208 - ISO 9227	humedad_SWITCHES	Evaluar el riesgo de condensación y su impacto en la durabilidad y el aislamiento del switch.

**Table 4-11:** Table of Variables for Network Sections

Variable	Descripción	Normativa	Documento	Sugerencia
KV NOM	Nivel de tensión nominal (kV).	- RETIE Artículo 3.20.6 - RETIE Artículo 3.20.5.1 - NTC 1340 - IEC 60038 - CHEC Calibres de Conductores - RETIE Artículo 3.20.6.4 - IEEE 835 - IEEE C2 - IEC 60909	KV NOM_REDMT	Validar compatibilidad con sistemas adyacentes y capacidad del conductor.
MATERIAL CONDUCTOR	Material del conductor.	- RETIE Artículo 3.20.6.1 - NTC 1329 - ASTM G85 - IEC 60468 - NTC 2244	material_REDMT	Asegurar resistencia y durabilidad según el entorno climático.
CALIBRE CONDUCTOR	Tamaño del conductor.	- RETIE Artículo 3.20.6.1 - NTC 1056 - CHEC Operador de Red - IEEE 738	CALIBRE CONDUCTOR_REDMT	Confirmar que soporta las demandas de corriente del sistema eléctrico.
GUARDA CONDUCTOR	Presencia de cable de guarda.	- RETIE Artículo 3.20.6.4 - NTC 2050 Capítulo 300.50 - IEEE 80 - IEC 60826	GUARDA CONDUCTOR_REDMT	Evaluar el diseño para mitigar riesgos de sobretensiones y rayos.
VELOCIDAD_VIENTO	Velocidad del viento en la región.	- RETIE Artículo 3.20.6.1 - CHEC 6.6 - IEC 60826 - NTC 2050	viento_REDMT	Asegurar que el diseño soporta la carga eólica máxima.
TEMPERATURA	Variación de temperatura ambiente.	- RETIE Artículo 3.20.6.1 - IEC 60287 - NTC 2050 - CHEC 6.5.2	temperatura_REDMT	Validar que el material soporta condiciones extremas.
HUMEDAD_RELATIVA	Nivel de humedad en el entorno.	- RETIE Artículo 3.20.6.3 - IEC 60502-2 - NTC 2050 - IEEE 1216	HUMEDAD_REDMT	Evaluar el impacto en la corrosión de conductores.
PRECIPITACIÓN	Intensidad y frecuencia de lluvias.	- RETIE Artículo 3.20.6.3 - IEC 60826 - CHEC - IEC 60507	precipitacion_REDMT	Analizar el riesgo de descargas y cortocircuitos por acumulación de agua.

**Table 4-12:** Table of Variables for Transformers

Variable	Descripción	Normativa	Documento	Sugerencia
KVA	Capacidad nominal del transformador.	- RETIE Artículo 3.17.28 - NTC 2050 - IEEE C57.12 - Normativa de CHEC	KVA_TRAFOS	Validar que el transformador cumpla con las demandas de carga del sistema eléctrico.
KV1	Tensión primaria nominal (kV).	- RETIE Artículo 3.17.28 - NTC 2050 - IEEE C57.12 - IEEE C57.91 - Normativa de CHEC	KV1_TRAFOS	Verificar compatibilidad con los niveles de tensión del sistema eléctrico.
IMPEDANCE	Impedancia del transformador (%).	- RETIE Artículo 3.17.28 - NTC 2050 - IEEE C57.12 - IEC 60076 - IEEE C57.152	IMPEDANCIA_TRAFOS	Analizar la contribución a las pérdidas y cortocircuitos.
TEMPERATURA	Rango de temperaturas máximas y mínimas.	- RETIE Artículo 3.17.28 - NTC 2050 - IEEE C57.12 - IEC 60076	temperatura_TRANSFORMADORES	Asegurar que el diseño del transformador soporte el rango de temperaturas especificado en la normativa.
HUMEDAD	Nivel promedio de humedad en la ubicación.	- RETIE Artículo 3.17.28 - NTC 2050 - IEEE C57.12 - IEC 60076	humedad_TRANSFORMADORES	Verificar materiales y protecciones frente a la corrosión y degradación.
ACEITE AISLANTE	Propiedades del aceite aislante.	- RETIE Artículo 3.17.28 - NTC 3436 - IEEE C57.104 - IEC 60296 - ASTM D974	aceite_TRANSFORMADORES	Realizar análisis periódicos y sustituir o reacondicionar si el aceite muestra degradación.

**Table 4-13:** Table of Variables for Utility Poles

Variable	Descripción	Normativa	Documento	Sugerencia
MATERIAL	Tipo de material del apoyo (concreto, madera, metal, fibra de vidrio, etc.).	- NTC 1065 (Postes Metálicos) - NTC 1329 (Postes de Concreto) - ASTM A123, A36, D4923 - IEEE 837 - CHEC: Líneas a 33 kV y 13.2 kV	MATERIAL_APOYOS_CARGA_TRABAJO_APOYOS	Verificar que cumple con resistencia, recubrimientos y tratamientos requeridos según normativa.
LONG_APOYO	Longitud total del apoyo en metros.	- Artículo 3.20.4 del RETIE - NTC 2050 - NTC 1329	APOYOS_LONG_APOYO	Validar que cumple con la longitud mínima requerida para la instalación en su categoría.
TIERRA_PIE	Resistencia del sistema de puesta a tierra (en $\Omega$ ).	- RETIE - IEEE Std 80 - NTC 2050	NORMATIVA_APOYOS_TIERRA_PIE	Verificar que la resistencia no exceda los valores permitidos.
VIENTO	Velocidad del viento en la zona donde está instalado el apoyo (en m/s o km/h).	- RETIE - NTC 1329 - ASCE 7	VIENTO_APOYOS	Validar que el diseño soporta la velocidad máxima registrada en la región.
ÍNDICE_RAYOS	Frecuencia promedio de descargas eléctricas en la región (rayos por km <sup>2</sup> por año).	- RETIE Artículo 3.20.5 - IEEE Std 998, 1410 - NTC 2050	RAYOS_APOYOS	Evaluar la necesidad de sistemas de protección contra rayos.
PRECIPITACIÓN	Media anual de precipitación en la zona (mm/año).	- RETIE Artículo 3.20.4 - NTC 1065, 1329 - CHEC Normativa	PRECIPITACION_APOYOS	Diseñar protecciones contra humedad y reforzar materiales expuestos.
RADIACIÓN_UV	Nivel de exposición a radiación ultravioleta ( $W/m^2$ ).	- ASTM G154 - NTC 6275 - CHEC Normativa	RADIACION_APOYOS	Verificar que los materiales sean resistentes a la degradación por UV.
TEMPERATURA	Rango de temperaturas operativas (-40°C a +50°C).	- RETIE 3.20.4 - ASTM D648 - NTC 2050	TEMPERATURA_APOYOS	Seleccionar materiales que soporten las condiciones térmicas extremas.

### 4.2.3 Confidentiality Statement on the Internal Database Use

The private database provided by CHEC, which contains historical records of medium-voltage power grid interruptions, was used strictly for informational validation purposes within this study. This database enabled the evaluation of the predictive and explanatory capabilities of the proposed methodology in a real operational context. For the purposes of this document, the database will be referenced solely as an internal validation point supporting the system’s results. No part of this database will be published or disclosed, and its use fully complies with established confidentiality agreements, thereby ensuring the protection of the company’s private and proprietary information.

## 4.3 Mathematical Background

In this section, we summarize the mathematical foundations underlying the proposed methodology. We progress from general to specific, covering machine learning models for tabular data (TabNet and other regressors), their loss functions and performance metrics, followed by large language models (LLMs), transformer architectures, pre-training strategies, evaluation metrics, and retrieval-augmented generation (RAG).

### 4.3.1 Classical Regressors

Classical regression models provide a solid baseline for supervised learning tasks involving tabular data. They differ in complexity, interpretability, and their ability to capture non-linear dependencies. In this subsection, we describe four representative regressors: *Linear Regression*, *Random Forest Regressor*, *Extreme Gradient Boosting (XGBoost)*, and *Deep Neural Networks (DNNs)* for regression.

**Linear Regression** As a foundational benchmark for regression, Ordinary Least Squares (OLS) presupposes a linear correlation between the exogenous variables, represented by an input matrix  $\mathbf{X} \in \mathbb{R}^{N \times P}$  (with  $N$  samples and  $P$  features), and a continuous target variable  $\mathbf{y} \in \mathbb{R}^{N \times 1}$ . A linear relationship is then established through the coefficient vector  $\boldsymbol{\theta} \in \mathbb{R}^{P \times 1}$ , as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad (4-2)$$

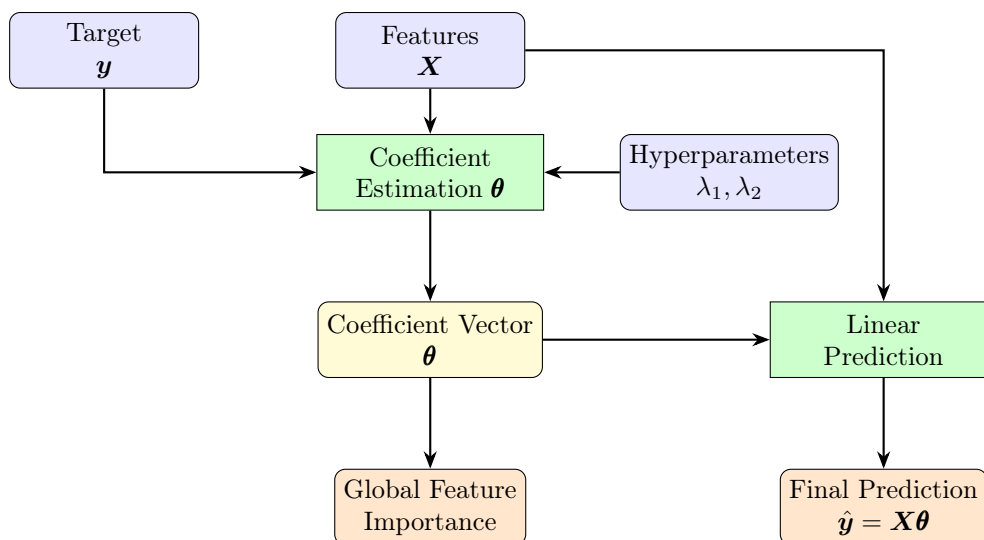
where  $\boldsymbol{\epsilon}$  denotes the residual noise. A common estimator of  $\boldsymbol{\theta}$  is obtained through the Moore–Penrose pseudoinverse:

$$\boldsymbol{\theta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}. \quad (4-3)$$

To mitigate overfitting or multicollinearity, regularized extensions are often used. A general penalized optimization can be written as:

$$\boldsymbol{\theta}^* = \arg \min_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2, \quad (4-4)$$

where  $\lambda_1, \lambda_2 \geq 0$  are regularization coefficients. When  $\lambda_1 > 0$  and  $\lambda_2 = 0$ , this reduces to LASSO regression [Ranstam & Cook, 2018]; when both terms are nonzero, it becomes Elastic Net regression [Murphy, 2022]. The overall pipeline is shown in Fig. 4-1.



**Figure 4-1:** Schematic representation of a linear modeling workflow: data input, coefficient estimation, and feature relevance extraction.

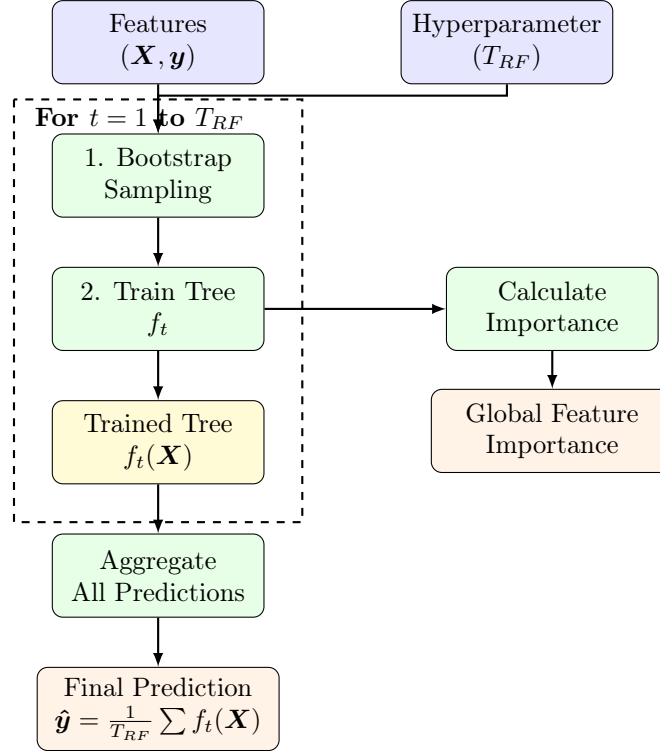
Despite its simplicity and interpretability, linear regression is limited to capturing linear patterns and may underperform when strong non-linearities exist in the data.

**Random Forest Regressor** Random Forests (RF) offer a non-linear alternative through ensemble learning. They construct multiple decision trees trained on bootstrap samples and aggregate their predictions:

$$\hat{\mathbf{y}} = \frac{1}{T_{RF}} \sum_{t=1}^{T_{RF}} f_t(\mathbf{X}), \quad (4-5)$$

where each tree  $f_t$  is trained on a random subset of features and samples. This ensemble approach reduces variance and improves generalization [Kumar et al., 2024]. Each tree learns piecewise-constant approximations by recursive partitioning to minimize node impurity, and global feature importance is derived

by aggregating reductions in squared error across all trees. The conceptual workflow of this process is illustrated in Fig. 4-2, which summarizes the bagging-based training loop, aggregation step, and importance computation.



**Figure 4-2:** Conceptual pipeline for Random Forest regression: bagging, tree training, ensemble averaging, and importance derivation.

*Extreme Gradient Boosting (XGBoost)* XGBoost extends ensemble learning by constructing trees sequentially, where each new tree corrects the residuals of the ensemble built so far [Uyar & Albayrak, 2025]. The prediction after  $T_{XGB}$  boosting rounds is:

$$\hat{\mathbf{y}}^{(T_{XGB})} = \sum_{t=1}^{T_{XGB}} \eta f_t(\mathbf{X}), \quad (4-6)$$

with  $\eta \in (0, 1]$  the learning rate. At each iteration, the model minimizes a regularized objective function:

$$\mathcal{J}^{(t)} \approx \sum_{n=1}^N [g_n f_t(\mathbf{x}_n) + \frac{1}{2} h_n f_t^2(\mathbf{x}_n)] + \Omega(f_t), \quad (4-7)$$

where  $g_n$  and  $h_n$  are the first and second derivatives of the loss with respect to current predictions, and  $\Omega(f_t) = \gamma L_t + \frac{\lambda}{2} \|\mathbf{w}^{(t)}\|_2^2$  is the regularization term controlling tree complexity. Fig. summarizes this process.

*Deep Neural Networks (DNNs) for Regression* Deep Neural Networks (DNNs) generalize linear models by introducing multiple layers of non-linear transformations, enabling them to capture complex dependencies among features. For an input  $\mathbf{x}$ , the  $l$ -th hidden layer computes:

$$\mathbf{h}^{(l)} = \sigma(W^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}), \quad (4-8)$$

where  $W^{(l)}$  and  $\mathbf{b}^{(l)}$  denote the weights and biases, and  $\sigma(\cdot)$  is a non-linear activation (e.g., ReLU). The final output layer produces:

$$\hat{y} = W^{(L+1)}\mathbf{h}^{(L)} + b^{(L+1)}. \quad (4-9)$$

The model is trained by minimizing a regression loss such as Mean Squared Error (MSE) or Mean Absolute Error (MAE):

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|. \quad (4-10)$$

While DNNs can approximate arbitrary functions and achieve strong performance in high-dimensional data, they require large datasets and careful regularization. Their interpretability is often limited compared to tree-based or linear models.

*Discussion* Linear models provide interpretability through explicit coefficients, Random Forests and XGBoost offer high accuracy and feature importance metrics derived from tree structures, and DNNs extend flexibility for complex, high-dimensional data. In subsequent sections, these regressors serve as performance and interpretability baselines against the proposed TabNet-based approach.

### 4.3.2 TabNet Architecture

Having established the performance and interpretability baselines with classical regressors, we now transition toward deep learning models specifically tailored for tabular data. Among these, **TabNet** [Arik & Pfister, 2021] stands out as a neural architecture that reconciles predictive power with intrinsic interpretability through sequential attention and sparse feature selection mechanisms.

Formally, let  $\mathbf{X} \in \mathbb{R}^{N \times P}$  denote the input data,  $\mathbf{y} \in \mathbb{R}^{N \times 1}$  the target vector, and  $\hat{\mathbf{y}} \in \mathbb{R}^{N \times 1}$  the prediction produced by a parametric mapping:

$$\hat{\mathbf{y}} = f(\mathbf{X}; \Theta) = (f_L \circ \check{f}_{L-1} \circ \cdots \circ \check{f}_1)(\mathbf{X}),$$

where  $\Theta$  represents the set of trainable parameters. TabNet realizes this mapping as a composition of *decision steps* that progressively refine representations through attention-based feature selection. Each step identifies a sparse subset of relevant features and transforms them into decision embeddings, leading to both high predictive performance and transparent feature attribution.

*Core Components* TabNet comprises three core components:

- **Attentive Feature Selection:** implemented through the Sparsemax activation, which enforces sparsity in the attention masks and allows the model to focus on a limited subset of features.
- **Sequential Decision Steps:** each decision step selectively transforms the chosen features, refining the representation over time.
- **Interpretability Mechanism:** accumulated attention masks yield both local and global feature importance measures, providing human-understandable explanations of the model’s reasoning.

*Sequential Attention and Sparsemax* At each step  $s$ , TabNet computes an attention mask  $\mathbf{Z}^{(s)} \in \mathbb{R}^{N \times P}$  that performs soft feature selection:

$$\mathbf{Z}^{(s)} = \text{sparsemax}(\mathbf{Q}^{(s-1)} \cdot \phi_s(\mathbf{c}^{(s-1)})), \quad (4-11)$$

where  $\mathbf{Q}^{(s-1)}$  tracks the prior scale of feature usage, and  $\phi_s(\cdot)$  is a trainable transformation of the previous feature context  $\mathbf{c}^{(s-1)}$ . The *sparsemax* activation [Martins & Astudillo, 2016] projects logits onto the probability simplex while promoting sparsity:

$$\text{sparsemax}(\mathbf{z}) = \underset{\mathbf{p} \in \Delta^{d-1}}{\arg \min} \|\mathbf{p} - \mathbf{z}\|^2, \quad \text{with } \Delta^{d-1} = \{\mathbf{p} \in \mathbb{R}^d \mid \mathbf{p} \geq 0, \sum_j p_j = 1\}. \quad (4-12)$$

The prior scale is updated recursively to prevent overuse of features:

$$\mathbf{Q}^{(s)} = \mathbf{Q}^{(s-1)} \odot (\nu - \mathbf{Z}^{(s)}), \quad (4-13)$$

where  $\nu \geq 1$  controls feature reuse. This encourages exploration of new features at each decision step.

*Feature Transformer (GLU Block)* The masked features  $\mathbf{F}^{(s)} = \mathbf{Z}^{(s)} \odot \mathbf{X}$  are processed by a feature transformer  $\mathcal{F}$ , which extracts nonlinear representations through Gated Linear Units (GLUs) [Dauphin et al., 2017]:

$$\text{GLU}(\mathbf{h}') = (\mathbf{W}_1 \mathbf{h}' + \mathbf{b}_1) \odot \sigma(\mathbf{W}_2 \mathbf{h}' + \mathbf{b}_2), \quad (4-14)$$

where  $\sigma$  is the sigmoid activation,  $\mathbf{W}_1, \mathbf{W}_2$  are weight matrices, and  $\mathbf{b}_1, \mathbf{b}_2$  are biases. Residual connections are normalized by  $\sqrt{0.5}$  to stabilize learning. The transformer outputs two vectors: a decision embedding  $\mathbf{d}^{(s)} \in \mathbb{R}^{N \times N_d}$  and an attention embedding  $\mathbf{c}^{(s)} \in \mathbb{R}^{N \times N_a}$ :

$$[\mathbf{d}^{(s)}, \mathbf{c}^{(s)}] = \mathcal{F}(\mathbf{F}^{(s)}), \quad (4-15)$$

where  $N_d$  and  $N_a$  are the dimensions of the decision and attention embeddings, respectively.

*Normalization and Aggregation* To handle large batch sizes, TabNet applies **ghost batch normalization**, splitting each batch into virtual mini-batches of size  $B_v$ :

$$\tilde{\mathbf{X}} = \frac{\mathbf{X} - \boldsymbol{\mu}_{B_v}}{\sqrt{\boldsymbol{\sigma}_{B_v}^2 + \epsilon}}, \quad (4-16)$$

ensuring stable normalization without excessive memory cost. The overall decision embedding aggregates contributions from all steps:

$$\hat{\mathbf{y}} = \left( \sum_{s=1}^S \text{ReLU}(\mathbf{d}^{(s)}) \right) \mathbf{W}_{\text{final}}, \quad (4-17)$$

where  $\mathbf{W}_{\text{final}}$  projects the aggregated embedding onto the regression target.

*Loss Function and Sparsity Regularization* The total objective function combines task-specific loss and a sparsity penalty:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{sparse}} \mathcal{L}_{\text{sparse}}, \quad (4-18)$$

with  $\lambda_{\text{sparse}}$  controlling the regularization strength. The regression loss is the Mean Squared Error:

$$\mathcal{L}_{\text{task}} = \frac{1}{N} \sum_{n=1}^N (y_n - \hat{y}_n)^2, \quad (4-19)$$

and the sparsity loss promotes compact feature selection:

$$\mathcal{L}_{\text{sparse}} = \sum_{s=1}^S \frac{-1}{N} \sum_{n=1}^N \sum_{p=1}^P \mathbf{Z}_{n,p}^{(s)} \log(\mathbf{Z}_{n,p}^{(s)} + \epsilon). \quad (4-20)$$

*Pretraining and Fine-Tuning Strategy* To enhance representation learning, TabNet was first configured as an **autoencoder**, where the encoder compressed the outage-related features into a latent representation  $Z \in \mathbb{R}^{n \times k}$ :

$$Z = \text{Encoder}_{\text{TabNet}}(\mathbf{X}), \quad (4-21)$$

and the decoder reconstructed the input, optimizing Mean Squared Error during pretraining. The pretrained encoder was then fine-tuned as a **regressor** to predict the System Average Interruption Duration Index (SAIDI):

$$\hat{y} = W \cdot Z + b, \quad (4-22)$$

minimizing the Mean Absolute Error:

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|. \quad (4-23)$$

*Hyperparameter Optimization* A Bayesian optimization via Optuna explored key hyperparameters: attention dimension  $n_a$ , decision dimension  $n_d$  (2–512), number of decision steps  $n_{\text{steps}}$  (3–10), and embedding sizes (3–70). Regularization terms such as  $\nu$  (feature reuse) and  $\lambda_{\text{sparse}}$  were tuned jointly with learning rates ( $10^{-7}$ – $10^{-1}$ ) and optimizers (Adam, AdamW, SGD, RMSprop). Batch sizes between 1024 and 4096, and virtual batches of 256–1024, were evaluated. Early stopping with patience of 40 epochs was applied, with training capped at 100 epochs. The validation metric was MAE.

*Interpretability and Feature Importance* TabNet’s interpretability stems from aggregating feature masks across decision steps. The importance score for feature  $p$  and sample  $n$  is:

$$\bar{Z}_{n,p} = \frac{\sum_{s=1}^S \xi_n^{(s)} \mathbf{Z}_{n,p}^{(s)}}{\sum_{p'=1}^P \sum_{s=1}^S \xi_n^{(s)} \mathbf{Z}_{n,p'}^{(s)}}, \quad (4-24)$$

where  $\xi_n^{(s)} = \sum_{k=1}^{N_d} \text{ReLU}(\mathbf{d}_{n,k}^{(s)})$  quantifies the contribution of step  $s$  to the prediction. Aggregating these masks yields global importance maps that reveal which electrical network components (transformers, switches, supports) most influence SAIDI. The top-20 features were visualized through bar plots, violin plots, and correlation heatmaps.

---

*Illustrative Diagram* Figure **4-3** depicts the full TabNet pipeline, highlighting the sequence of attentive masks, feature transformers, residual aggregations, and the predictor head.

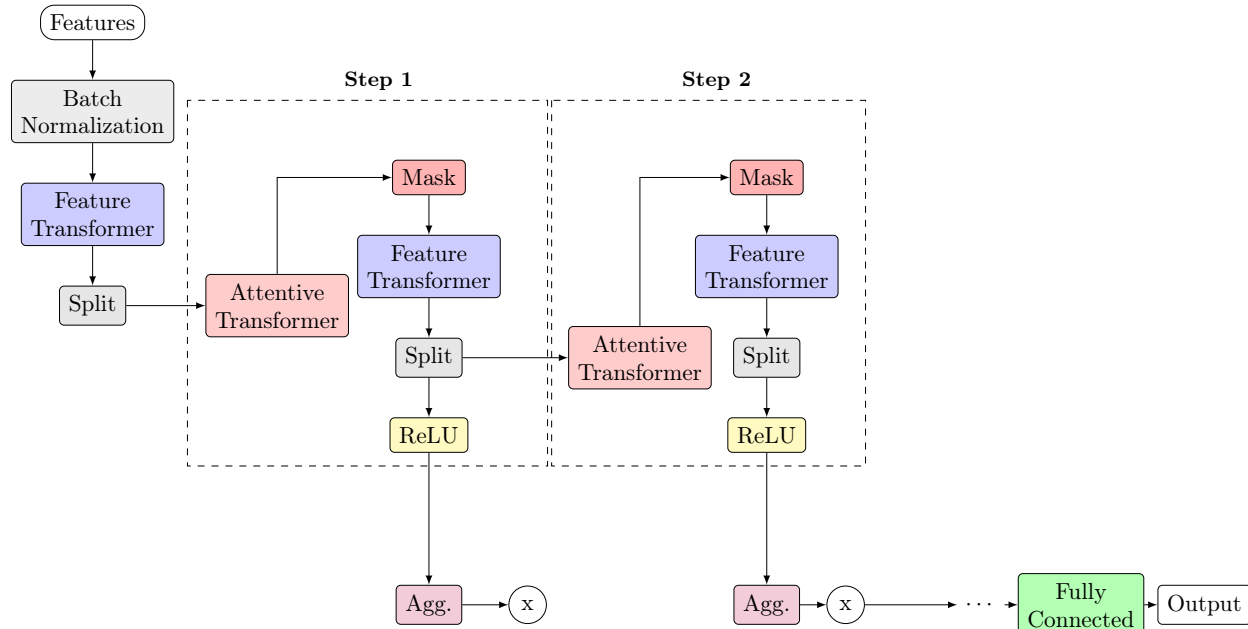


Figure 4-3: TabNet step-wise architecture: each decision step applies attentive feature selection, transformation, and residual aggregation before producing the final prediction.

**Summary** In summary, TabNet bridges the gap between traditional regressors and deep architectures by introducing sparsity-driven attention for interpretability, efficient feature reuse via prior scaling, and GLU-based transformations for expressive modeling. Its dual-stage use as an autoencoder and regressor enabled both robust representation learning and transparent prediction of SAIDI, positioning it as the central model in this study.

### 4.3.3 Loss Functions for Regression

This section builds upon standard formulations of regression losses and regularization techniques commonly described in the machine learning literature [Jadon, 2022, Bishop, 2006].

Loss functions are crucial in regression tasks since they define how prediction errors are quantified and guide the optimization process. Different loss functions emphasize distinct error properties, and the choice can significantly impact model performance and robustness. Below, we describe the most widely used loss functions for regression: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Huber Loss, along with common regularization terms (L1, L2, Elastic Net).

**Mean Squared Error (MSE)** The Mean Squared Error is the most common loss function in regression. It penalizes larger deviations more strongly due to the squared term. Given predictions  $\hat{y}_i$  and targets  $y_i$  for  $N$  samples, the MSE is defined as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2. \quad (4-25)$$

MSE is differentiable and convex, which makes it suitable for optimization using gradient-based methods. However, it is sensitive to outliers, as large errors disproportionately influence the overall loss.

*Mean Absolute Error (MAE)* The Mean Absolute Error measures the average magnitude of errors without considering their direction. It is defined as:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|. \quad (4-26)$$

Unlike MSE, MAE treats all deviations equally, making it more robust to outliers. However, the absolute value introduces non-differentiability at zero, which can complicate optimization. In practice, sub-gradient methods are used to overcome this issue.

*Huber Loss and Regularization Terms (L1, L2, Elastic Net)* The Huber Loss combines the benefits of MSE and MAE, being quadratic for small residuals and linear for large residuals. It is defined as:

$$\mathcal{L}_{\delta}(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |y - \hat{y}| \leq \delta, \\ \delta|y - \hat{y}| - \frac{1}{2}\delta^2, & \text{otherwise.} \end{cases} \quad (4-27)$$

Here,  $\delta > 0$  is a threshold that controls the transition between quadratic and linear regions. This loss is less sensitive to outliers than MSE while retaining differentiability, making it a popular choice in practice.

To prevent overfitting, regression models often incorporate **regularization terms**:

- **L1 Regularization (Lasso):**

$$\Omega_{L1} = \lambda \sum_{j=1}^d |w_j|, \quad (4-28)$$

which encourages sparsity in the weight vector  $\mathbf{w}$ .

- **L2 Regularization (Ridge):**

$$\Omega_{L2} = \lambda \sum_{j=1}^d w_j^2, \quad (4-29)$$

which penalizes large weights and improves generalization.

#### - Elastic Net:

combines both penalties:

$$\Omega_{EN} = \alpha \Omega_{L1} + (1 - \alpha) \Omega_{L2}, \quad (4-30)$$

where  $\alpha \in [0, 1]$  controls the trade-off. This formulation benefits from both sparsity (L1) and stability (L2), making it especially effective in high-dimensional data scenarios.

### 4.3.4 Performance Metrics for Regression

The metrics presented here are adapted from established references in regression model evaluation and performance analysis [*Chicco et al., 2021, Jadon, 2022*].

To evaluate the predictive capability of regression models, several performance metrics are used. These metrics quantify the difference between predicted values  $\hat{y}_i$  and true targets  $y_i$  across  $N$  samples. Below, we summarize the most common metrics for regression tasks.

**MAE, MSE, RMSE Mean Absolute Error (MAE):** measures the average magnitude of errors without considering their direction. It is robust to outliers compared to squared-error metrics.

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|. \quad (4-31)$$

**Mean Squared Error (MSE):** emphasizes larger deviations by squaring the error terms, making it more sensitive to outliers.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2. \quad (4-32)$$

**Root Mean Squared Error (RMSE):** is the square root of the MSE and expresses the error in the same units as the target variable.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2}. \quad (4-33)$$

While MSE and RMSE penalize large errors more heavily, MAE provides a more balanced view when outliers are present. RMSE is particularly useful when interpretability in the original scale of the target variable is desired.

**Coefficient of Determination ( $R^2$ )** The coefficient of determination,  $R^2$ , measures the proportion of variance in the target variable that is explained by the regression model. It is defined as:

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}, \quad (4-34)$$

where  $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$  is the mean of the observed values.

An  $R^2$  value close to 1 indicates that the model explains most of the variance in the data, while values near 0 suggest poor explanatory power. Negative values imply that the model performs worse than a simple mean-based predictor.

$R^2$  is widely used but should be interpreted with caution: it does not capture bias, overfitting, or whether predictions are systematically shifted.

### 4.3.5 Transformer Architecture

The following description is primarily based on the original Transformer architecture proposed by Vaswani *et al.* [Vaswani *et al.*, 2017], along with later architectural refinements for large-scale models [Dosovitskiy *et al.*, 2021].

The Transformer architecture has become the foundation of modern large language models (LLMs) by replacing recurrence with attention, thus enabling efficient parallelization, long-range dependency modeling, and scalability. Its design is modular and consists of stacked encoder and decoder blocks, where each block is composed of attention mechanisms, feed-forward networks, residual connections, and normalization layers.

**Self-Attention and Scaled Dot-Product** Given an input sequence represented as  $X \in \mathbb{R}^{n \times d_{\text{model}}}$ , the Transformer projects it into queries  $Q = XW^Q$ , keys  $K = XW^K$ , and values  $V = XW^V$ , where

$W^Q, W^K, W^V \in \mathbb{R}^{d_{\text{model}} \times d_k}$  are trainable parameters. The scaled dot-product attention is defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \quad (4-35)$$

The scaling factor  $\sqrt{d_k}$  prevents extremely large dot products, which would otherwise push the softmax function into saturation regions with vanishing gradients. This mechanism allows every token to attend to all other tokens, capturing global contextual dependencies.

*Positional Encoding* Since the self-attention mechanism is permutation-invariant, positional encodings are introduced to inject order information into the sequence. These encodings are added to the token embeddings at the input layer:

$$\tilde{X} = X + P, \quad (4-36)$$

where  $P \in \mathbb{R}^{n \times d_{\text{model}}}$  is a positional encoding matrix. In the original Transformer, sinusoidal encodings were defined as:

$$\text{PE}(t, 2i) = \sin\left(t/10000^{2i/d_{\text{model}}}\right), \quad (4-37)$$

$$\text{PE}(t, 2i + 1) = \cos\left(t/10000^{2i/d_{\text{model}}}\right), \quad (4-38)$$

where  $t$  denotes the position index and  $i$  the embedding dimension. Learned positional embeddings are also widely adopted in modern LLMs.

*Multi-Head Attention* Instead of computing attention in a single space, the Transformer uses  $h$  attention heads, each operating in a different subspace:

$$\text{head}_j = \text{Attention}(QW_j^Q, KW_j^K, VW_j^V), \quad j = 1, \dots, h, \quad (4-39)$$

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O. \quad (4-40)$$

This allows the model to jointly capture diverse relationships, such as syntactic and semantic dependencies, across multiple representational subspaces.

*Residual Connections and Layer Normalization* Each sub-layer within the Transformer is wrapped with residual connections followed by layer normalization. For an input  $X$ , the pre-norm variant, widely used for stability in deep training, computes:

$$Y = X + \text{Dropout}(\text{MHA}(\text{LN}(X))), \quad (4-41)$$

$$Z = Y + \text{Dropout}(\text{FFN}(\text{LN}(Y))). \quad (4-42)$$

Residual connections mitigate vanishing gradients, while LayerNorm stabilizes training by re-centering and scaling activations.

*Feed-Forward Network (FFN)* Each encoder and decoder block contains a position-wise feed-forward network (FFN), applied independently to each token:

$$\text{FFN}(x) = W_2 \phi(W_1 x + b_1) + b_2, \quad (4-43)$$

where  $\phi$  is a non-linear activation function, typically ReLU or GELU. The hidden dimension  $d_{\text{ff}}$  is usually set to  $4d_{\text{model}}$ , providing additional representational capacity.

*Encoder, Decoder, and Encoder-Decoder Variants* The **encoder** consists of  $N$  stacked layers of multi-head self-attention and FFN modules, producing contextualized representations  $Z = \text{Encoder}(X)$ .

The **decoder** contains three sub-layers: (i) masked multi-head self-attention over the target sequence (to preserve autoregressive causality), (ii) encoder-decoder attention over  $Z$ , and (iii) a feed-forward network. Given partial target inputs  $Y$ , the decoder generates predictions:

$$\hat{Y} = \text{Decoder}(Y, Z). \quad (4-44)$$

Variants include encoder-only models (e.g., BERT) specialized in representation learning, decoder-only models (e.g., GPT) optimized for autoregressive generation, and encoder-decoder architectures (e.g., T5, BART) suited for sequence-to-sequence tasks.

*Attention Masking* Attention masking ensures correct sequence handling. Padding masks prevent attending to empty tokens, while causal masks enforce autoregressive constraints:

$$\text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V, \quad (4-45)$$

where

$$M_{ij} = \begin{cases} -\infty & \text{if position } j \text{ is padding or } j > i \text{ (causal mask),} \\ 0 & \text{otherwise.} \end{cases} \quad (4-46)$$

This ensures that the model cannot peek at future tokens and that padding tokens do not interfere with attention computations.

### 4.3.6 Training Objectives for LLMs

The training paradigms discussed below are derived from seminal works introducing masked and autoregressive objectives [Devlin et al., 2019, Radford et al., 2018], and from reinforcement learning frameworks aligning models with human feedback [Christiano et al., 2017].

Large Language Models (LLMs) are trained using different objectives depending on whether the goal is bidirectional understanding, autoregressive generation, or alignment with human preferences. Below, we summarize the most important training paradigms.

*Masked Language Modeling (MLM)* Masked Language Modeling is the training objective popularized by BERT and related models. Given an input sequence of tokens  $\{x_1, \dots, x_n\}$ , a subset  $\mathcal{M} \subset \{1, \dots, n\}$  is randomly replaced by a special token [MASK]. The model is then trained to recover the original tokens at those positions.

Formally, the objective is to minimize the negative log-likelihood of the masked tokens:

$$\mathcal{L}_{MLM} = - \sum_{i \in \mathcal{M}} \log P_{\theta}(x_i | x_{/i}), \quad (4-47)$$

where  $x_{/i}$  denotes the sequence with the masked token hidden. MLM enables bidirectional context learning but introduces a pretrain–finetune mismatch, since [MASK] tokens do not appear at inference.

*Next Sentence Prediction (NSP)* Next Sentence Prediction is a complementary task introduced in BERT to capture inter-sentence relationships. Given two segments  $A$  and  $B$ , the model must predict whether  $B$  is the actual next sentence following  $A$  in the corpus or a randomly sampled sentence.

The loss is defined as a binary cross-entropy classification:

$$\mathcal{L}_{NSP} = - \left[ y \log P_{\theta}(\text{IsNext} | A, B) + (1 - y) \log P_{\theta}(\text{NotNext} | A, B) \right], \quad (4-48)$$

where  $y = 1$  if  $B$  follows  $A$ , and  $y = 0$  otherwise. Later work (e.g., RoBERTa) showed that NSP is not always essential, but it remains an important early objective.

*Causal Language Modeling (CLM)* Causal Language Modeling, used in GPT-style models, trains the network autoregressively by predicting the next token given all previous ones. For a sequence  $\{x_1, \dots, x_n\}$ , the objective is:

$$\mathcal{L}_{CLM} = - \sum_{i=1}^n \log P_{\theta}(x_i | x_1, \dots, x_{i-1}). \quad (4-49)$$

This approach naturally aligns with left-to-right generation and avoids the pretrain–finetune mismatch of MLM. However, it lacks explicit bidirectional context, which is instead inferred from large-scale training.

*Reinforcement Learning with Human Feedback (RLHF)* Reinforcement Learning with Human Feedback has become central to aligning LLMs with human preferences. The process generally involves three steps: (i) pretraining a base LLM with MLM or CLM, (ii) training a reward model  $R_{\phi}(y | x)$  from human preference data, and (iii) fine-tuning the LLM with reinforcement learning, typically Proximal Policy Optimization (PPO).

The RLHF objective seeks to maximize expected reward:

$$\mathcal{L}_{RLHF} = -\mathbb{E}_{y \sim \pi_{\theta}(\cdot | x)} [R_{\phi}(y | x)], \quad (4-50)$$

where  $\pi_{\theta}$  is the LLM policy and  $R_{\phi}$  is the reward model. PPO constrains policy updates to remain close to the pretrained distribution while improving alignment.

RLHF enables models not only to generate coherent text but also to follow human instructions, reduce harmful outputs, and exhibit more helpful conversational behavior.

### 4.3.7 Loss Functions for LLMs

The loss formulations described here follow standard objectives used in large-scale language modeling and fine-tuning via reinforcement learning with human feedback [Schulman *et al.*, 2017, Ouyang *et al.*, 2022].

Training and fine-tuning Large Language Models (LLMs) require carefully designed loss functions that guide the optimization process. These functions determine how the model parameters are updated to improve predictive accuracy and alignment with human preferences.

*Cross-Entropy Loss* The primary loss function for language modeling tasks is the token-level cross-entropy loss. Given a training dataset consisting of token sequences  $\{x_1, \dots, x_n\}$ , the model is trained to minimize the negative log-likelihood of the true next token.

Formally, the loss is defined as:

$$\mathcal{L}_{CE} = -\frac{1}{n} \sum_{i=1}^n \log P_{\theta}(x_i | x_1, \dots, x_{i-1}), \quad (4-51)$$

where  $P_{\theta}(x_i | x_{<i})$  is the probability assigned by the model with parameters  $\theta$  to the true token  $x_i$ .

Cross-entropy directly penalizes incorrect predictions, encouraging the model to assign higher probabilities to the correct tokens. This loss function is used in both masked language modeling (MLM) and causal language modeling (CLM).

*Kullback–Leibler Divergence (for RLHF fine-tuning)* In Reinforcement Learning with Human Feedback (RLHF), fine-tuning requires balancing two objectives: maximizing rewards assigned by the human-trained reward model while preventing the updated policy from drifting too far from the pretrained distribution. This trade-off is formalized using the Kullback–Leibler (KL) divergence.

Given a pretrained policy  $\pi_{\theta_0}$  and an updated policy  $\pi_{\theta}$ , the KL divergence is defined as:

$$D_{KL}(\pi_{\theta} \parallel \pi_{\theta_0}) = \sum_y \pi_{\theta}(y | x) \log \frac{\pi_{\theta}(y | x)}{\pi_{\theta_0}(y | x)}. \quad (4-52)$$

In practice, the KL term is added as a regularization component to the reinforcement learning objective:

$$\mathcal{L}_{RLHF} = -\mathbb{E}_{y \sim \pi_{\theta}(\cdot | x)} [R_{\phi}(y | x)] + \beta D_{KL}(\pi_{\theta} \parallel \pi_{\theta_0}), \quad (4-53)$$

where  $R_{\phi}$  is the reward model, and  $\beta$  controls the strength of the penalty.

This ensures that fine-tuning improves alignment with human preferences while maintaining linguistic fluency and avoiding catastrophic deviations from the pretrained base model.

### 4.3.8 Performance Metrics for LLMs

The evaluation metrics summarized in this section are inspired by widely used approaches in natural language generation and text similarity assessment [Papineni et al., 2002, Zhang et al., 2020].

Evaluating Large Language Models (LLMs) requires multiple complementary metrics, as no single measure can fully capture fluency, coherence, and semantic accuracy. Below, we present the most commonly used metrics in both pretraining and downstream evaluation.

*Perplexity* Perplexity (PPL) measures how well a language model predicts a sequence of tokens. It is defined as the exponential of the average negative log-likelihood of the true tokens given the model’s predictions:

$$\text{PPL} = \exp\left(-\frac{1}{n} \sum_{i=1}^n \log P_{\theta}(x_i | x_{<i})\right), \quad (4-54)$$

where  $n$  is the sequence length, and  $P_{\theta}(x_i | x_{<i})$  is the probability assigned by the model to the token  $x_i$ .

Lower perplexity indicates better predictive performance, with values closer to 1 suggesting higher confidence in generating the correct sequence.

*BERTScore* BERTScore evaluates text generation quality by comparing candidate and reference sentences in a contextual embedding space obtained from pretrained transformers (e.g., BERT, RoBERTa).

For each token embedding  $\mathbf{h}_i$  in the candidate and  $\mathbf{h}'_j$  in the reference, cosine similarity is used to align tokens and compute precision, recall, and F1 scores:

$$\text{Precision} = \frac{1}{|C|} \sum_{i \in C} \max_{j \in R} \cos(\mathbf{h}_i, \mathbf{h}'_j), \quad (4-55)$$

$$\text{Recall} = \frac{1}{|R|} \sum_{j \in R} \max_{i \in C} \cos(\mathbf{h}_i, \mathbf{h}'_j), \quad (4-56)$$

where  $C$  and  $R$  denote the sets of tokens in the candidate and reference sentences, respectively.

The BERTScore F1 is then computed as the harmonic mean of precision and recall. Unlike BLEU or ROUGE, BERTScore captures semantic similarity rather than relying only on n-gram overlap.

*Cosine Similarity in Embedding Space* Cosine similarity is a widely used metric for evaluating semantic retrieval and embedding-based tasks. Given two embedding vectors  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ , the similarity is defined as:

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}. \quad (4-57)$$

Values range from  $-1$  (completely dissimilar) to  $1$  (identical direction in embedding space).

In retrieval-augmented generation (RAG) systems, cosine similarity is critical for ranking candidate documents by measuring their closeness to the query embedding. High similarity values indicate that the retrieved document is semantically aligned with the query.

### 4.3.9 Vector Embeddings and Semantic Spaces

The representation principles discussed here are grounded in foundational embedding models such as Word2Vec and GloVe [Mikolov *et al.*, 2013, Pennington *et al.*, 2014], and later contextual approaches like BERT [Devlin *et al.*, 2019].

Vector embeddings are numerical representations of tokens, words, sentences, or documents in a continuous vector space, typically  $\mathbb{R}^d$ . They are learned through neural networks to capture semantic and syntactic properties, such that semantically similar items lie closer together.

Formally, let  $\mathcal{E} : \mathcal{X} \rightarrow \mathbb{R}^d$  be an embedding function that maps an input text  $x \in \mathcal{X}$  to a dense vector  $\mathbf{e} \in \mathbb{R}^d$ . The embedding space allows operations such as clustering, semantic search, and retrieval by leveraging geometric relationships between vectors.

Embeddings are central to both classical NLP tasks (e.g., Word2Vec, GloVe) and modern transformer-based models (e.g., BERT, GPT), where contextual embeddings dynamically represent meaning depending on surrounding context.

### 4.3.10 Similarity Search Metrics

The similarity measures presented are based on established concepts in information retrieval and vector search [Manning *et al.*, 2008, Johnson *et al.*, 2019].

Similarity metrics in embedding space determine how close or related two vectors are, enabling semantic retrieval, clustering, and ranking. The three most widely used metrics are cosine similarity, dot product, and Euclidean distance.

**Cosine Similarity** Cosine similarity measures the cosine of the angle between two vectors, focusing on direction rather than magnitude. Given two embeddings  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ , the similarity is:

$$\cos(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|}, \quad (4-58)$$

where  $\mathbf{u} \cdot \mathbf{v}$  denotes the dot product and  $\|\cdot\|$  the Euclidean norm. Values range from  $-1$  (opposite direction) to  $1$  (identical direction). This metric is widely adopted in retrieval-augmented generation (RAG) systems and semantic search.

**Dot Product and Euclidean Distance** The **dot product** directly measures the unnormalized similarity between two vectors:

$$\mathbf{u} \cdot \mathbf{v} = \sum_{i=1}^d u_i v_i. \quad (4-59)$$

It is sensitive to both vector direction and magnitude, making it useful in attention mechanisms within transformers (e.g., scaled dot-product attention).

The **Euclidean distance** (or  $\ell_2$  distance) instead quantifies dissimilarity as the straight-line distance between vectors:

$$d(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2 = \sqrt{\sum_{i=1}^d (u_i - v_i)^2}. \quad (4-60)$$

Smaller distances indicate stronger similarity. While Euclidean distance is intuitive, it may be less effective in high-dimensional spaces due to the "curse of dimensionality," where distances tend to concentrate.

In practice, cosine similarity is often preferred for high-dimensional embeddings, whereas dot product is used internally in attention mechanisms, and Euclidean distance is applied in clustering tasks.

### 4.3.11 Retrieval Performance Metrics

The following evaluation criteria follow standard practices in information retrieval and ranking systems [Manning *et al.*, 2008, Guo *et al.*, 2019].

Evaluating retrieval systems requires metrics that capture how effectively relevant items are identified and

ranked. Unlike classical regression or classification metrics, retrieval metrics account for ordered candidate lists and relevance judgments. Below we present the most widely used: Recall@k, Mean Reciprocal Rank (MRR), and Precision/F1 adapted to retrieval contexts.

**Recall@k** Recall@k measures the proportion of queries for which at least one relevant document appears among the top- $k$  retrieved results. Formally, for a set of  $Q$  queries:

$$\text{Recall@k} = \frac{1}{Q} \sum_{q=1}^Q \mathbb{1}[\exists d \in D_q^{rel} \cap D_q^{(k)}], \quad (4-61)$$

where  $D_q^{rel}$  is the set of relevant documents for query  $q$ ,  $D_q^{(k)}$  is the top- $k$  retrieved documents, and  $\mathbb{1}[\cdot]$  is the indicator function. A higher Recall@k indicates better ability to surface at least one relevant item in the top results.

**Mean Reciprocal Rank (MRR)** MRR evaluates ranking quality by measuring the position of the first relevant document in the retrieved list. It is defined as:

$$\text{MRR} = \frac{1}{Q} \sum_{q=1}^Q \frac{1}{\text{rank}_q}, \quad (4-62)$$

where  $\text{rank}_q$  is the position of the first relevant document for query  $q$ . This metric heavily rewards systems that rank relevant documents at the very top of the list.

**Precision and F1 in Retrieval Contexts** **Precision@k** measures the fraction of retrieved documents in the top- $k$  that are actually relevant:

$$\text{Precision@k} = \frac{|D_q^{rel} \cap D_q^{(k)}|}{k}. \quad (4-63)$$

Precision is useful when the goal is to minimize irrelevant results shown to users.

The **F1-score** balances precision and recall in retrieval, defined as the harmonic mean:

$$F1@k = 2 \cdot \frac{\text{Precision@k} \cdot \text{Recall@k}}{\text{Precision@k} + \text{Recall@k}}. \quad (4-64)$$

F1 is particularly informative when systems must not only retrieve relevant documents (recall) but also avoid noise (precision).

These metrics together provide a comprehensive view of retrieval quality: Recall@k for coverage, MRR for ranking effectiveness, and Precision/F1 for balance between accuracy and completeness.

### 4.3.12 Combined Loss in RAG

The following evaluation criteria follow standard practices in information retrieval and ranking systems [Manning *et al.*, 2008, Guo *et al.*, 2019].

Retrieval-Augmented Generation (RAG) integrates both retrieval and generation modules. Training such systems requires a combined objective function that aligns the retriever with the generator, while ensuring that both modules contribute to the final task performance. This is achieved by combining retrieval loss, generation loss, and multi-task optimization strategies.

**Retrieval Loss** The retrieval component is trained to maximize the probability of selecting relevant documents given a query. Typically, retrieval loss is formulated as a contrastive loss over positive (relevant) and negative (irrelevant) documents:

$$\mathcal{L}_{\text{retrieval}} = - \sum_{q \in Q} \log \frac{\exp(\text{sim}(q, d^+))}{\exp(\text{sim}(q, d^+) + \sum_{d^-} \exp(\text{sim}(q, d^-))}, \quad (4-65)$$

where  $\text{sim}(q, d)$  is a similarity function (e.g., dot product or cosine similarity) between query  $q$  and document  $d$ ,  $d^+$  is the positive document, and  $d^-$  are sampled negatives. This loss encourages the retriever to score relevant documents higher than irrelevant ones.

**Generation Loss (Cross-Entropy)** The generation module, usually based on a Transformer decoder, is trained with a cross-entropy loss to maximize the likelihood of the target sequence given the query and retrieved documents:

$$\mathcal{L}_{\text{generation}} = - \sum_{t=1}^T \log P(y_t | y_{<t}, q, D^{(k)}; \theta), \quad (4-66)$$

where  $y_t$  is the  $t$ -th token of the target sequence,  $y_{<t}$  denotes previously generated tokens,  $q$  is the query,  $D^{(k)}$  are the top- $k$  retrieved documents, and  $\theta$  are the generator parameters. This loss ensures that the language model produces fluent, coherent, and contextually relevant outputs.

*Multi-Task Optimization* In RAG, the total objective is a combination of retrieval and generation losses, typically weighted to balance the two tasks:

$$\mathcal{L}_{\text{RAG}} = \alpha \cdot \mathcal{L}_{\text{retrieval}} + \beta \cdot \mathcal{L}_{\text{generation}}, \quad (4-67)$$

where  $\alpha, \beta \in \mathbb{R}^+$  are hyperparameters controlling the trade-off.

Multi-task optimization allows the retriever to improve document selection while the generator refines contextualized text production. Joint optimization ensures end-to-end training, aligning retrieval relevance with generative quality for downstream tasks.

## 5 TabNet for Reliability-Oriented Criticality Analysis in MV Power Networks

In this objective, we propose the implementation of **TabNet**, a deep neural network architecture specifically designed for tabular data, to model reliability indices and conduct criticality analysis of medium-voltage (MV) distribution networks.

The rationale behind this choice is TabNet’s ability to combine efficiency, interpretability, and sequential feature selection, which allows capturing both endogenous and exogenous drivers of interruptions. **Endogenous variables** refer to technical and structural attributes of the electrical network, such as transformer type, line length, switch configuration, or asset age. **Exogenous variables** correspond to environmental and external conditions, including weather factors (e.g., precipitation, wind gusts, humidity, lightning density) and geographic/topographic information (e.g., altitude, terrain). By integrating both dimensions, the model can uncover how intrinsic infrastructure properties interact with external stressors to explain reliability outcomes.

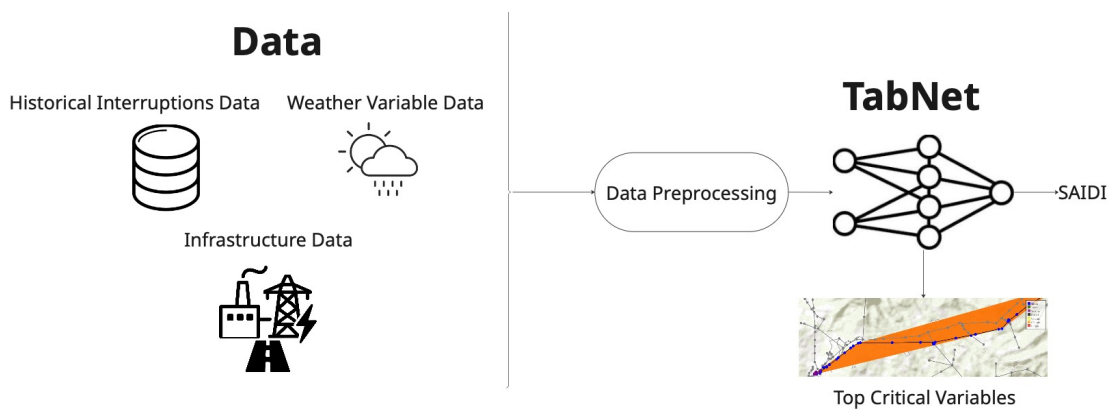
This dual perspective is particularly relevant in the context of outage prediction and criticality diagnosis, where it is necessary not only to achieve predictive accuracy but also to interpret the contribution of each input factor. In this sense, TabNet’s sequential attention masks enable a fine-grained attribution of importance to both endogenous and exogenous features, supporting transparent and actionable insights for asset management and maintenance planning.

### 5.1 Methodology

The methodological pipeline of this objective is depicted in Figure 5-1. The analysis integrates three main data sources: (i) historical interruption records (**EVENTS** database), which capture the occurrence, duration, and affected assets of each outage; (ii) meteorological and environmental datasets, providing **exogenous variables** such as precipitation, wind gusts, humidity, lightning density, temperature, and solar radiation; and (iii) infrastructure and asset registries, containing **endogenous variables** such as transformer characteristics, line length, switch configuration, and network topology.

Starting from the **EVENTS** database as the primary source, data preprocessing is performed to standardize temporal, categorical, and numerical attributes across all sources. Features are encoded through **LabelEncoder** for categorical variables and normalized using **MinMaxScaler** for continuous attributes. Missing values are imputed according to data type and contextual relationships between assets, ensuring consistency between outage records, weather conditions, and infrastructure descriptors. This integration of heterogeneous datasets provides a comprehensive representation of both technical and environmental drivers of reliability, which are then modeled by TabNet for criticality assessment.

TabNet acts as a regressor, predicting the target reliability metric (SAIDI) based on a sequential attention mechanism. The model learns dynamic feature masks ( $M[i]$ ) generated through the Sparsemax activation, ensuring sparsity and interpretability of feature contributions. A cumulative importance scale ( $P[i]$ ) is employed to diversify feature usage across decision steps. Each decision step transforms inputs through gated linear units (GLUs), batch normalization, and non-linear activations.



**Figure 5-1:** Pipeline for TabNet-based criticality analysis. Preprocessing integrates data imputation, encoding, and normalization; training optimizes TabNet hyperparameters with Bayesian search; interpretability is provided through attention masks and sparsity regularization.

## 5.2 Experimental Framework

The dataset was split into 64% training, 16% validation, and 20% testing. The target (SAIDI) was categorized into quartiles to balance partitions during optimization. Hyperparameter tuning was carried out using **Optuna** with Bayesian search. Structural hyperparameters included: attention and activation dimensions ( $n_d, n_a \in [2, 512]$ ), decision steps ( $n_{steps} \in [3, 10]$ ), and embedding size ( $emb \in [3, 70]$ ). Regularization terms explored values for  $\gamma$  and  $\lambda_{sparse}$  across log-uniform ranges  $[1e^{-12}, 1e^2]$ . Training used optimizers Adam, AdamW, SGD, and RMSprop with adaptive schedulers. Batch sizes (1024–4096) and virtual batch sizes (256–1024) were compared. Early stopping was applied with patience = 40.

The evaluation metrics included:

- Mean Absolute Error (MAE), as primary objective.
- Mean Squared Error (MSE) and Root Mean Squared Error (RMSE).
- $R^2$  coefficient of determination, as a measure of explained variance.

## 5.3 Results and Discussion

**Computational Environment.** Experiments were executed in two complementary environments. The predictive pipeline was implemented and trained on **Google Colab**, using an **NVIDIA A100 GPU** (40.0 GB VRAM) and 83.5 GB of system RAM. All experiments were performed using **Python 3.12** with a global random seed of 42 to ensure reproducibility. The implementation relied on **PyTorch v2.8.0** and **pytorch-tabnet v4.1.0**, with deterministic computation enforced by enabling cuDNN deterministic kernels (v91002) and disabling non-deterministic algorithms in PyTorch. Supporting libraries included **NumPy v2.0.2**, **cuML v25.06.00**, **cuPy v13.3.0**, and **XGBoost v3.1.1**. All source code and preprocessed datasets are available at <https://github.com/UN-GCPDS/CRITAIR> (accessed on October 30, 2025).

### 5.3.1 Convergence Behavior

Figure 5-2 shows the convergence of the training and validation loss. TabNet converged after  $\sim 75$  epochs, achieving stable reconstruction error. This indicates the model's robustness in learning complex feature interactions across technical and environmental variables.

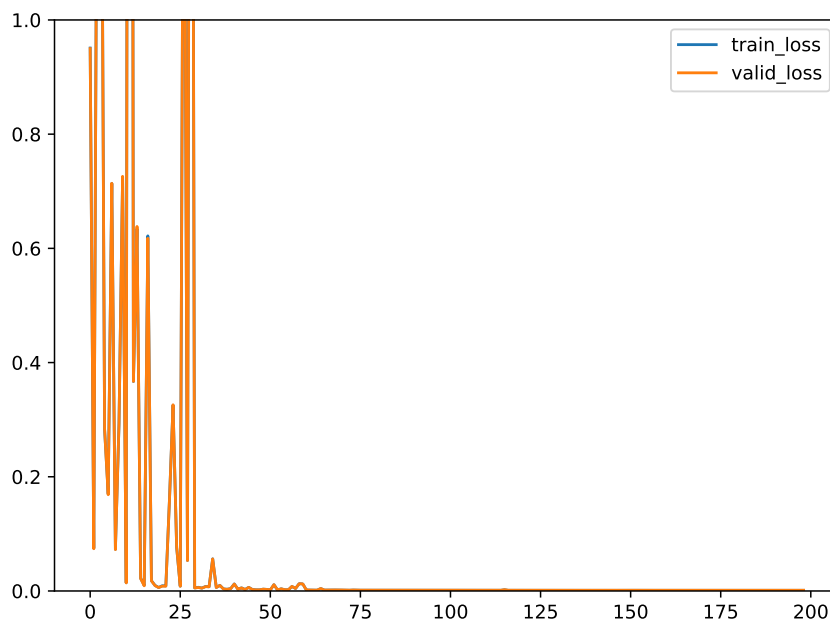


Figure 5-2: Convergence of TabNet regression loss during training.

### 5.3.2 Comparison with Baseline Regressors

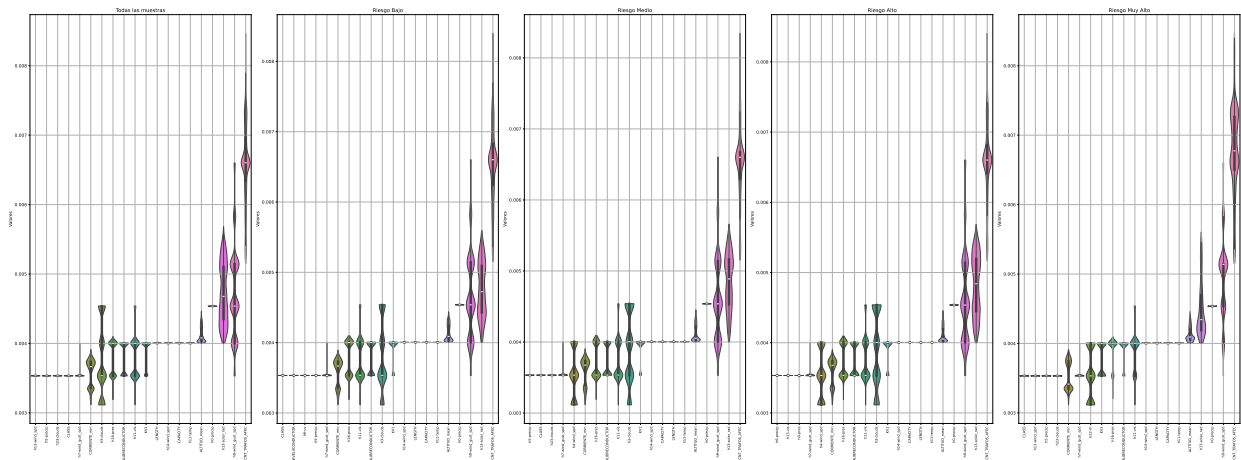
To contextualize TabNet’s performance, we compared it against three widely used regression models: a Deep Neural Network (DNN), Random Forest, and Linear Regression. Table 5-1 summarizes the results. TabNet outperformed all baselines with an  $R^2$  of 0.88, confirming its predictive superiority. Beyond accuracy, TabNet provides a distinctive advantage: the interpretability of its feature selection masks, which classical models lack.

**Table 5-1:** Performance comparison between TabNet and baseline regressors. TabNet achieved the best  $R^2$ , with the added value of interpretability.

Model	$R^2$ Score
TabNet	<b>0.94</b>
XGBoost	0.86
Random Forest	0.79
Deep Neural Network (DNN)	0.77
Linear Regression	0.7

### 5.3.3 Feature Importance and Interpretability

Attention masks were analyzed to generate global feature importance. Figure 5-3 shows violin plots for variable relevance across different risk categories (low, medium, high, very high). Environmental features (wind gusts, relative humidity, precipitation) consistently dominated higher risk levels, while structural attributes (circuit length, transformer density) were influential in lower risk ranges.



**Figure 5-3:** Feature importance distribution across risk levels derived from TabNet attention masks.

Correlation heatmaps (Figure 5-4) highlighted clusters of interacting variables, validating known operational dependencies (e.g., wind speed with pole failures, precipitation with grounding systems). Focused analysis (Figure 5-5) showed that SAIDI had the strongest correlation with transformer density, wind gust intensity, and average altitude.

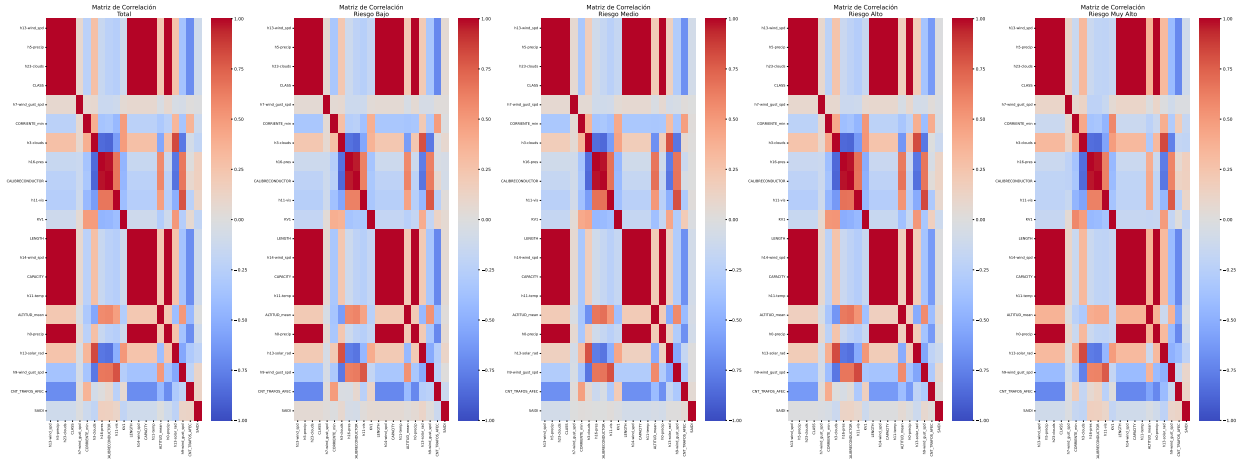


Figure 5-4: Correlation heatmap across environmental and structural variables, showing clusters of interdependent drivers influencing SAIDI.

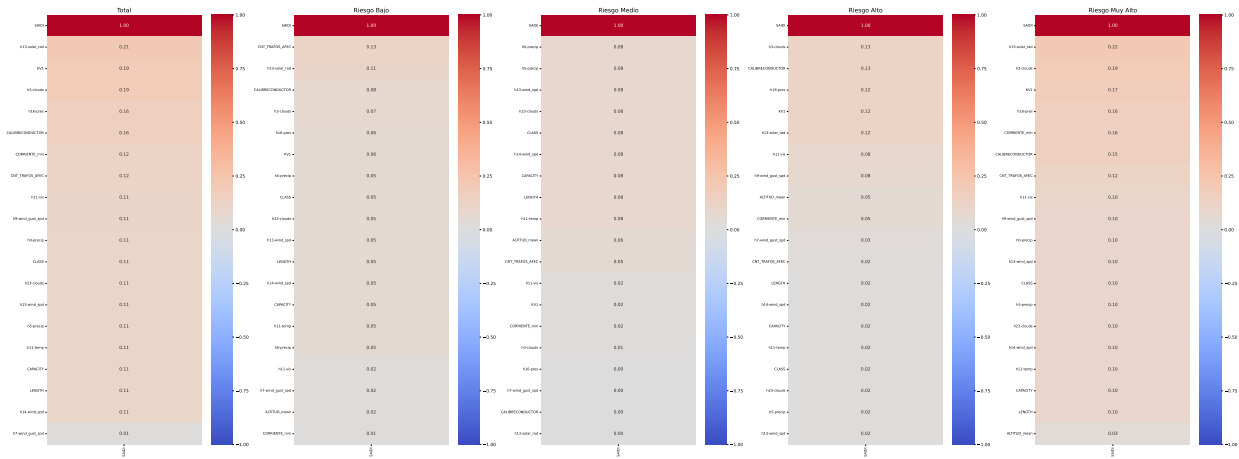


Figure 5-5: Focused correlation analysis between SAIDI and top explanatory variables across different risk categories.



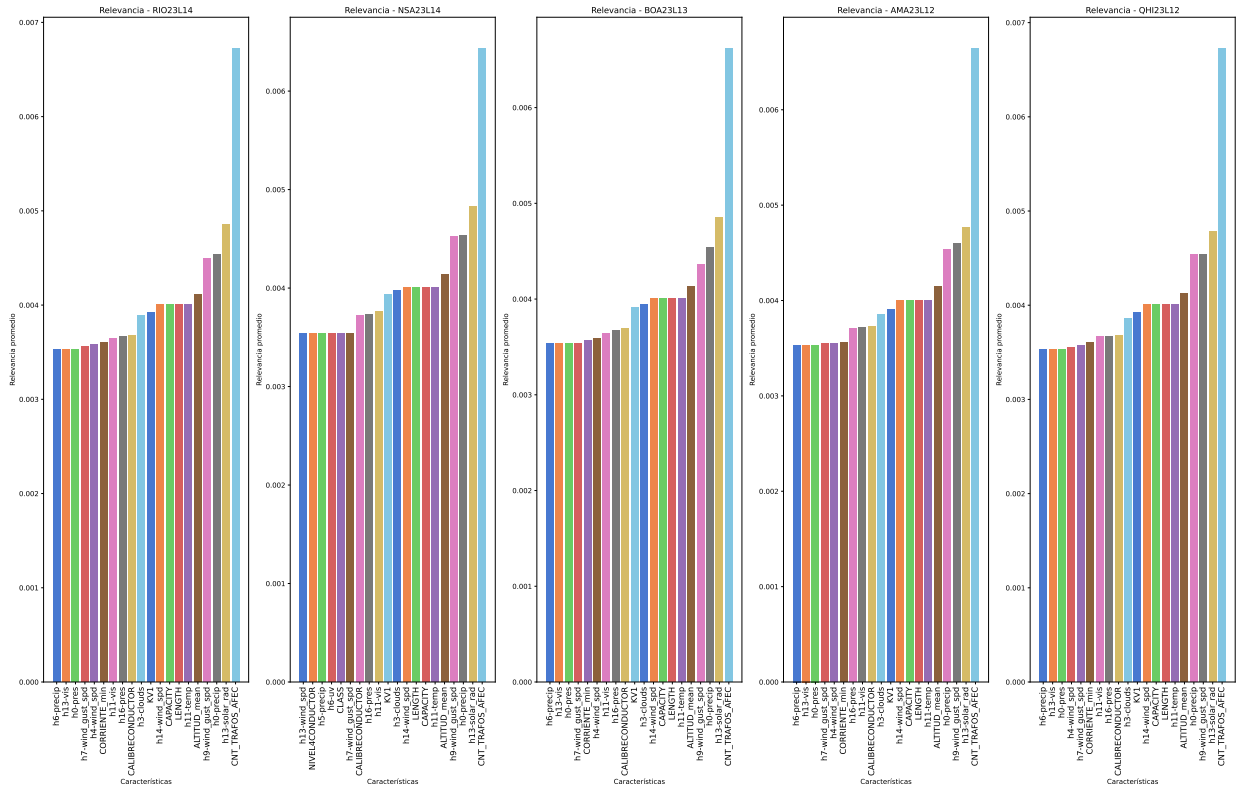


Figure 5-7: Criticality analysis by circuit, highlighting environmental and structural interactions.

These analyses reveal that municipalities located in mountainous regions (e.g., Manizales, Riosucio) and circuits exposed to frequent storms exhibit systematically higher SAIDI values. The dominant drivers were exogenous factors such as wind gusts, humidity, and precipitation, although endogenous factors such as transformer density and line length were also significant. This validates the dual-variable approach and highlights the practical value of combining technical and environmental perspectives.

## 5.4 Summary

This objective demonstrated the effectiveness of TabNet as a regression-based methodology for reliability-oriented criticality analysis in MV networks. The model provided competitive predictive accuracy ( $R^2 = 0.88$ ) while offering interpretable feature selection. Criticality patterns were successfully linked to both environmental (storms, altitude, precipitation) and structural (circuit length, asset density) drivers.

Compared with baseline regressors (DNN, Random Forest, Linear Regression), TabNet consistently outperformed in predictive power and, most importantly, provided interpretable attention masks that allowed the attribution of risk to specific variables at municipal and circuit levels.

The TabNet-based pipeline therefore constitutes a robust foundation for downstream asset prioritization, maintenance planning, and regulatory compliance, offering both predictive accuracy and transparency in reliability analysis.

## 6 Agentic RAG for Structured Data Queries, Normative Compliance, and Criticality-Based Recommendations in MV Power Networks

In this objective, we proposed an **Agentic Retrieval-Augmented Generation (RAG)** system designed to support intelligent querying and recommendation in medium-voltage (MV) distribution networks. The system integrates structured outage records, regulatory documentation, and model-driven criticality analysis into a single conversational framework. By combining retrieval modules with generative large language models (LLMs), the system provides **contextualized, explainable, and actionable responses** to both technical and normative queries.

This objective bridges three complementary dimensions: (i) natural language access to structured outage databases; (ii) retrieval of normative and regulatory guidelines (e.g., RETIE); (iii) recommendation pipelines leveraging TabNet-based criticality scores.

All experiments were carried out using the **expert-guided Q&A bitácoras** developed with CHEC engineers, described in the **Datasets** section of Chapter 4.2. These bitácoras defined the ground-truth references for evaluating accuracy and interpretability across the three query types.

---

### 6.1 System Architecture

The overall system architecture is depicted in Figure 6-1. It follows the Agentic RAG paradigm, in which the query is decomposed into retrieval and generation stages. The retrieval stage fetches either structured records or text passages, while the generation stage uses an LLM to produce a coherent answer.

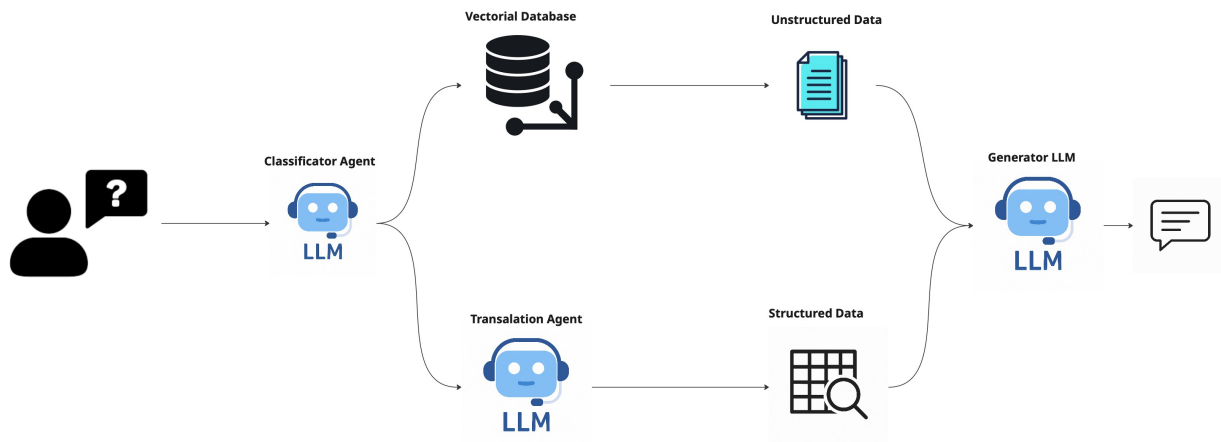


Figure 6-1: General architecture of the Agentic RAG-based chatbot for structured and unstructured queries.

The architecture is divided into three main modules, each adapted to a specific query type:

### 6.1.1 Structured Queries: Interruption Databases

For queries related to switches and transformer outages, the pipeline (see Table 6-1) translates user questions into executable pandas queries, executes them on structured tables, and returns results through a generative LLM.

Process	Description
Question translation	Natural language input is parsed by an LLM and converted into a pandas query.
Query execution	The query is executed on switch/transformer outage tables to extract relevant data.
Prompt construction	A prompt is built with user query, retrieved data, instructions, and desired output format.
Answer generation	The LLM generates a coherent, natural-language response contextualized by retrieved data.

Table 6-1: Pipeline for structured outage queries.

### 6.1.2 Unstructured Queries: Regulatory Documents

For normative and regulatory compliance (e.g., RETIE), the pipeline (see Table 6-2) retrieves text passages from a vector database of standards and uses them as context for the LLM.

## 6. Agentic RAG-based Conversational System for Asset Queries and Recommendations System Architecture

Process	Description
Query reception	The user submits a natural language query to the system.
Retrieval module	The system searches the vector database to extract the most relevant passages.
Prompt construction	The prompt includes instructions, user query, retrieved fragments, and output format.
Answer generation	The LLM generates a response grounded in the retrieved normative content.

Table 6-2: Pipeline for normative queries.

### 6.1.3 Criticality-Based Recommendations

For recommendations, the system integrates the TabNet-based criticality analysis (Figure 6-2). Predictions of SAIDI at the asset level are used to identify critical elements. Recommendations are then generated by retrieving technical/normative knowledge relevant to the most influential variables.

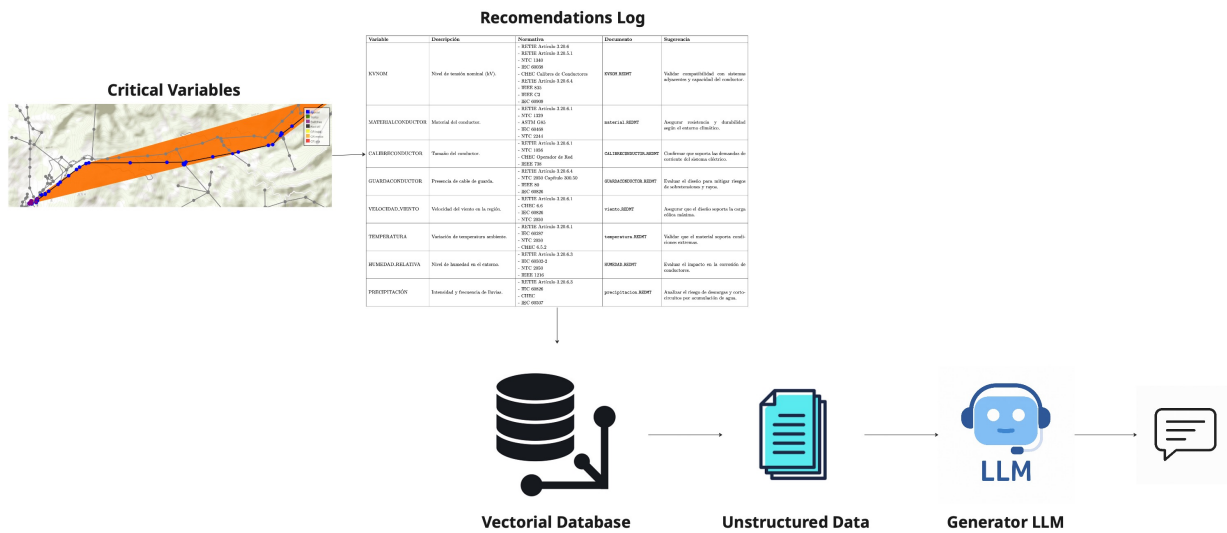


Figure 6-2: Architecture of the recommendation system combining TabNet outputs with RAG.

Process	Description
Input data structure	The element (transformer, switch, line, pole) is represented with its key variables and TabNet criticality scores.
Retrieval module	The system retrieves relevant fragments (e.g., RETIE clauses, technical guidelines) linked to the critical variables.
Prompt construction	A recommendation prompt integrates retrieved context with TabNet-derived criticality analysis.
Recommendation generation	The LLM produces actionable recommendations tailored to the asset and its most critical factors.

Table 6-3: Pipeline for criticality-based recommendations.

## 6.2 Experimental Framework

To evaluate the system, we designed three categories of queries:

1. Structured queries (outage databases).
2. Unstructured queries (regulatory documents).
3. Recommendation queries (critical assets).

A total of ten LLMs were tested as backbones for response generation. Of these, only six could be executed locally (others were tested only via API). Evaluation combined:

- **BERTScore**: semantic similarity between generated and reference answers.
- **Inference time**: computational cost for real-time responses (measured only on local models).

This dual evaluation allowed us to assess both **performance** (accuracy of answers) and **efficiency** (latency).

**Implementation Platform and Model Configuration.** All open-weight language models evaluated in this study were executed locally using the **Ollama** platform, a lightweight runtime that supports efficient inference and management of quantized large language models (LLMs). This framework was selected for its compatibility with modern architectures, its transparent configuration of inference parameters, and its reproducibility across experiments.

Each model was run in its **chat variant**, consistent with the conversational nature of the Agentic-RAG system, which requires contextual continuity and multi-turn reasoning. For every model, the default configuration provided by Ollama was employed, including its respective **quantization scheme** (e.g., Q4-K-M, Q4-K-S, or equivalent, depending on the model). These defaults were verified against the official Ollama documentation to ensure consistent and comparable experimental conditions.

The hyperparameters summarized in **Table 4.9** (“Configuration of evaluated LLMs”) specify the effective context length, maximum generation length (“Max Tokens”), and quantization method used during testing. In this context, “Max Tokens” refers to the maximum number of tokens that each model could generate as an output sequence, while “Context Length” indicates the window size explicitly configured for prompt and retrieved content.

All experiments were executed on a **local workstation** running **Ubuntu 22.04**, equipped with an **Intel Core i9-11900 CPU, 64 GB RAM**, and an **NVIDIA RTX 3070 Ti GPU (8 GB VRAM)**. Deterministic behavior was enforced using a global random seed of 42 and fixed temperature  $T=0.0$ .

For models that expose explicit reasoning traces (e.g., `<think> . . . </think>` blocks in the *DeepSeek-R1* family), only the final user-oriented portion of the response was considered during evaluation, excluding internal reasoning text to ensure fair semantic comparison.

Finally, none of the generated responses exceeded the 8192-token embedding limit of the *text-embedding-3-small* model used for **BERTScore** computation. Consequently, no truncation or segmentation procedures were required. Should longer generations arise in future work, a fragment-based embedding averaging strategy would be applied to preserve semantic consistency without biasing similarity metrics.

All source code and preprocessed datasets are available at <https://github.com/UN-GCPDS/CRITAIR> (accessed on October 30, 2025).

## 6.3 Results and Discussion

### 6.3.1 Structured Data Queries

Figure 6-3 reports results for all ten models. The highest semantic accuracy was achieved by **GPT-3.5 Turbo**, with a BERTScore of **0.956**, but with a non-negligible inference time since it was executed via API. The best trade-off between accuracy and inference time was provided by **Llama 3.2:1B**, reaching a BERTScore of **0.736** with an inference time of **3.04 seconds**.

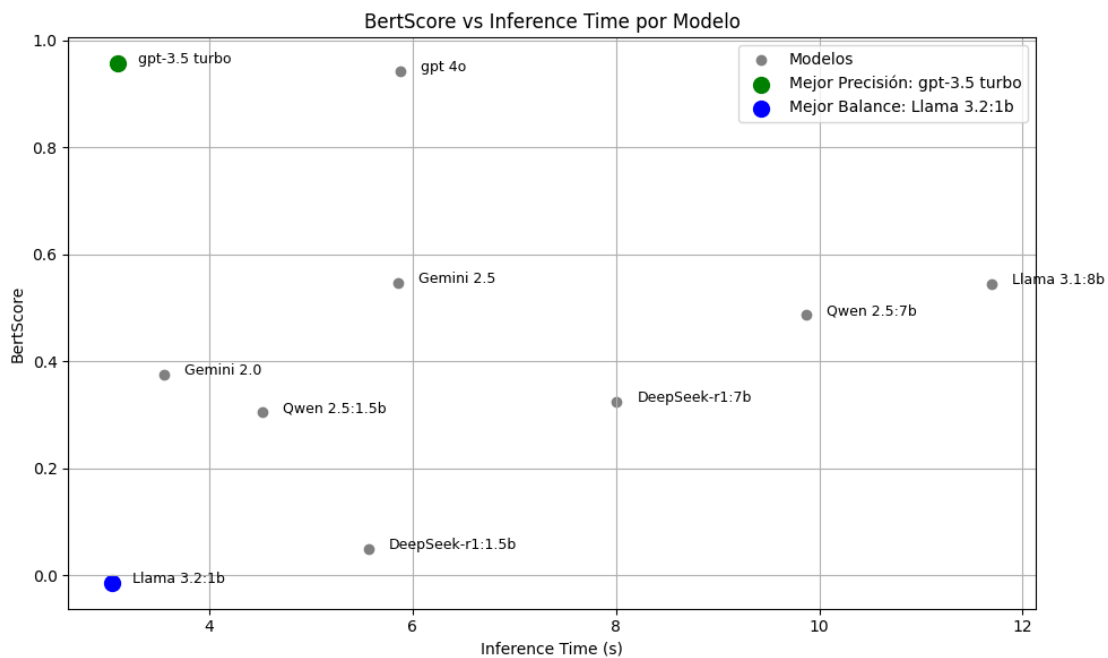


Figure 6-3: Performance of ten LLMs in structured outage queries (BERTScore vs inference time).

Restricting the comparison to the six models that could be executed locally (Figure 6-4), the best balance remains with **Llama 3.2:1B**, since GPT-3.5 Turbo cannot be considered under identical execution conditions.

### 6.3. Results and Discussion: Asynchronous RAG-based Conversational System for Asset Queries and Recommendations

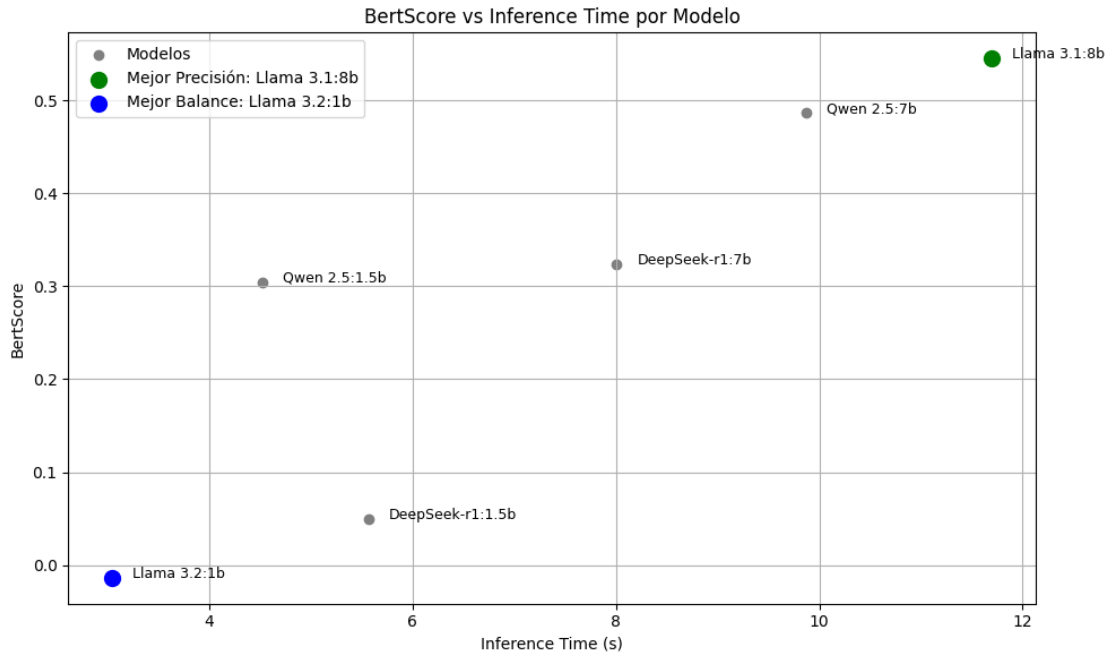


Figure 6-4: Performance of six LLMs executed locally in structured outage queries.

### 6.3.2 Unstructured Normative Queries

For queries grounded in normative documents, **GPT-3.5 Turbo** again obtained the highest semantic accuracy, with a BERTScore of **0.9847** and an inference time of **11.54 seconds** (Figure 6-5). This model also provided the best accuracy-efficiency balance, but since it was executed via API, comparisons under local conditions are necessary.

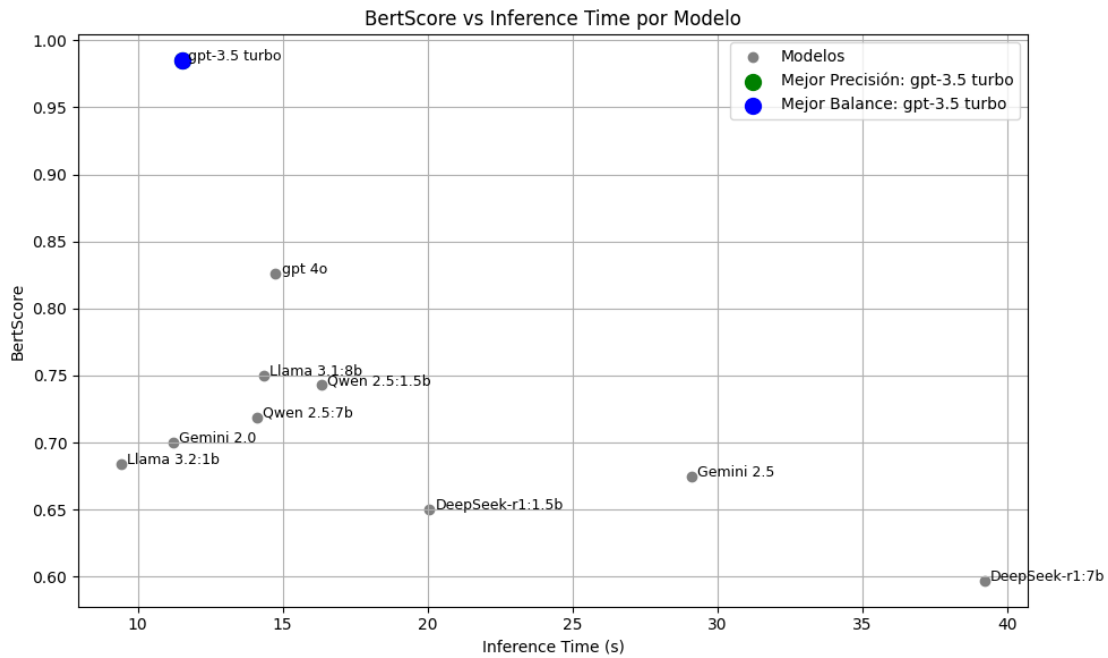


Figure 6-5: Performance of ten LLMs in unstructured normative queries.

Considering only locally executed models (Figure 6-6), the best balance was achieved by **Llama 3.2:1B**, with a BERTScore of **0.684** and an inference time of **9.43 seconds**. While lower in accuracy than GPT-3.5 Turbo, this result highlights its practicality for real-time deployment in on-premise conditions.

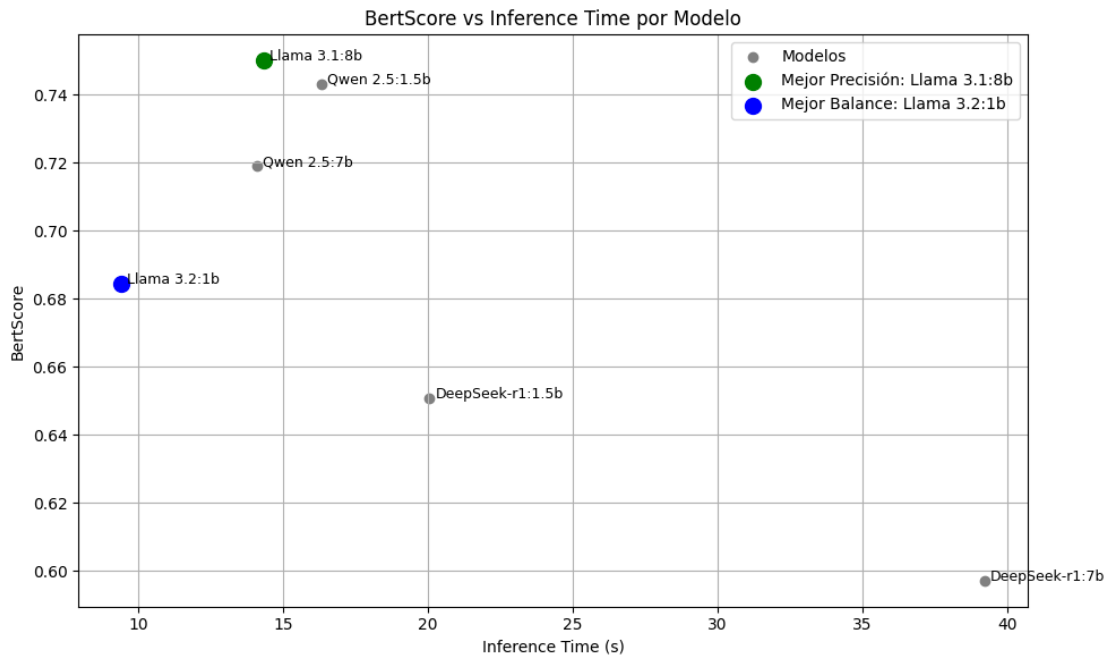


Figure 6-6: Performance of six LLMs executed locally in unstructured normative queries.

### 6.3.3 Recommendation Queries

In the recommendation setting, integrating TabNet outputs with normative retrieval, performance differences between models widened (Figure 6-7). The best semantic accuracy was achieved by **Gemini 2.5**, with a BERTScore of **0.823**, but with elevated computational cost. The best balance among the ten models was offered by **GPT-3.5 Turbo**, which achieved a BERTScore of **0.743** with an inference time of **2.23 seconds**.

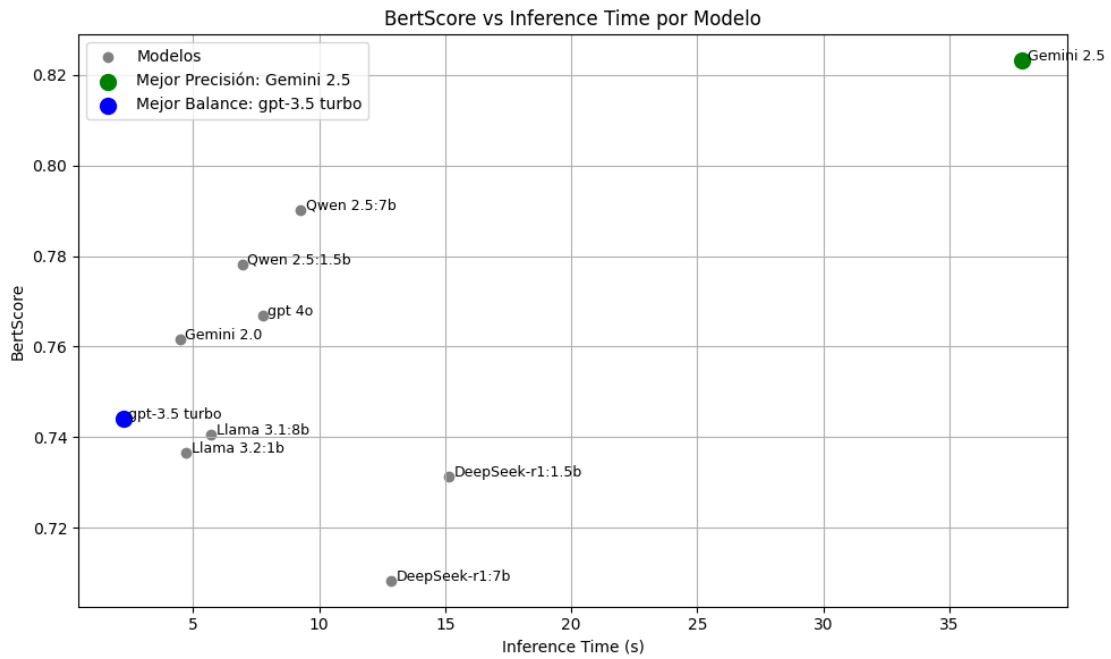


Figure 6-7: Performance of ten LLMs in recommendation queries.

Since GPT-3.5 Turbo was executed only via API, in local conditions the best balance was achieved by **Llama 3.2:1B**, which reached a BERTScore of **0.736** with an inference time of **4.74 seconds** (Figure 6-8).

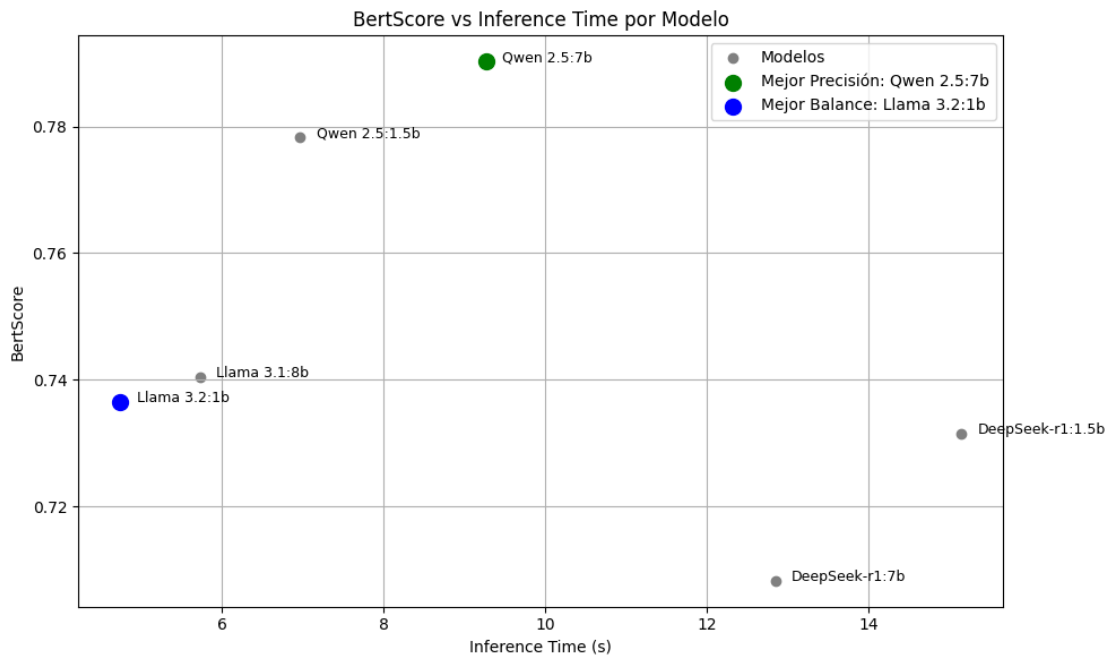


Figure 6-8: Performance of six LLMs executed locally in recommendation queries.

### 6.3.4 Discussion

The results highlight the importance of jointly evaluating **semantic accuracy** and **computational efficiency**.

- **Structured queries:** GPT-3.5 Turbo was the most accurate, but Llama 3.2:1B offered the best balance locally.
- **Unstructured queries:** GPT-3.5 Turbo excelled in both accuracy and trade-off overall, but locally Llama 3.2:1B was the most practical.
- **Recommendation queries:** Gemini 2.5 achieved the highest accuracy, GPT-3.5 Turbo the best global trade-off, and Llama 3.2:1B the best local trade-off.

These findings demonstrate the feasibility of deploying Agentic RAG pipelines with heterogeneous data sources, but also emphasize that model selection must balance accuracy and deployment constraints. While API-based models deliver higher semantic accuracy, locally deployable models like Llama 3.2:1B provide robust performance with acceptable latency, making them suitable for operational integration within CHEC's infrastructure.

—

## 6.4 Summary

This objective validated the implementation of a RAG-based system for MV networks, integrating structured outage data, normative guidelines, and TabNet-driven criticality analysis. Results demonstrated:

- The feasibility of using LLMs for real-time technical and normative queries.
- The superiority of certain models in terms of BERTScore, balanced with inference cost.
- The value of combining TabNet outputs with RAG for interpretable, actionable recommendations.

By enabling transparent and data-driven recommendations, the system supports both compliance (RETIE) and operational resilience, offering a robust tool for decision-making in power distribution networks.

## 7 Interpretable Reasoning Graphs for Transparent Recommendations in MV Power Networks

In this objective, we address the challenge of **explainability and trust** in AI-driven recommendation systems for power networks. While predictive models such as TabNet provide accurate estimations of reliability indices and asset criticality, utility engineers require not only predictions but also **transparent reasoning paths** that clarify how recommendations are derived. To meet this need, we designed and implemented **interpretable reasoning graphs** that explicitly represent the logical flow followed by the recommendation system.

This objective therefore complements Objectives 1 and 2 by focusing on **auditability and interpretability**, ensuring that recommendations for asset management are transparent, traceable, and aligned with regulatory and operational standards.

---

### 7.1 End-to-End Methodology and Reasoning Flow

The full reasoning pipeline begins once the TabNet model has processed the data from a network section affected by an outage. TabNet outputs the top three assets with the highest estimated contribution to the SAIDI index. For each of these assets, the five most influential variables are extracted based on feature attribution scores.

As an illustrative example, consider the following TabNet output for a particular event:

- **Asset ID: B38007 Type: Pole Top 5 critical variables:** {'h13-solar\_rad': 740.0, 'h14-temp': 23.7, 'CNT\_TRAFOS\_AFEC': 1.0, 'h22-rh': 76.0, 'h6-precip': 0.0} **Coordinates:** 5.2811, -75.0925

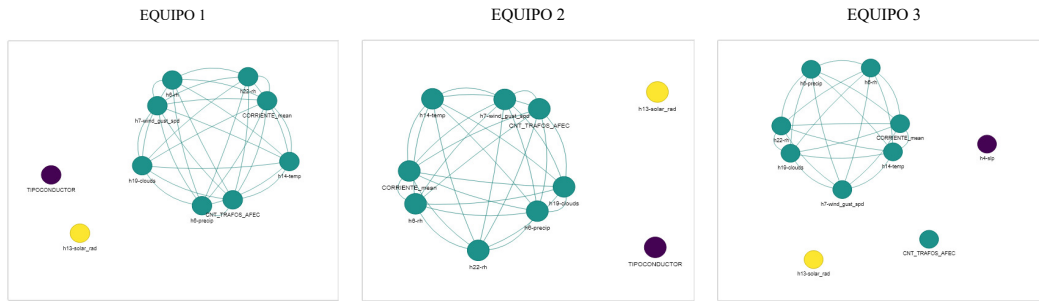
This information provides a technical snapshot of the asset and the potential environmental and operational

7.1. End-to-End Methodology and Reasoning Flow

Interpretable Reasoning Graphs for Trust and Auditability

variables associated with the failure.

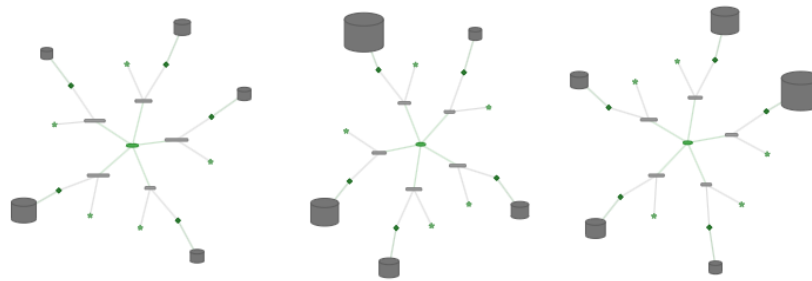
To better interpret these results, the system generates a **graph of variable relevance** across the top three critical assets, highlighting interdependencies and relative importance. Figure 7-1 shows this graph, where yellow nodes represent the most critical variables (importance close to 1), and blue nodes indicate minimal influence.



**Figure 7-1:** Graph illustrating the relationship between the most influential variables across the top three critical assets. Color intensity denotes normalized relevance.

Next, the system initiates the **reasoning phase** using a large language model (LLM). For each critical variable, the system retrieves relevant content from domain-specific regulatory documents (e.g., RETIE, NTC 1329, ASCE 7), identifies evaluation criteria, and produces a recommendation.

To make this reasoning path explicit, we designed visual **reasoning graphs** (Figure 7-2) that show the internal decision logic of the LLM-based recommendation module.



**Figure 7-2:** Reasoning graph showing how the LLM processes critical variables to generate technical recommendations.

## 7. Interpretable Reasoning Graphs for Trust and Auditability: End-to-End Methodology and Reasoning Flow

For instance, for the variable **Pole Length** of asset B38007, the system retrieved the clauses from RETIE, specifically NTC 1329 and ASCE 7 (Figure 7-3), which provide guidance on structural specifications for poles.

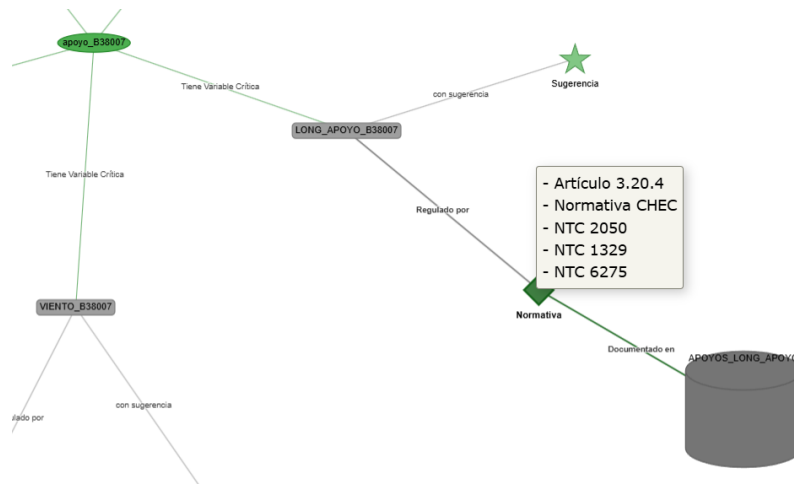


Figure 7-3: Retrieval of relevant regulatory sources used by the LLM to contextualize the variable Pole Length.

Subsequently, as seen in Figure 7-4, the LLM determines which specific evaluation criteria should be verified. In this case, minimum compliance requirements such as height, material properties, and resistance are extracted and linked to the criticality detected by TabNet.

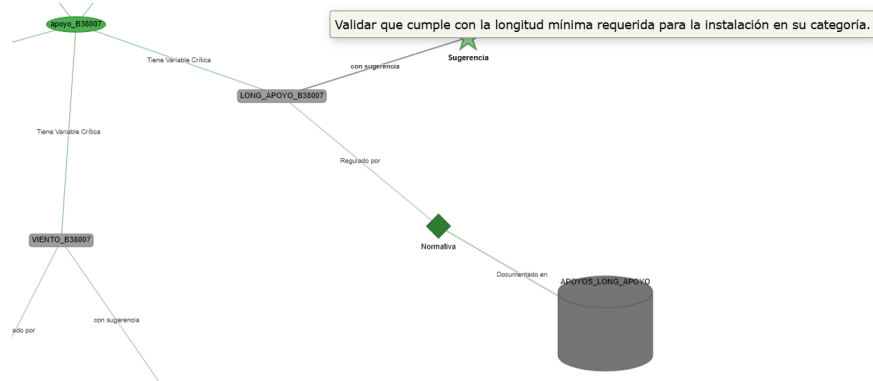
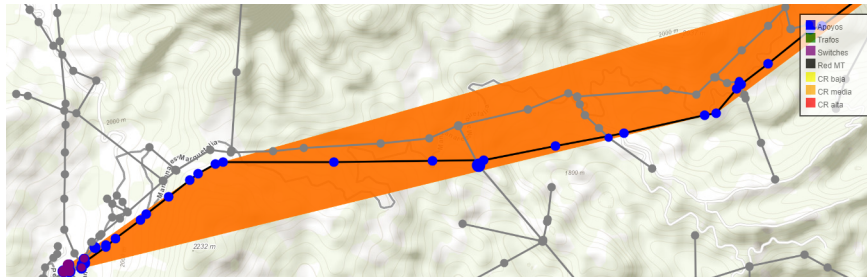


Figure 7-4: Interpretation of specific compliance criteria for the variable Pole Length.

Finally, the recommendations are contextualized within the affected network section. Figure 7-5 shows the spatial layout of assets involved in the event, where the radius of each circle denotes its contribution to the SAIDI index. This visualization supports prioritization for inspection or preventive maintenance.



**Figure 7-5:** Graph showing asset contribution to SAIDI for the event. Larger circles indicate higher impact.

Through this end-to-end workflow—from TabNet output to regulatory validation—the system not only identifies failure-prone assets but also delivers actionable, traceable, and technically grounded recommendations.

## 7.2 Results and Expert Validation

Unlike the previous objectives that were evaluated through quantitative benchmarks (e.g., regression metrics or BERTScore), this component was validated through a **qualitative review with domain experts from CHEC**. Engineers were invited to audit the full reasoning chain—starting from TabNet’s variable outputs to the final LLM-generated recommendation and its supporting documentation.

Key findings from expert feedback include:

- The **reasoning graphs** improved confidence in the AI system, especially by making each decision step explicit and auditable.
- Domain experts valued the clear distinction between **evidence retrieval** (i.e., regulatory text) and the **interpretation of actionable criteria**.
- The integration of technical knowledge with model predictions mirrors their actual diagnostic workflows, making the system both interpretable and operationally relevant.

## 7.3 Summary

This objective demonstrated that incorporating **interpretable reasoning graphs** and document-grounded recommendations significantly enhances the trust and auditability of AI systems in power networks. By integrating graph-based visualizations, regulatory clause retrieval, and criteria interpretation, the system transforms black-box predictions into transparent decision-support tools.

---

Field validation with engineers confirmed that the recommendations were not only interpretable but also **aligned with their maintenance practices and regulatory requirements**. These results highlight the value of combining data-driven learning with domain knowledge to enable more responsible and effective AI in critical infrastructure.

## 8 Final Remarks

### 8.1 Conclusions

This thesis presented a novel **hybrid AI methodology** that combines predictive modeling, regulatory retrieval, and interpretable reasoning to support reliability-oriented decision-making in medium-voltage (MV) power networks. The core contribution lies in integrating **TabNet-based regression**, **Agentic Retrieval-Augmented Generation (RAG)**, and **reasoning graphs**, enabling both accurate predictions and transparent, regulation-aware recommendations. Compared with the state of the art, this methodology advances the field by unifying structured and unstructured data sources into a single explainable framework, bridging the gap between technical analytics and regulatory compliance.

Regarding **Specific Objective 1**, the implementation of TabNet demonstrated strong predictive power for the SAIDI index, achieving superior accuracy compared to classical regressors (e.g., Random Forest, Linear Regression, DNNs) while maintaining intrinsic interpretability through sequential feature masks. The integration of both endogenous variables (infrastructure characteristics) and exogenous variables (meteorological and environmental factors) provided actionable insights into the drivers of reliability, allowing CHEC engineers to trace outage patterns to both technical and environmental causes.

For **Specific Objective 2**, the design of an Agentic RAG system enabled natural language access to structured outage data and normative documents, as well as criticality-based recommendations. Evaluation using expert-guided Q&A bitácoras showed that large language models can effectively combine heterogeneous knowledge sources, with performance measured through semantic similarity (BERTScore) and efficiency (inference time). This system provides a practical decision-support tool, empowering operators to query technical databases, validate RETIE compliance, and receive data-driven recommendations in real time.

With respect to **Specific Objective 3**, interpretable reasoning graphs were introduced as a visualization layer to enhance trust and auditability. These graphs explicitly traced the reasoning steps of the agent, showing how critical variables identified by TabNet were linked to regulatory clauses and evaluation criteria. Expert validation by CHEC engineers confirmed the usefulness of these reasoning processes in real operational scenarios, emphasizing that transparency and traceability are key to adopting AI in high-stakes domains such as energy distribution.

## 8.2 Future Work

Future research directions extend and deepen the contributions developed across the three objectives of this thesis. The following lines summarize potential areas of exploration:

- **Objective 1 — Enhanced Predictive Modeling:**

- *Integration of ensemble and multimodal architectures:* Combine TabNet with other models (e.g., Random Forests, Gradient Boosting, or deep CNNs) to capture both tabular and visual information. This could improve robustness by leveraging complementary data modalities.
- *Incorporation of geospatial and image-based data:* Fuse traditional tabular features with high-resolution satellite imagery, LiDAR data, or drone-based inspections to better represent environmental and infrastructural contexts influencing reliability.
- *Temporal generalization and transfer learning:* Evaluate model adaptability over time using domain adaptation and transfer learning strategies to ensure consistent performance under evolving network and climatic conditions.

- **Objective 2 — Expansion of the RAG Architecture:**

- *Transition to multi-agent RAG systems:* Develop a multi-agent pipeline where distinct agents specialize in document retrieval, normative interpretation, and recommendation synthesis, coordinated through a supervisory agent.
- *Continuous learning and feedback loops:* Implement human-in-the-loop mechanisms allowing engineers to correct or refine recommendations, progressively improving the system’s accuracy and trustworthiness.
- *Knowledge graph enrichment:* Integrate the RAG system with dynamic knowledge graphs that store normative relationships, contextual metadata, and temporal evolution of assets.

- **Objective 3 — Advanced Interpretability and Simulation:**

- *Interactive reasoning dashboards:* Extend reasoning graphs into interactive visualization tools that allow users to explore decision pathways, parameter sensitivity, and underlying evidence.
- *Causal inference integration:* Incorporate causal discovery modules (e.g., do-calculus, structural causal models) to enable “what-if” scenario simulations and counterfactual explanations for outage mitigation.
- *Trust and regulatory compliance frameworks:* Evaluate the interpretability and fairness of recommendations against industry standards (e.g., IEEE, NTC, RETIE) to ensure alignment with operational and ethical guidelines.

- **Towards an End-to-End Reliability Assistant:**

- *Unified AI system for predictive and prescriptive analytics:* Integrate the outcomes of the three objectives into a comprehensive intelligent assistant capable of predicting outages, explaining reasoning, and suggesting preventive or corrective actions.
- *Scalability and real-time deployment:* Adapt the full pipeline for integration with existing utility platforms and SCADA systems, supporting continuous data ingestion and real-time recommendations.
- *Human-AI collaboration:* Explore mechanisms for joint decision-making, where the assistant provides transparent justifications, uncertainty estimates, and ranked alternatives to empower engineers’ situational awareness.

---

Together, these future research lines pave the way toward an **end-to-end intelligent reliability assistant** that not only predicts outages but also explains, recommends, and justifies interventions, thereby advancing the state of AI applications in electrical power systems.

## References

- [cre, 2008] , 2008; Resolución creg 097 de 2008: Régimen de calidad del servicio de energía eléctrica; *Informe técnico*; Comisión de Regulación de Energía y Gas (CREG); URL <https://www.creg.gov.co>.
- [iee, 2012a] , 2012a; Ieee guide for electric power distribution reliability indices; doi:10.1109/IEEESTD.2012.6209381.
- [iee, 2012b] , 2012b; Ieee guide for electric power distribution reliability indices; doi:10.1109/IEEESTD.2012.6209381.
- [cre, 2018] , 2018; Resolución creg 015 de 2018: Actualización del régimen de calidad del servicio de energía eléctrica; *Informe técnico*; Comisión de Regulación de Energía y Gas (CREG); URL <https://www.creg.gov.co>.
- [ret, 2022] , 2022; Reglamento técnico de instalaciones eléctricas (retie) — actualización 2022; URL <https://www.minenergia.gov.co>.
- [Aldhubaib et al., 2023a] **Aldhubaib, A.; Al-Gailani, S. & Yaseen, Z.** , 2023a; Climate change impacts on electrical power distribution networks: A review; *Energy Reports*; **9**: 256–272; doi:10.1016/j.egy.2022.11.275.
- [Aldhubaib et al., 2023b] **Aldhubaib, H. A.; Hassan Ahmed, M. & Salama, M. M.** , 2023b; A weather-based power distribution system reliability assessment; *Alexandria Engineering Journal*; **78**: 256–264; doi:<https://doi.org/10.1016/j.aej.2023.07.033>; URL <https://www.sciencedirect.com/science/article/pii/S1110016823006154>.
- [Arik & Pfister, 2021] **Arik, S. Ö. & Pfister, T.** , 2021; Tabnet: Attentive interpretable tabular learning; en *Proceedings of the AAAI conference on artificial intelligence*, tomo 35; págs. 6679–6687.
- [Bai et al., 2025] **Bai, Y.; Wei, J.; Zhang, Z.; Yu, X.; Wang, M.; Liu, Y.; Zhao, W. et al.** , 2025; Deepseek-v2: Towards language agents with world models; *arXiv preprint arXiv:2503.06233*.
- [Billinton & Allan, 1996a] **Billinton, R. & Allan, R. N.** , 1996a; Reliability evaluation of power systems; *Springer*; doi:10.1007/978-1-4899-1860-0.
- [Billinton & Allan, 1996b] **Billinton, R. & Allan, R. N.** , 1996b; *Reliability Evaluation of Power Systems*; Springer; 2<sup>a</sup> edición; doi:10.1007/978-1-4899-1860-4.
- [Bishop, 2006] **Bishop, C. M.** , 2006; *Pattern Recognition and Machine Learning*; Springer.
- [Böckling et al., 2025] **Böckling, M.; Paulheim, H. & Iana, A.** , 2025; Walk&retrieve: Simple yet effective zero-shot retrieval-augmented generation via knowledge graph walks; *arXiv preprint arXiv:2505.16849*.
- [Bouadi et al., 2025] **Bouadi, H.; Singh, A. & Patel, R.** , 2025; Kg-smile: Knowledge graph-guided semantic attribution for interpretable rag systems; *arXiv preprint arXiv:2509.03626*; URL <https://arxiv.org/abs/2509.03626>.
- [Brown et al., 2020] **Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A. et al.** , 2020; Language models are few-shot learners; *Advances in neural information processing systems*; **33**: 1877–1901.
- [Chatterjee & Dethlefs, 2020] **Chatterjee, J. & Dethlefs, N.** , 2020; Xai4wind: A multimodal knowledge graph database for explainable decision support in operations & maintenance of wind turbines; *arXiv preprint arXiv:2012.10489*.
- [Chen et al., 2022] **Chen, X.; Wang, Y. & Liu, H.** , 2022; Application of knowledge graph technology in fault diagnosis of power systems; *Frontiers in Energy Research*; **10**: 988280; doi:10.3389/fenrg.2022.988280.
- [Chen et al., 2025] **Chen, Y.; Zhang, E.; Yan, L.; Wang, S.; Huang, J.; Yin, D. & Mao, J.** , 2025; Mao-arag: Multi-agent orchestration for adaptive retrieval-augmented generation; *arXiv preprint arXiv:2508.01005*.
- [Chicco et al., 2021] **Chicco, D.; Warrens, M. J. & Jurman, G.** , 2021; The coefficient of determination  $r^2$  and adjusted  $r^2$  in regression analysis; *WIREs Data Mining and Knowledge Discovery*.
- [Christiano et al., 2017] **Christiano, P. F. et al.** , 2017; Deep reinforcement learning from human preferences; en *NeurIPS*.
- [Dauphin et al., 2017] **Dauphin, Y. N.; Fan, A.; Auli, M. & Grangier, D.** , 2017; Language modeling with gated convolutional networks; en *Proceedings of the 34th International Conference on Machine Learning (ICML)*; PMLR; págs. 933–941; URL <https://proceedings.mlr.press/v70/dauphin17a.html>.

- [dechgummarn et al., 2023] dechgummarn, Y.; Fuangfoo, P. & Kampeerawat, W.: , 2023; Predictive reliability analysis of power distribution systems considering the effects of seasonal factors on outage data using weibull analysis combined with polynomial regression; *IEEE Access*; **PP**: 1–1; doi:10.1109/ACCESS.2023.3340515.
- [Dehghanian et al., 2011] Dehghanian, P.; Fotuhi-Firuzabad, M. & Razi-Kazemi, A.: , 2011; An approach for critical component identification in reliability-centered maintenance of power distribution systems based on analytical hierarchical process.
- [Delavechia et al., 2023] Delavechia, R.; Petry Ferraz, B.; Weiand, R.; Silveira, L.; Ramos, M.; Santos, L.; Bernardon, D. & Garcia, R.: , 2023; Electricity supply regulations in south america: A review of regulatory aspects; *Energies*; **16**: 915; doi:10.3390/en16020915.
- [Devlin et al., 2018] Devlin, J.; Chang, M.-W.; Lee, K. & Toutanova, K.: , 2018; Bert: Pre-training of deep bidirectional transformers for language understanding; *arXiv preprint arXiv:1810.04805*.
- [Devlin et al., 2019] Devlin, J.; Chang, M.-W.; Lee, K. & Toutanova, K.: , 2019; Bert: Pre-training of deep bidirectional transformers for language understanding; en *NAACL-HLT*.
- [Dimitriou & Tsakalidis, 2020] Dimitriou, N. & Tsakalidis, A.: , 2020; A new batch normalization technique for deep neural networks; *Neural Computing and Applications*; **32** (18): 14511–14523; doi:10.1007/s00521-020-04843-5.
- [Dorji et al., 2025] Dorji, J.; Yangzom, S.; Phuntsho, T.; Choden, D. & Sd, D.: , 2025; *THE ROLE OF KNOWLEDGE GRAPHS IN EXPLAINABLE AI*.
- [Dosovitskiy et al., 2021] Dosovitskiy, A. et al.: , 2021; An image is worth 16x16 words: Transformers for image recognition at scale; en *ICLR*.
- [Du et al., 2025] Du, C.; Zhang, L. & Chen, M.: , 2025; Foundations and applications of random forest regression for data-driven modeling; *Expert Systems with Applications*; **239**: 122466; doi:10.1016/j.eswa.2024.122466.
- [Ghasemkhani et al., 2024] Ghasemkhani, B.; Kut, R. A.; Yilmaz, R.; Birant, D.; Arıkök, Y. A.; Güzeyol, T. E. & Kut, T.: , 2024; Machine learning model development to predict power outage duration (pod): A case study for electric utilities; *Sensors*; **24** (13); doi:10.3390/s24134313; URL <https://www.mdpi.com/1424-8220/24/13/4313>.
- [Guo et al., 2019] Guo, J.; Fan, Y.; Ai, Q. & Croft, W. B.: , 2019; A deep look into neural ranking models for information retrieval; *Information Processing & Management*.
- [Izacard & Grave, 2021] Izacard, G. & Grave, E.: , 2021; Leveraging passage retrieval with generative models for open-domain question answering; en *ACL*.
- [Jadon, 2022] Jadon, S.: , 2022; A comprehensive review of loss functions in machine learning: From regression to generative adversarial networks; *arXiv preprint arXiv:2208.04874*.
- [Johnson et al., 2019] Johnson, J.; Douze, M. & Jégou, H.: , 2019; Billion-scale similarity search with gpus; *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- [Jørgensen & Ma, 2025] Jørgensen, B. N. & Ma, Z. G.: , 2025; Regulating ai in the energy sector: A scoping review of eu laws, challenges, and global perspectives; *Energies*; **18** (9); doi:10.3390/en18092359; URL <https://www.mdpi.com/1996-1073/18/9/2359>.
- [Karpukhin et al., 2020] Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P. S.; Wu, L.; Edunov, S.; Chen, D. & Yih, W.-t.: , 2020; Dense passage retrieval for open-domain question answering.; en *EMNLP (1)*; págs. 6769–6781.
- [Kostopoulos et al., 2024] Kostopoulos, G.; Davrazos, G. & Kotsiantis, S.: , 2024; Explainable artificial intelligence-based decision support systems: A recent review; *Electronics*; **13** (14); doi:10.3390/electronics13142842; URL <https://www.mdpi.com/2079-9292/13/14/2842>.
- [Krstivojević & Stojković Terzić, 2025] Krstivojević, J. & Stojković Terzić, J.: , 2025; Enhancing reliability performance in distribution networks using monte carlo simulation for optimal investment option selection; *Applied Sciences*; **15** (8); doi:10.3390/app15084209; URL <https://www.mdpi.com/2076-3417/15/8/4209>.
- [Kumar et al., 2024] Kumar, A.; Sharma, R. & Singh, A.: , 2024; Random forest algorithms: A comprehensive review and future directions; *Artificial Intelligence Review*; **57** (3): 1881–1909; doi:10.1007/s10462-023-10568-2.
- [Lewis et al., 2019] Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V. & Zettlemoyer, L.: , 2019; Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension; *arXiv preprint arXiv:1910.13461*.
- [Lewis et al., 2020] Lewis, P. et al.: , 2020; Retrieval-augmented generation for knowledge-intensive nlp tasks; en *NeurIPS*.
- [Li et al., 2023] Li, J.; Zhang, L. & Zhou, P.: , 2023; Knowledge graph construction for fault diagnosis in power systems; *Electronics*; **12** (23): 4808; doi:10.3390/electronics12234808.
- [Lin et al., 2025] Lin, J.; Xie, R.; Lin, H.; Guo, X.; Mao, Y. & Fang, Z.: , 2025; A study on the key factors influencing power grid outage restoration times: A case study of the jixi area; *Processes*; **13** (9); doi:10.3390/pr13092708; URL <https://www.mdpi.com/2227-9717/13/9/2708>.
- [Liu et al., 2019] Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L. & Stoyanov, V.: , 2019; Roberta: A robustly optimized bert pretraining approach; *arXiv preprint arXiv:1907.11692*.
- [Löwenmark et al., 2025] Löwenmark, K.; Strömbergsson, D.; Liu, C.; Liwicki, M. & Sandin, F.: , 2025; Agent-based condition monitoring assistance with multimodal industrial database retrieval augmented generation; *arXiv preprint arXiv:2506.09247*.

- [Manning et al., 2008] Manning, C. D.; Raghavan, P. & Schütze, H.: , 2008; *Introduction to Information Retrieval*; Cambridge University Press.
- [Martins & Astudillo, 2016] Martins, A. F. T. & Astudillo, R. F.: , 2016; From softmax to sparsemax: A sparse model of attention and multi-label classification; en *Proceedings of the 33rd International Conference on Machine Learning (ICML)*; PMLR; págs. 1614–1623; URL <https://proceedings.mlr.press/v48/martins16.html>.
- [Mikolov et al., 2013] Mikolov, T. et al.: , 2013; Efficient estimation of word representations in vector space; *arXiv preprint arXiv:1301.3781*.
- [Mortensen, 2024] Mortensen, L.: , 2024; Data-driven proactive maintenance and asset management for energy distribution networks; doi:10.21996/ffxv-jx48.
- [Murphy, 2022] Murphy, K. P.: , 2022; *Probabilistic Machine Learning: An Introduction*; MIT Press; ISBN 978-0-262-04792-9.
- [Nachouki & Benbrahim, 2023] Nachouki, Y. & Benbrahim, H.: , 2023; Student performance prediction using random forest and explainable machine learning methods; *Education and Information Technologies*; **28** (5): 5669–5692; doi:10.1007/s10639-022-11569-4.
- [Ouyang et al., 2022] Ouyang, L. et al.: , 2022; Training language models to follow instructions with human feedback; *arXiv preprint arXiv:2203.02155*.
- [Papineni et al., 2002] Papineni, K.; Roukos, S.; Ward, T. & Zhu, W.-J.: , 2002; Bleu: a method for automatic evaluation of machine translation; en *ACL*.
- [Pennington et al., 2014] Pennington, J.; Socher, R. & Manning, C. D.: , 2014; Glove: Global vectors for word representation; en *EMNLP*.
- [Radford et al., 2018] Radford, A.; Narasimhan, K.; Salimans, T. & Sutskever, I.: , 2018; Improving language understanding by generative pre-training; *OpenAI*.
- [Raffel et al., 2020] Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W. & Liu, P. J.: , 2020; Exploring the limits of transfer learning with a unified text-to-text transformer; *Journal of Machine Learning Research*; **21** (140): 1–67.
- [Ranstam & Cook, 2018] Ranstam, J. & Cook, J. A.: , 2018; Lasso regression; *Journal of British Surgery*; **105** (10): 1348–1348; doi:10.1002/bjs.10895.
- [Rudin, 2019] Rudin, C.: , 2019; Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead; *Nature machine intelligence*; **1** (5): 206–215.
- [Saeed & Omlin, 2023] Saeed, W. & Omlin, C.: , 2023; Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities; *Knowledge-Based Systems*; **263**: 110273; doi:<https://doi.org/10.1016/j.knosys.2023.110273>; URL <https://www.sciencedirect.com/science/article/pii/S0950705123000230>.
- [Schulman et al., 2017] Schulman, J. et al.: , 2017; Proximal policy optimization algorithms; *arXiv preprint arXiv:1707.06347*.
- [Seppälä & Järventausta, 2024] Seppälä, J. & Järventausta, P.: , 2024; Analyzing supply reliability incentive in pricing regulation of electricity distribution operators; *Energies*; **17** (6); doi:10.3390/en17061451; URL <https://www.mdpi.com/1996-1073/17/6/1451>.
- [Shadi et al., 2025] Shadi, M. R.; Mirshekari, H. & Shaker, H. R.: , 2025; Explainable artificial intelligence for energy systems maintenance: A review on concepts, current techniques, challenges, and prospects; *Renewable and Sustainable Energy Reviews*; **216**: 115668; doi:<https://doi.org/10.1016/j.rser.2025.115668>; URL <https://www.sciencedirect.com/science/article/pii/S1364032125003417>.
- [Shalev-Shwartz & Ben-David, 2014] Shalev-Shwartz, S. & Ben-David, S.: , 2014; *Understanding machine learning: From theory to algorithms*; Cambridge university press.
- [Singh et al., 2025] Singh, A.; Ehtesham, A.; Kumar, S. & Khoei, T. T.: , 2025; Agentic retrieval-augmented generation: A survey on agentic rag; *arXiv preprint arXiv:2501.09136*.
- [Tang et al., 2019] Tang, Y.; Xu, Q. & Zhao, Y.: , 2019; Building a power equipment knowledge graph for intelligent maintenance; *arXiv preprint arXiv:1904.12242*; URL <https://arxiv.org/abs/1904.12242>.
- [Team, 2024] Team, N. R.: , 2024; Graphrag: Enhancing retrieval-augmented generation with knowledge graphs; <https://neo4j.com/blog/developer/graphrag-and-agentic-architecture-with-neoconverse/>.
- [Teixeira et al., 2025] Teixeira, B.; Carvalhais, L.; Pinto, T. & Vale, Z.: , 2025; Explainable ai framework for reliable and transparent automated energy management in buildings; *Energy and Buildings*; **347**: 116246; doi:<https://doi.org/10.1016/j.enbuild.2025.116246>; URL <https://www.sciencedirect.com/science/article/pii/S0378778825009764>.
- [Touvron et al., 2023] Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F. et al.: , 2023; Llama: Open and efficient foundation language models; *arXiv preprint arXiv:2302.13971*.
- [Trancasanchai, 2024] Trancasanchai, S.: , 2024; *Improving Question Answering Systems with Retrieval Augmented Generation*; Tesis Doctoral; University of Helsinki.
- [Troncia et al., 2023] Troncia, M.; Ruggeri, S.; Soma, G. G.; Pilo, F.; Ávila, J. P. C.; Muntoni, D. & Gianinoni, I. M.: , 2023; Strategic decision-making support for distribution system planning with flexibility alternatives; *Sustainable Energy, Grids and Networks*; **35**: 101138; doi:<https://doi.org/10.1016/j.segan.2023.101138>; URL <https://www.sciencedirect.com/science/article/pii/S2352467723001467>.

- [U.S. Commercial Service, 2021] **U.S. Commercial Service**: , 2021; Colombia retie technical standards; URL <https://www.trade.gov/market-intelligence/colombia-retie-technical-standards>; accessed: 2025-09-15.
- [U.S. Energy Information Administration, 2023a] **U.S. Energy Information Administration**: , 2023a; Annual electric power industry report: Reliability metrics (saidi, saifi); URL [https://www.eia.gov/electricity/annual/html/epa\\_11\\_01.html](https://www.eia.gov/electricity/annual/html/epa_11_01.html).
- [U.S. Energy Information Administration, 2023b] **U.S. Energy Information Administration**: , 2023b; Annual electric power industry report: Reliability metrics (saifi, saidi); URL [https://www.eia.gov/electricity/annual/html/epa\\_11\\_01.html](https://www.eia.gov/electricity/annual/html/epa_11_01.html).
- [Uyar & Albayrak, 2025] **Uyar, A. & Albayrak, S.**: , 2025; Interpretable gradient boosting machines for regression and classification; *Applied Intelligence*; **55** (1): 432–451; doi:10.1007/s10489-024-05442-7.
- [Vaswani et al., 2017] **Vaswani, A. et al.**: , 2017; Attention is all you need; *Advances in Neural Information Processing Systems (NeurIPS)*.
- [Wang et al., 2025] **Wang, D.; Maharjan, S.; Zheng, J.; Liu, L. & Wang, Z.**: , 2025; Data-driven quantification and visualization of resilience metrics of power distribution system; *arXiv preprint arXiv:2508.12408*.
- [Zhan et al., 2024] **Zhan, J.; Wu, C.; Yang, C.; Miao, Q. & Ma, X.**: , 2024; Hfn: Heterogeneous feature network for multivariate time series anomaly detection; *Information Sciences*; **670**: 120626.
- [Zhang et al., 2023] **Zhang, K.; Liu, J. & Huang, R.**: , 2023; Rule-enhanced cognitive graph for power grid knowledge reasoning; en *International Conference on Knowledge Graph and Semantic Computing*; Springer; págs. 701–714; doi: 10.1007/978-981-99-4761-4\_59.
- [Zhang et al., 2020] **Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q. & Artzi, Y.**: , 2020; Bertscore: Evaluating text generation with bert; en *ICLR*.
- [Zhou et al., 2024] **Zhou, Z.; Li, Y.; Guo, Z.; Yan, Z. & Chow, M.-Y.**: , 2024; A white-box deep-learning method for electrical energy system modeling based on kolmogorov-arnold network; *arXiv preprint arXiv:2409.08044*.
- [Zhu et al., 2021] **Zhu, S.; Yao, R.; Xie, Y.; Qiu, F.; Wu, X. et al.**: , 2021; Quantifying grid resilience against extreme weather using large-scale customer power outage data; *arXiv preprint arXiv:2109.09711*.