6th International Conference on Electrical, Computer and Energy Technologies (ICECET 2026)
6–9 July 2026, Rome, Italy

Bridging Raw Sensor Data and BI Systems: An AI-Driven Metadata-Aware Mapping Approach

1st Konstantinos Kaimakis
LIBRA AI Technologies
Athens, Greece
konstantinos.kaimakis@libramli.ai

2nd Stelios Georgaras
LIBRA AI Technologies
Athens, Greece
stelios.georgaras@libramli.ai

3rd Eleanna Kouletou
LIBRA AI Technologies
Athens, Greece
eleanna.kouletou@libramli.ai

4th Zisis Flokas
LIBRA AI Technologies
Athens, Greece
zisis.flokas@libramli.ai

5th Georgios Siokas
LIBRA AI Technologies
Athens, Greece
george.siokas@libramli.ai

6th Athanasios Balomenos
LIBRA AI Technologies
Athens, Greece
athanasios.balomenos@libramli.ai

7th Yannis Kopsinis
LIBRA AI Technologies
Athens, Greece
yannis.kopsinis@libramli.ai

***Abstract*—This paper presents an AI-driven, metadata-aware framework for automating the integration of heterogeneous sensor time-series into Business Intelligence (BI) systems. The method is industry-agnostic, fusing time-series representations with semantic cues from metadata (e.g., column names). Concretely, time-series features from ROCKET and Chronos-T5 are combined with MiniLM-based sentence embeddings of metadata and classified with gradient-boosting models to predict sensor type–unit pairs across 40 distinct classes. On a diverse, multi-source dataset, the approach achieves 86.4% Top-1 accuracy, demonstrating the benefit of jointly leveraging time-series and semantic patterns. To illustrate practical use, a Streamlit application was developed and deployed in a mining case study, showcasing its usability in the field and enabling scalable, expert-assisted mapping pipelines for heterogeneous sensor data.**

***Keywords*—sensor data integration, time-series classification, human-in-the-loop, Business Intelligence**

I. Introduction

The growing adoption of data-centric strategies across industrial domains has intensified the need to integrate heterogeneous sensor data into Business Intelligence (BI) and Decision Support Systems (DSS) [1]. Time-series data from sensors and Industrial Internet of Things (IIoT) devices are essential to such systems but remain difficult to integrate due to structural variability and a lack of semantic standardisation [2], [3]. Traditional data mapping approaches rely on manual effort or static rules that are hard to maintain and scale in dynamic environments.

Prior research on data integration has mainly focused on schema matching and tabular alignment. The Valentine framework [4] exemplifies this direction, benchmarking algorithms across structured datasets. However, such methods target static tables and are not suited to time-series data that combine temporal dynamics with heterogeneous, non-standard schemas. Advances in artificial intelligence provide alternative approaches for automating data preparation. Time-series classification extracts meaningful temporal patterns for tasks such as monitoring and predictive maintenance [5], while metadata-aware learning leverages auxiliary information, e.g., column names and measurement units, to infer semantic context [6]. Frameworks such as stEELlm [7] highlight the potential of language models for semantic annotation but remain limited to static tabular settings and struggle to generalise to raw sensor streams.

To address this gap, an AI-driven, metadata-aware mapping framework is proposed that fuses temporal feature extraction with semantic embeddings for sensor data. The model combines univariate time-series representations with natural-language embeddings of metadata to infer both sensor type and measurement unit. It supports operation with or without metadata and includes a human-in-the-loop validation step to ensure practical usability. A prototype implementation, developed in Streamlit [8], enables interactive validation and integration of raw sensor data into BI and DSS platforms. Although domain-agnostic, the framework has been validated on mining datasets, demonstrating its applicability in complex industrial environments.

II. Methodology

A. Dataset

Publicly available industrial and environmental sensor datasets were used together with a private dataset from the mining industry [9]–[16]. Each dataset contains multivariate time-series with varying sampling rates and inconsistent metadata. To ensure comparability, all data were resampled at an hourly frequency, harmonised to common units (e.g., temperature in °C, pressure in mbar), and segmented into fixed-length univariate windows of 100 samples. For each segment, the associated column name was preserved as optional metadata. A multilingual dictionary of synonymous and abbreviated column names was compiled to increase

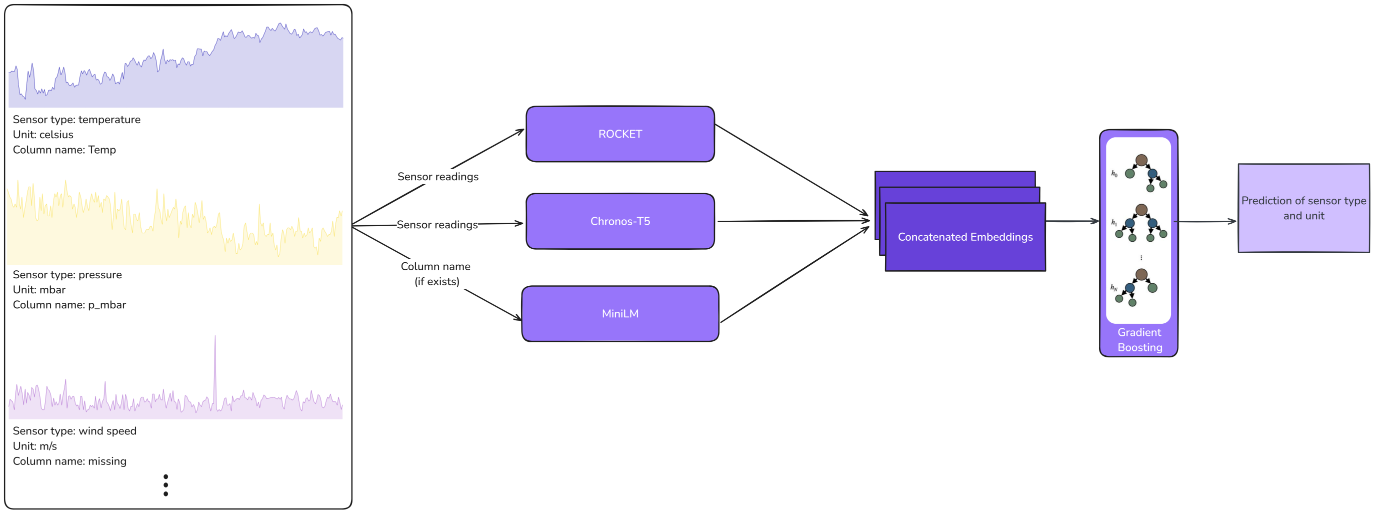

Fig. 1. Model architecture. Embeddings are generated by time-series encoders from sensor readings and by a language model from column names. The concatenated embeddings are used to train a gradient-boosting classifier to predict sensor type and units.

robustness to naming variation. The resulting balanced dataset covers a diverse set of sensor types and units representative of typical IIoT deployments. The classification task is defined over a Cartesian label space constructed from the combination of sensor types and their corresponding measurement units; each class represents a unique $(\text{sensor type}, \text{unit})$ pair, such as (temperature, °C) or (pressure, mbar). In total, the dataset covers more than 20 sensor types and over 40 valid combinations, reflecting the heterogeneity typically encountered in industrial IIoT environments (see Table I).

B. Model Architecture

The proposed framework aims to automate the semantic integration of heterogeneous sensor time-series data by combining temporal feature extraction with metadata-aware semantic embeddings. The overall workflow (Fig. 1) encodes both numerical and textual information into a unified representation space, where each signal is jointly characterised by its temporal dynamics and contextual metadata. These representations are then used to train supervised classifiers capable of predicting the sensor type and the measurement unit. **Feature extraction:** To capture time-series patterns, two complementary encoders were employed: ROCKET [17] and Chronos-T5 [18]. ROCKET (Random Convolutional Kernel Transform) provides a lightweight yet expressive transformation of time series into high-dimensional feature vectors by applying a large number of random convolutional kernels and computing simple summary statistics. It is particularly effective for fast, task-agnostic time-series classification. Chronos-T5 extends the transformer architecture to time-series representation learning. Building on the Text-to-Text Transfer Transformer (T5) model, it leverages masked-reconstruction pretraining to learn contextual dependencies and long-term temporal structure. In this framework, Chronos-T5 embeddings complement the handcrafted statistical features of ROCKET, capturing richer sequence-level semantics.

TABLE I
SENSOR TYPES AND CORRESPONDING MEASUREMENT UNITS SUPPORTED BY THE CLASSIFICATION FRAMEWORK.

Sensor Type	**Measurement Units**
Temperature	°C, °F, K
Relative Humidity	%
UV Index	a.u.
Visibility	m
Wind Direction	degrees
Wind Gust	m/s
Wind Speed	m/s, km/h, knots
Cloudiness	%
Dew Point	°C
Feels Like	°C
Flow Rate	L/s, m^3/h
Pressure	atm, bar, mbar
Solar Radiation	W/m^2
Rainfall Rate	mm/h
Snowfall Rate	mm/h
CO Concentration	$\mu g/m^3$, ppm, ppb
SO_2 Concentration	$\mu g/m^3$, ppm, ppb
NO Concentration	$\mu g/m^3$, ppm, ppb
NO_2 Concentration	$\mu g/m^3$, ppm, ppb
O_3 Concentration	$\mu g/m^3$, ppm, ppb
NH_3 Concentration	$\mu g/m^3$
$PM_{2.5}$	$\mu g/m^3$
PM_{10}	$\mu g/m^3$
Displacement	cm/m, mm/m
Electrical Conductance	mS
Evapotranspiration Rate	mm/h
Height	cm, mm, m
Altitude	masl
Relative Water Level Elevation (CA)	m

Metadata embeddings: When available, column names and other textual metadata provide valuable contextual cues. These were embedded using a multilingual MiniLM-based Sentence Transformer [19], which generates dense vector representations that preserve semantic similarity across abbreviations and

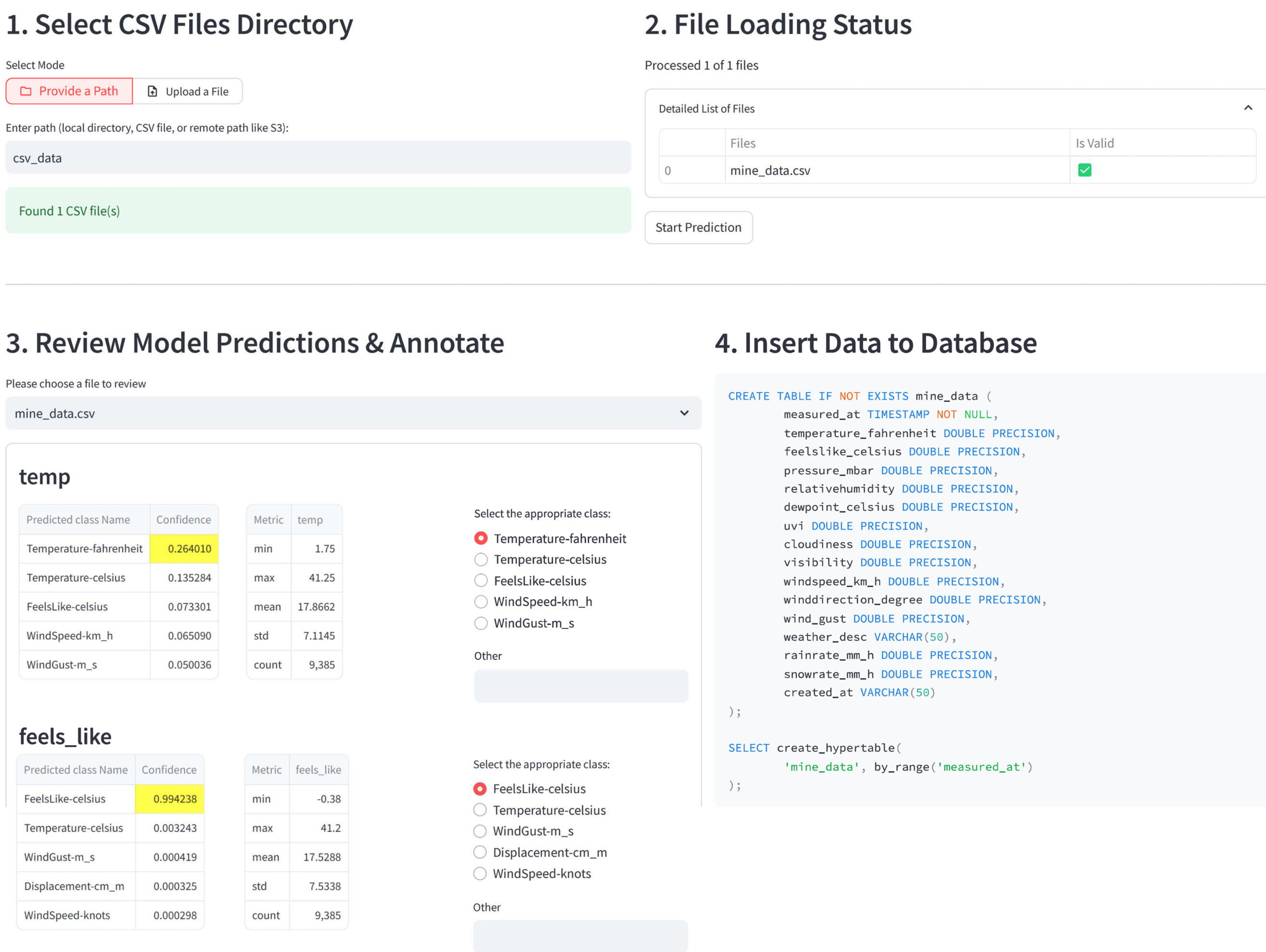

Fig. 2. User interface of the AI-assisted sensor data mapping tool. The Streamlit-based application allows experts to upload sensor files, review model predictions, validate mappings, and insert verified results into a central database, enabling human-in-the-loop integration within BI pipelines.

multilingual expressions. Using a distilled transformer model ensures computational efficiency while maintaining semantic expressiveness. When metadata are absent, the framework relies solely on the temporal features described above.

Feature fusion and classification: The temporal and semantic embeddings are concatenated into a unified feature vector representing each sensor signal. This fused representation is then used for supervised classification, where the goal is to infer both the sensor type and its associated measurement unit. Gradient-boosting classifiers were selected for this task owing to their robustness and high performance on structured, heterogeneous data. Specifically, XGBoost [20] and LightGBM [21] were adopted. Both algorithms efficiently handle high-dimensional inputs, support parallel computation, and offer strong regularisation capabilities. While XGBoost is known for its stability and accuracy across a range of datasets, LightGBM provides improved scalability and faster training via histogram-based and leaf-wise growth strategies. Combining the two models enables a comparative evaluation of predictive performance and computational efficiency.

This hybrid approach enables the model to operate under varying metadata-availability conditions, supporting both fully annotated and metadata-sparse datasets. By jointly learning from signal dynamics and semantic context, the framework provides a scalable and domain-agnostic solution for automated sensor data mapping.

C. Experimental Setup

All experiments were implemented in Python using PyTorch, Scikit-learn, and the Hugging Face Transformers library. Training and inference were executed on a workstation equipped with an NVIDIA RTX 3090 GPU. Hyperparameters for the gradient-boosting models (learning rate, tree depth, and number of estimators) were tuned via grid search on the validation set. The Chronos-T5 encoder was used in its *tiny* configuration for efficiency, and experiments were tracked using MLflow. Evaluation metrics (Top-1 accuracy, Top-5 accuracy, and F1-score) were averaged across three random seeds to ensure stability. A stratified split was applied to create disjoint training, validation, and test sets; segments from the same original data source never appeared in multiple splits, preventing data leakage.

D. UI/UX Application

A user-oriented application was developed to support sensor data mapping and integration in mining operations (Fig. 2). The tool combines automated time-series classification with expert validation, addressing challenges related to inconsistent or missing metadata in heterogeneous sensor datasets and improving the efficiency of data preparation for BI and decision-support systems.

The application was implemented as a web-based interface using Streamlit. Users can upload sensor datasets in CSV format, after which the system automatically detects the time column, resamples high-frequency signals to an hourly resolution, and separates numerical from categorical variables. Numerical signals are processed by the proposed classification framework, which outputs the top-five most probable sensor-type and measurement-unit predictions. Categorical variables, such as equipment identifiers or locations, are identified and made available for optional renaming or exclusion.

Depending on metadata availability, the system applies either a hybrid configuration combining ROCKET, Chronos-T5, and column-name embeddings, or a fallback configuration relying solely on time-series features. Inference is parallelised to ensure responsiveness, and validated results are stored in a backend database for traceability. The application was evaluated on mining datasets involving environmental and operational sensors, demonstrating its practical applicability in real industrial settings.

III. Results and Discussion

A. Quantitative Results

Table II reports the performance of the proposed framework under different feature combinations and classifiers. Evaluation was based on Top-1 and Top-5 accuracy as well as F1-score across all sensor-type and measurement-unit classes. Overall, combining the features extracted by ROCKET and Chronos-T5 with metadata embeddings significantly improved predictive performance. The best configuration achieved a Top-1 accuracy of **0.864** and an F1-score of **0.865** using LightGBM with fused temporal and textual features. When metadata were omitted, performance decreased by approximately two percentage points, confirming that column names provide meaningful semantic cues for sensor identification. XGBoost and LightGBM exhibited comparable behaviour, with LightGBM offering slightly higher efficiency during training. The consistently high Top-5 accuracy values (> 0.97) demonstrate that the model's most probable predictions almost always include the correct label.

The results confirm that combining metadata-based semantic embeddings with temporal representations improves sensor classification. When column names are available, the model leverages contextual cues about measurement type and unit to disambiguate signals with similar temporal patterns. The joint use of ROCKET and Chronos-T5 proved complementary: ROCKET captured local variations while Chronos-T5 modelled long-term dependencies, yielding more stable and generalisable performance. A slight drop in Top-5 accuracy, linked to a higher confidence concentration on the top prediction, aligns with known calibration effects in multi-class classifiers [22]. In practice, this trade-off is beneficial: more decisive Top-1 predictions enhance automated mapping with optional human validation.

B. Application Case Study: Mining Sensor Data Mapping

The proposed framework was evaluated on real-world sensor data from a large-scale mining operation that operates a monitoring infrastructure across its production and environmental systems. The deployed sensors capture a variety of operational and environmental measurements, including flow rate, electrical conductance, water level, pressure, and displacement. The collected datasets reflect realistic industrial conditions, exhibiting heterogeneous sensor configurations, irregular sampling rates, inconsistent and multilingual column-naming conventions, and partial or missing metadata.

During model development, mining-related sensor data were incorporated into training in order to improve the representation of domain-specific temporal patterns. Evaluation was subsequently performed on temporally disjoint datasets collected from the same industrial environment but not observed during training. While the sensor types were known to the model, the evaluation data differed in signal realisations, naming conventions, units, and acquisition conditions, thereby testing the model's ability to correctly classify unseen instances.

Table III presents representative Top-5 prediction outputs for selected sensor variables from these evaluation datasets. Highly distinctive signals, such as flow-rate measurements, yield predictions with near-perfect confidence and can be reliably accepted with minimal expert intervention. In contrast, variables with overlapping temporal characteristics, such as electrical conductance and relative water-level elevation, produce multiple plausible class candidates with lower confidence scores. In such cases, human-in-the-loop validation is essential to resolve ambiguity and ensure correct semantic mapping. Overall, this case study demonstrates the practical applicability of the proposed framework in realistic industrial settings, where the primary challenge lies in the robust classification of heterogeneous and imperfectly annotated sensor data. By combining automated inference with confidence-aware expert validation, the approach significantly reduces manual integration effort while retaining control over final labelling decisions.

IV. Conclusion

This paper introduced an AI-driven, metadata-aware framework for automating the semantic mapping of heterogeneous sensor time-series data into Business Intelligence systems. By combining temporal representations extracted from raw sensor signals with semantic embeddings derived from column-level

TABLE II
PERFORMANCE OF TESTED CONFIGURATIONS ON THE EVALUATION SET.

Time-series encoder	Classifier	Column names	Top-1 acc	F1	Top-5 acc
ROCKET	XGBoost	No	0.816	0.809	0.977
ROCKET	XGBoost	Yes	0.833	0.832	0.973
ROCKET	LightGBM	No	0.811	0.804	0.982
ROCKET	LightGBM	Yes	0.841	0.840	0.971
ROCKET + Chronos-T5	XGBoost	No	0.858	0.850	0.984
ROCKET + Chronos-T5	XGBoost	Yes	0.863	0.863	0.980
ROCKET + Chronos-T5	LightGBM	No	0.854	0.846	**0.993**
ROCKET + Chronos-T5	LightGBM	Yes	**0.864**	**0.865**	0.985

TABLE III
ROW-WISE TOP-5 PREDICTIONS FOR REPRESENTATIVE SENSOR VARIABLES FROM UNSEEN THARSIS MINE DATASETS. THE MOST PROBABLE CLASS (TOP-1) IS HIGHLIGHTED IN BOLD, ILLUSTRATING CONFIDENCE-DRIVEN HUMAN-IN-THE-LOOP VALIDATION.

Sensor Column	Predicted Class	Confidence
Flow_ls	**Flow–L/s**	**0.99982**
	Wind Speed–m/s	0.00043
	Relative Water Level Elevation (CA)–m	0.00042
	Electrical Conductance–mS	0.00012
	Displacement–mm/m	0.00011
ElecCondMeasure	**Electrical Conductance–mS**	**0.79223**
	Relative Water Level Elevation (CA)–m	0.07804
	Wind Gust–m/s	0.01232
	Rainfall Rate–mm/h	0.00728
	Ozone–ppb	0.00665
Altura	**Height–m**	**0.91202**
	Displacement–cm/m	0.05502
	Height–mm	0.00983
	Height–cm	0.00363
	Rainfall Rate–mm/h	0.00241
Norm_elev	**Relative Water Level Elevation (CA)–m**	**0.4581**
	WindSpeed–km/h	0.1055
	Pressure–atm	0.0572
	RainRate–mm/h	0.0351
	Pressure–bar	0.0275

metadata, the proposed approach effectively addresses common challenges in industrial data integration, including inconsistent schemas, diverse naming conventions, and incomplete annotations.

The framework demonstrated strong performance across multi-source datasets and was validated in a real-world mining case study, where it successfully generalised to previously unseen sensor data under realistic operating conditions. The integration of automated inference with expert validation supports reliable deployment in industrial environments, enabling scalable and consistent sensor data integration while preserving human oversight.

ACKNOWLEDGMENT

This research was financially supported by the European Union's Horizon Europe research and innovation programme under the topic HORIZON-CL4-2022-RESILIENCE-01-06, called MASTERMINE, with grant agreement number 101091895.

REFERENCES

[1] A. Cuzzocrea, I.-Y. Song, and K. C. Davis, "Analytics over large-scale multidimensional data: the big data revolution!," in *Proc. ACM 14th Int. Workshop Data Warehousing and OLAP*, New York, NY, USA, 2011, pp. 101–104.

[2] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM Comput. Surv.*, vol. 46, no. 4, art. 44, 2014.

[3] R. K. Verma, K. K. Pattanaik, S. Bharti, and D. Saxena, "In-network context inference in IoT sensory environment for efficient network resource utilization," *J. Netw. Comput. Appl.*, vol. 130, pp. 89–103, 2019.

[4] C. Koutras, G. Siachamis, A. Ionescu, K. Psarakis, J. Brons, M. Fragkoulis, C. Lofi, A. Bonifati, and A. Katsifodimos, "Valentine: Evaluating matching techniques for dataset discovery," in *Proc. 2021 IEEE 37th Int. Conf. Data Eng. (ICDE)*, 2021, pp. 468–479.

[5] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Min. Knowl. Discov.*, vol. 33, no. 4, pp. 917–963, 2019.

[6] P. Ristoski and H. Paulheim, "Semantic Web in data mining and knowledge discovery: a comprehensive survey," *J. Web Semant.*, vol. 36, pp. 1–22, 2016.

[7] M. Cremaschi, F. D'Adda, and A. Maurino, "stEELlm: An LLM for generating semantic annotations of tabular data," *ACM Trans. Intell. Syst. Technol.*, 2025.

[8] Snowflake Inc., "Streamlit documentation," 2025. [Online]. Available: https://docs.streamlit.io/

[9] City of Chicago, "Beach weather stations (automated sensors)." Accessed: Sep. 2, 2024. [Online]. Available: https://data.cityofchicago.org/Parks-Recreation/Beach-Weather-Stations-Automated-Sensors/k7hf-8y75/about_data

[10] S. Feinberg, R. Williams, G. S. W. Hagler, J. Rickard, R. Brown, D. Garver, G. Harshfield, P. Stauffer, E. Mattson, R. Judge, and S. Garvey, "Long-term evaluation of air sensor technology under ambient conditions in Denver, Colorado," *Atmos. Meas. Tech.*, vol. 11, no. 8, pp. 4605–4615, 2018.

[11] Max Planck Institute for Biogeochemistry, "Weather data from the Max Planck Institute for Biogeochemistry," 2025. [Online]. Available: https://www.bgc-jena.mpg.de/wetter/

[12] R. M. Duvall, R. W. Long, M. R. Beaver, K. G. Kronmiller, M. L. Wheeler, and J. J. Szykman, "Performance evaluation and community application of low-cost sensors for ozone and nitrogen dioxide," *Sensors*, vol. 16, no. 10, art. 1698, 2016.

[13] S. Kimbrough, S. Krabbe, R. Baldauf, T. Barzyk, M. Brown, S. Brown, C. Croghan, M. Davis, P. Deshmukh, R. Duvall, S. Feinberg, V. Isakov, R. Logan, T. McArthur, and A. Shields, "The Kansas City Transportation and Local-Scale Air Quality Study (KC-TRAQS): integration of low-cost sensors and reference grade monitoring in a complex metropolitan area. Part 1: Overview of the project," *Chemosensors*, vol. 7, no. 2, art. 26, 2019.

[14] City of Chicago, "Smart green infrastructure monitoring sensors—historical," 2018. [Online]. Available: https://data.cityofchicago.org/Environment-Sustainable-Development/

Smart-Green-Infrastructure-Monitoring-Sensors-Hist/ggws-77ih/about_data
[15] OpenWeather, "OpenWeatherMap API," 2025. [Online]. Available: https://openweathermap.org/api
[16] U.S. Environmental Protection Agency, 2025. [Online]. Available: https://www.epa.gov/
[17] A. Dempster, F. Petitjean, and G. I. Webb, "ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels," *Data Min. Knowl. Discov.*, vol. 34, no. 5, pp. 1454–1495, 2020.
[18] A. F. Ansari, L. Stella, C. Turkmen, X. Zhang, P. Mercado, H. Shen, O. Shchur, S. S. Rangapuram, S. P. Arango, S. Kapoor, J. Zschiegner, D. C. Maddix, H. Wang, M. W. Mahoney, K. Torkkola, A. G. Wilson, M. Bohlke-Schneider, and Y. Wang, "Chronos: Learning the language of time series," 2024.
[19] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou, "MiniLM: deep self-attention distillation for task-agnostic compression of pre-trained transformers," 2020.
[20] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, New York, NY, USA, 2016, pp. 785–794.
[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: a highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
[22] A. Le Coz, S. Herbin, and F. Adjed, "Confidence calibration of classifiers with many classes," in *Proc. 38th Annu. Conf. Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=ebBnKVxMcZ