AIR QUALITY MONITOING

Validation of air quality data with AI

Jérôme LOUAT

- ENVEA overview
- Introduction to data validation
- Technical choices
- Data used for training
- Performance
- Conclusion

ENVEA overview

GLOBAL PRESENCE

Headquarters
Poissy, France

ENVEA has established a global presence, widely recognized as reference for ambient air monitoring with systems installed in 110+ countries and regions, including Beijing, Istanbul, Paris, Mecca, New Delhi, etc.

Overall, our equipment is compliant with 50 regulations across the world, with constant discussions with local authorities to ensure future compliance and drive innovation

With 10 local entities, ENVEA sells its systems worldwide:

SOLUTIONS

DIGITAL SOLUTIONS

INFORM

Gas monitors

Particulate monitors

Atmospheric monitoring & research

Miniature stations

MINIMIZE

Fixed Gas monitors

Particulate monitors

Portable Gas monitors

Mercury monitors

PREVENT

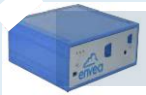
Gas monitor

Flow measurement

Moisture measurement

Silo protection

Dust measurement



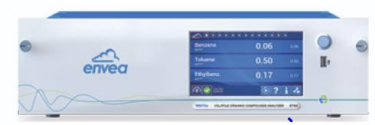
Leakage

ENVIRONMENTAL AIR MONITORING

MONITORING SOLUTIONS TO UNDERSTAND ATMOSPHERIC POLLUTION ORIGINS AND MONITOR AIR QUALITY FOR ENVIRONMENTAL AND PUBLIC HEALTH

- MICRO-SENSORS
- STANDART GAS & PARTICULATE MONITORS
- EXPERT INSTRUMENTS
- ATMOSPHERIC INSTRUMENTS
- DATA MANAGEMENT & REPORTING

- **40k+ air quality monitors** in major cities: Beijing, Istanbul, Paris, Mecca, New Delhi
- New deployments driven by countries upgrading to comply with **new regulations**
- **Strong presence** in both mature markets and leader in fast-growing emerging markets



Data validation

introduction

Introduction on data validation

- **Comprehensive Dual Validation**:
  - Rigorous two-tier validation process ensures relevance and compliance with environmental standards.
  - Technical validation & weekly environmental validation

- **Specialized Expert Teams**: Distinct groups of experts handle each validation phase.

- **Guaranteeing Data Accuracy for Regulatory Bodies**:
  - To certify the exactness of the data

- **Commitment to In-Depth Analysis**:
  - The validation process ensures that all data meets the highest standards of quality and reliability before dissemination.

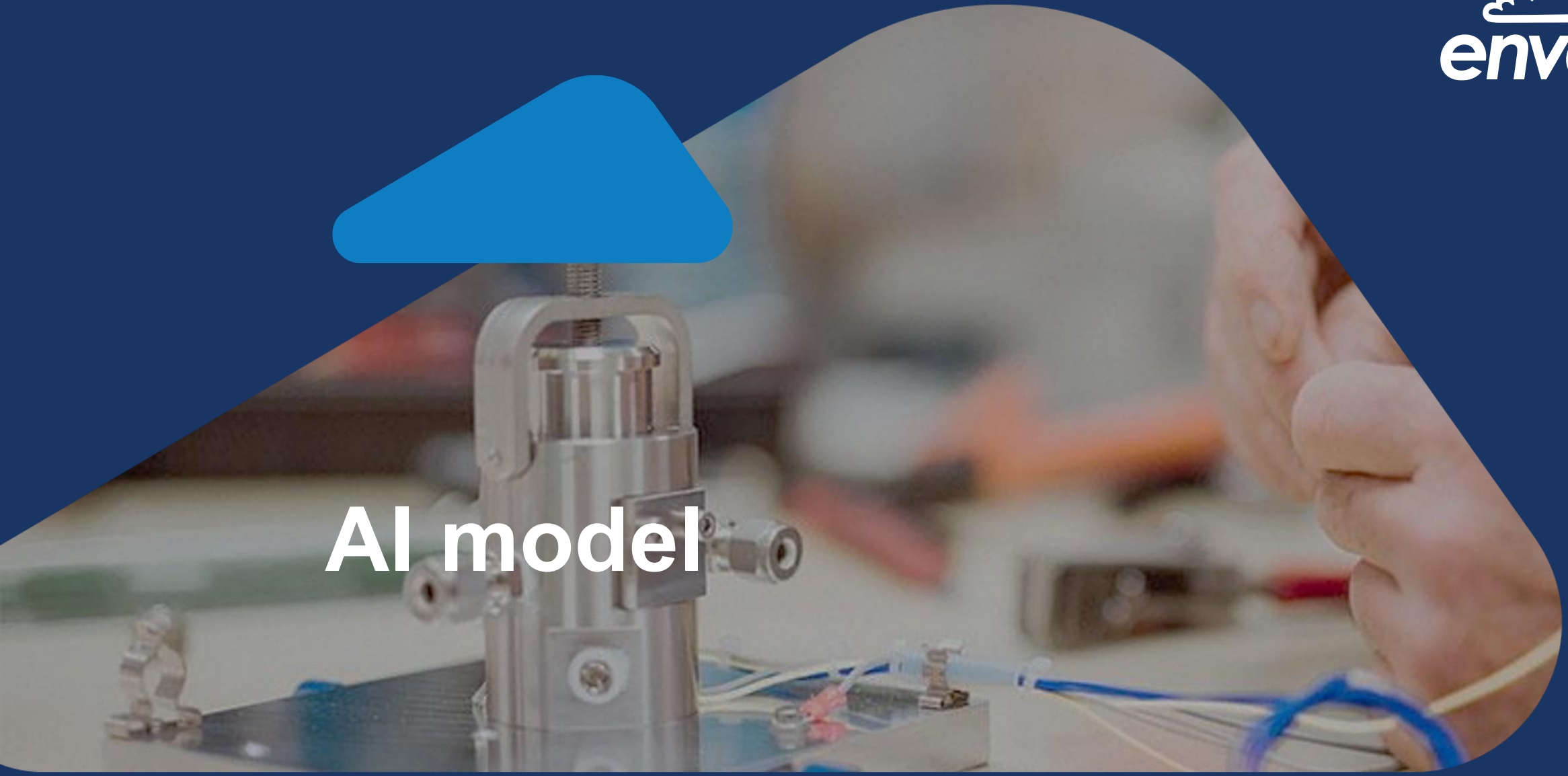
Air quality data : brief history

- **Beginning of data management at scale in the 1990's and early 2000's**
  - Iseo released the first version of XR in 1995, acquired by ENVEA in 2007

- **Up to the 2010's:**
  - Primary focus on metadata collect and flagging interface to support air quality engineer work

- **2015: release of first version of ADVAL by ENVEA, first module to provide automatic pre-validation tool to air quality experts**
  - Operator work efficiency
  - Reduced risk linked to online real time data publication

- **2025: development of ADVAL next gen through an AI based pre validation module**



AI model

Objective of implementing AI tool

- **Automated Validation of Air Quality Data**:
  - Automatically validates **90% of air quality data**
  - Significantly increases efficiency
  - Reduces the potential for human error

- **Reduced Manual Verification**: Only **10% of the data requires manual review**
  - Allows experts to focus on critical issues.

- **Comprehensive Pollutant Coverage**: The tool includes validation for key pollutants
  - **O_3, NO_2, PM_{10}, $PM_{2.5}$, CO, SO_2**

- **Continuous Improvement and Seamless Integration**:
  - **XR interface** now more intuitive and efficient user experience.

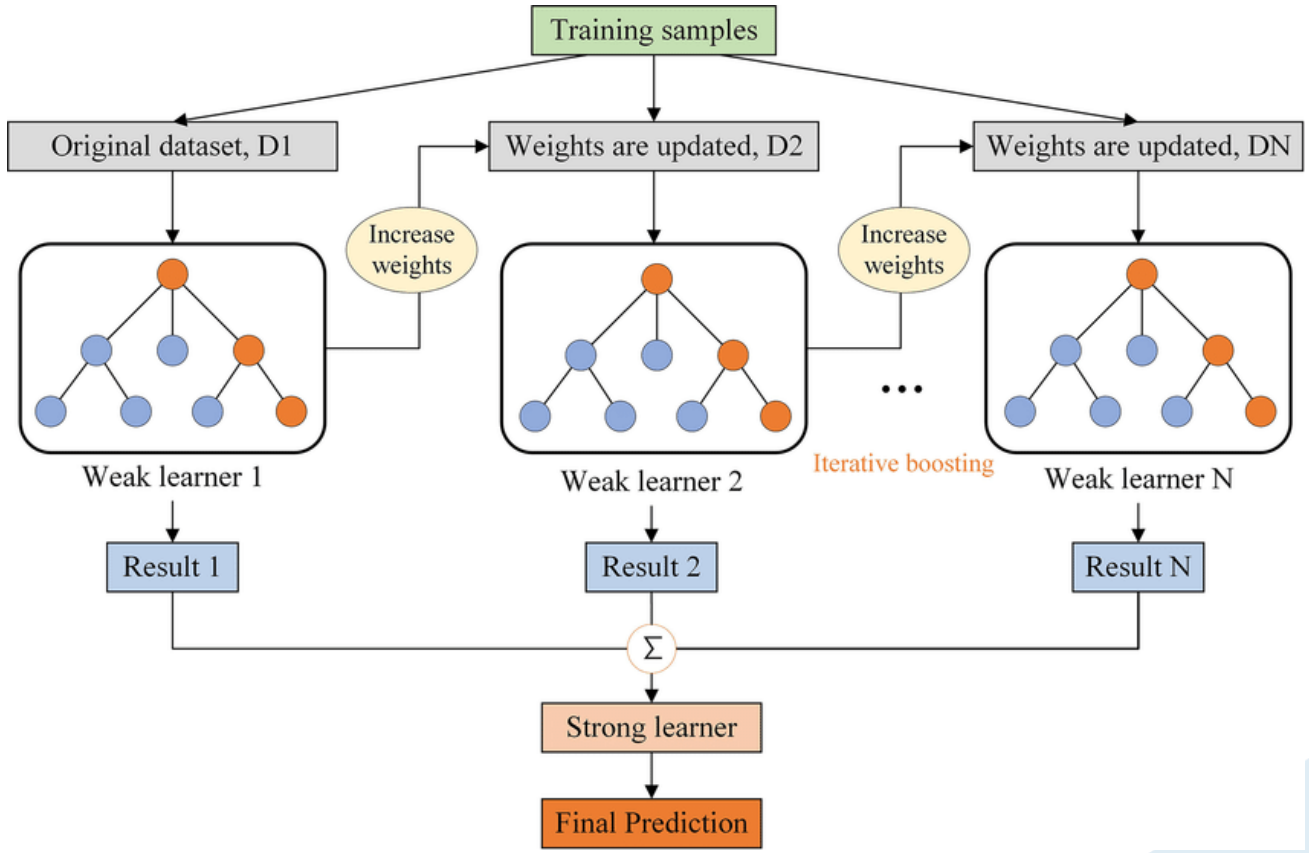
Application objectives

- **Provide Clear Visual Aids**:
  - Allows users to quickly assess and interpret information immediately

- **Continuous and Near-Instantaneous Operation**:
  - **Just one hour after data collection**
  - Enables timely responses and decision-making in air quality monitoring

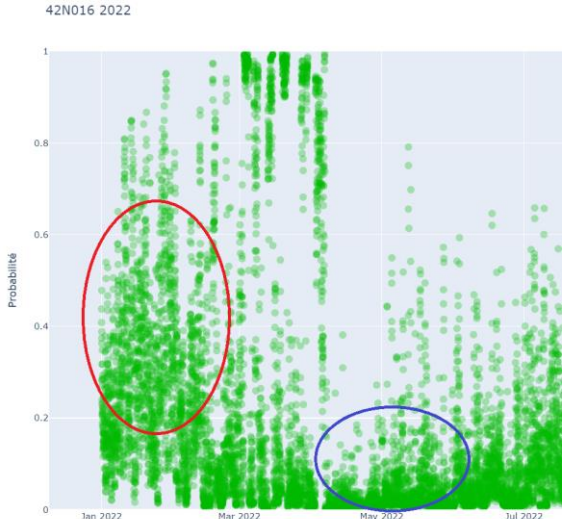
- **Enhance User Engagement Through Intuitive Interface**:
  - Facilitates easy navigation
  - Data validation process efficient and accessible

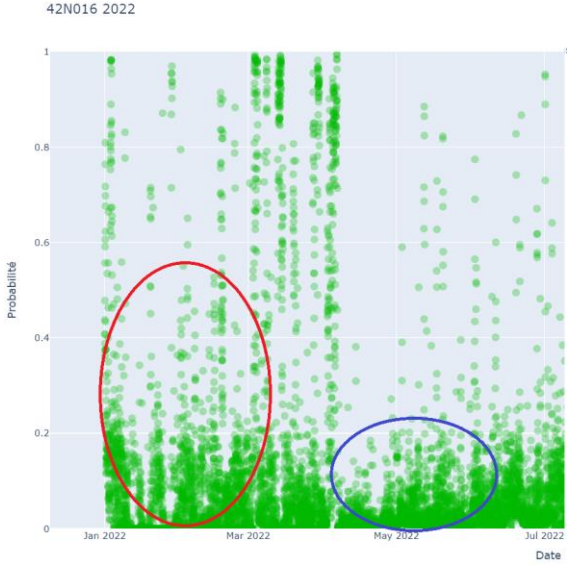
Model generalities



- **Combination of Decision Tree Algorithms**: Based on the open-source **LightGBM** library developed by Microsoft.
- **Fine-Tuned Parameters for Specialization**: Parameters are meticulously adjusted to specialize in the validation of **atmospheric pollutant concentration data**, improving the model's precision in this specific domain.
- **First Algorithm—Establishing General Invalidation Rules**
- **Second Algorithm—Site-Specific Fine-Tuning**
- **Third Algorithm—Context-Specific Activation**

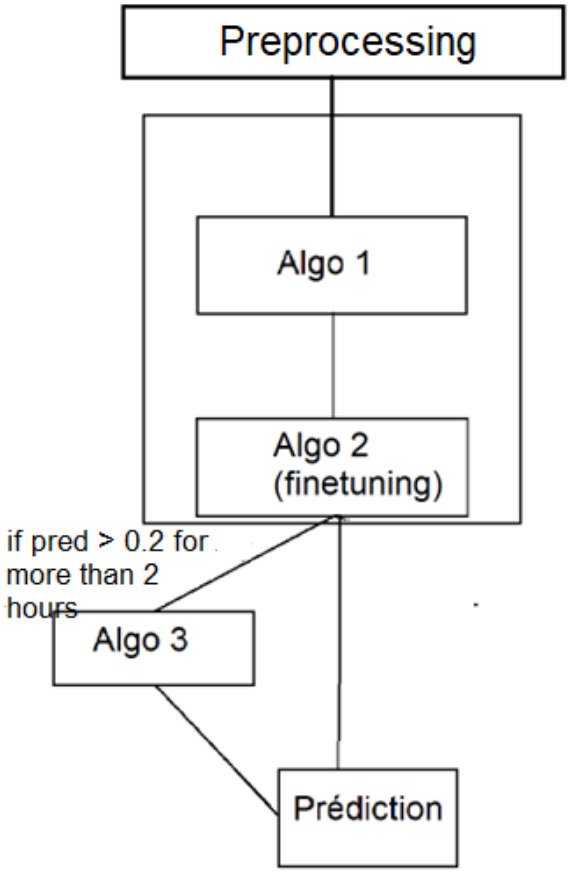
Model : long term invalidity case





- **Leverages Predictions from Previous Models with Targeted Variables**:
  - Focuses on a specific selection of variables
  - Designed to identify instances where forecasts remain elevated over an extended period.
- **Addresses Prolonged High Prediction Scenarios**:
  - Illustrates the invalidity probabilities before applying the third algorithm.
- **Objective to Eliminate Persistent High Predictions Without Reducing Invalid Detection Rate**: The goal of this model is to ensure both accuracy and reliability.
- **Activated Under Specific Conditions**:
  - Specifically addresses sustained anomalies.

Data comparison



Preprocessing:

- Pollutant Concentration Retrieval
- Calculation of Historical Site Averages
- Normalization Procedures
- Meteorological Data Integration
- Nearby Site Data Collection
- Moving Variable Computations
- Correlation Coefficients Between Pollutants

Data processing

Model Processing Workflow

- First Model Application
- Second Model Activation
- Final Prediction Output

Training data

Set of data

- **Extensive Network of Stations**
- **Vast Historical Dataset**:
  - **300 million data points**
- **Satellite Meteorological Data**:
  - Temperature, wind speed, humidity, & atmospheric pressure
- **Detailed Geographical and Topographical Data**:
  - Background sites, rural areas, traffic density, and altitude

Performances

4 metrics are used to evaluate the model:

1. **Rate of Isolated Invalid Data Identified:**
   - Corresponds to one-off errors

2. **Rate of Invalid Periods Detected:**
   - Assesses the model's effectiveness in identifying data sequences

3. **Rate of Data with High Probability of Error:**
   - Quantifies the percentage of data that suggests a need for detailed human examination.
   - This enables experts to target data for analysis, optimizing time

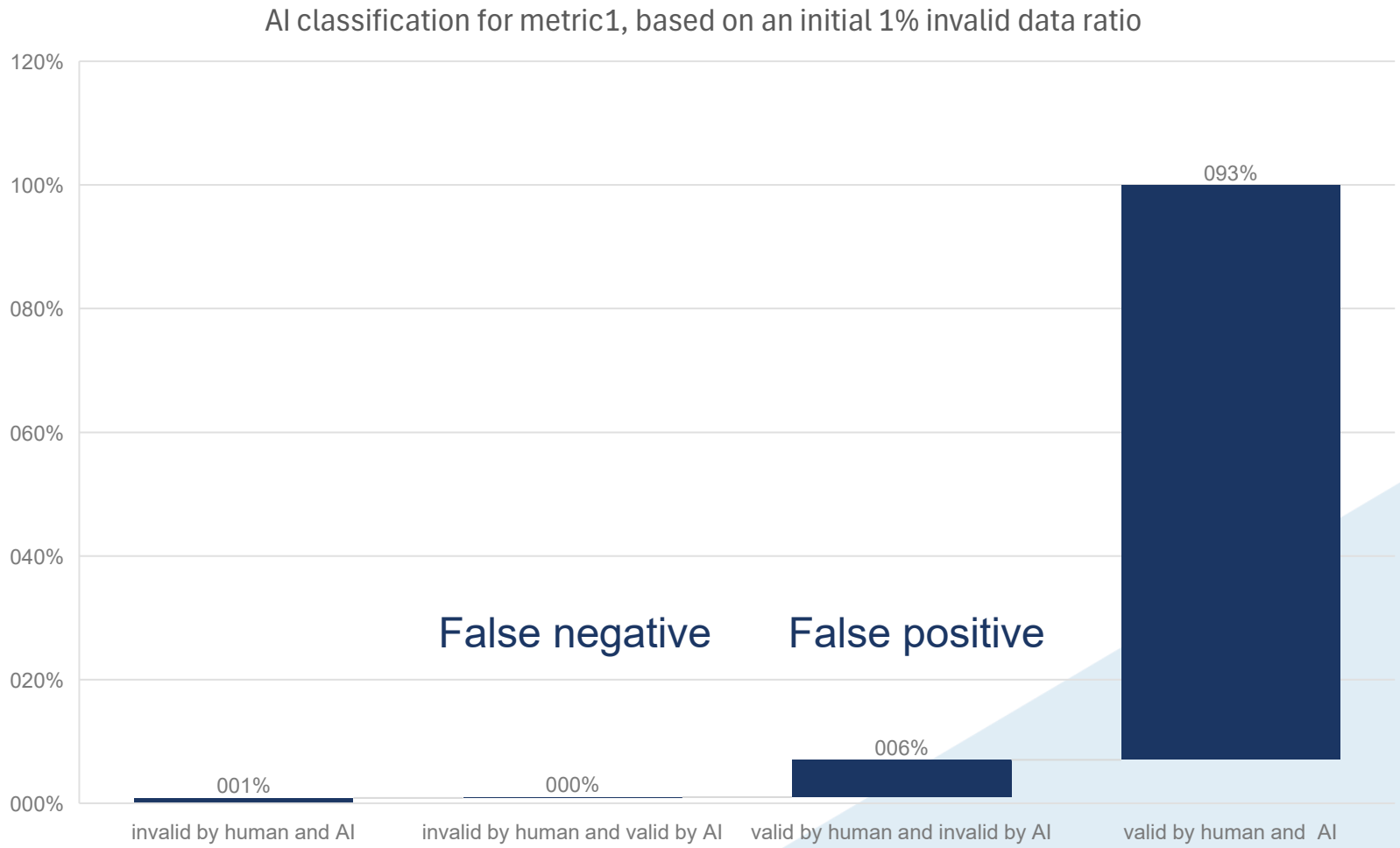
4. **Overall Performance Analysis:**
   - Provides a general overview of model performance

Model evaluation

Pollutant	Metric 1 (one-off)	Metric 2 (periods)	Metric 3 (high probability of error)	Métrique 4 (overall performance)
O_3	92%	78%	8%	0,875
NO_2	86%	79%	9%	0,855
NO	87%	84%	7%	0,880
PM10	92%	90%	8%	0,910
PM2.5	94%	87%	8%	0,910
CO	83%	80%	11%	0,840

Skipping local training step will reduce accuracy by a few %
Effective invalide rate is below 1%

Data classification for NO



Interpretation:

AI classifies 7% of the data as invalid or suspicious, while 93% has been classified as valid

Regarding all human invalidated data (in average <1% of the total), 87% have been classified as invalid by the AI

It means 13% of 1% of the data have been classified as valid while human invalidated it.

Conclusion

User feedback and follow up

Initial performance review made user interested to test it on real time data

Several improvements are planned, especially proper labelling of invalid data reason (why the AI classified it as invalid)

Question on how to consider the uncertainty on the human validation

ENVEA air quality monitoring application manager

Jérôme LOUAT

j.louat@envea.global

+33 6 63 98 49 26


envea