Predictive anomaly detection using Optimal Transport and Time Series Forcasting

Adil AHIDAR[1], Issiaka KANAZOE[1*],
Landry Yves Joel SEBEOGO[1†], Namou Karl Alban KADEWA[1†],
Palingwend Sosthène ZONGO[1†], Thomas Christian BADOLO[1†]

[1*]Ecole Centrale Casablanca, Morocco.

*Corresponding author(s). E-mail(s):
issiaka.kanazoe@centrale-casablanca.ma;
Contributing authors: adil.ahidar@centrale-casablanca.ma;
landry.sebeogo@centrale-casablanca.ma; alban.kadewa@centrale-casablanca.ma; sosthene.zongo@centrale-casablanca.ma;
christian.badolo@centrale-casablanca.ma;
[†]These authors contributed equally to this work.

Abstract

The contemporary industrial landscape, driven by rapid technological advancements, necessitates effective predictive maintenance strategies to ensure operational readiness. This paper proposes a novel approach to predictive anomaly detection, combining time series forecasting and Sinkhorn Optimal Transport. Leveraging the unique advantages of Sinkhorn Optimal Transport, which efficiently aligns probability distributions, our framework enhances predictive accuracy in identifying abnormal events in various industrial data types. The method is compared with the established One-Class SVM, demonstrating its effectiveness in anomaly detection. The integration of evolutionary representations, specifically time series forecasting, provides a comprehensive solution for proactive maintenance in technology-dependent industrial settings. The proposed framework showcases multi-feature anomaly detection capabilities, addressing challenges in monitoring diverse system metrics. Experimental results on a comprehensive dataset validate the robustness and accuracy of the approach, establishing its potential for real-world industrial applications.

Keywords: predictive maintenance, anomaly detection, time series forecasting, optimal transport, One-Class SVM, sinkhorn algorithm.

1 Introduction

The contemporary society, driven by rapid industrial advancements, stands at a pivotal moment marked by unprecedented innovations. These technological strides offer diverse solutions that propel the evolution of our digital landscape. Nevertheless, the reliance on technology presents several challenges, and part of those challenges lies in ensuring the operational readiness of machinery. Indeed, it is becoming critical in technology-dependent industrial settings to avoid production disruptions. Termed as "predictive maintenance," this approach is recognized and seen as the most optimal one within the global machine learning community.

Predictive maintenance involves leveraging data-driven insights and machine learning algorithms to anticipate potential equipment failures in industries, allowing proactive maintenance actions to be taken. A key point in such prediction is anomaly detection. One can utilize this method to oversee critical system metrics like machinery temperature, pressure, or voltage. An anomaly is defined when deviations from standard performance occur.

The nature of input data is a crucial factor in anomaly detection [1]. Data representation significantly impacts the applicability of anomaly detection techniques . Instances are categorized into three forms: metric, evolutionary, and multistructured representations. In our work, the focus has been on evolutionary representations, especially time series, where predicting upcoming data and analyzing current trends is a necessity.

For anomaly detection techniques applicable to various data representations, common approaches are presented in the following table.

Table 1 Overview of Anomaly Detection Techniques [1]

Approach	Description	Example
Statistical Techniques	Use a statistical model to classify data instances. Parametric techniques assume a known or assumed distribution, while non-parametric techniques do not.	Parametric: Regression models, kernel-based models. Non-parametric: Histograms.
Classification Techniques	Assign objects to predefined categories. Construct predictive models to classify instances into normal or anomaly classes.	One-class SVM, Bayesian networks, Rule-based techniques.
Nearest Neighbors	Assume normal instances occur in dense areas, while anomalies occur in sparse areas. Utilize distance or similarity measures.	k-NN (k-Nearest Neighbors), LOF (Local Outlier Factor).
Clustering Techniques	Group instances closely related. Anomalies either outside all clusters or far from the centroid of the nearest cluster.	DBSCAN, K-means, EM (Expectation-Maximization), CBLOF (Cluster-Based Local Outlier Factor).

One-Class SVM, a classification technique, is considered the current state-of-the-art method for our project, offering robust anomaly detection capabilities. This technique distinguishes itself from other SVM methods by being specifically designed for single-class classification.

Introducing **Sinkhorn Optimal Transport**, a novel technique leveraging optimal transport principles. Sinkhorn Optimal Transport differentiates itself by employing advanced optimization methods to efficiently align probability distributions. This approach proves promising for our anomaly detection task, offering unique advantages in capturing complex data relationships and enhancing predictive accuracy. Some authors studied such novelty by investigating on the effective identification of abnormal spectral densities in one sound recording signals. **Sinkhorn's distance** was employed to compute the distances that represent **deviations from a reference signal**.

Our proposed framework integrates time series forecasting, Sinkhorn Optimal Transport, and multi-feature prediction. In this context, our models are applied to various types of industrial data with the goal of identifying future abnormal events. The real-time series are forecasted, and a criterion based on sinkhorn's distance is employed to classify them as normal or abnormal. Our framework can be drawn as followed:

- **Combining Time Series Forecasting with novel unsupervised detection technique:** Our framework combines time series forecasting techniques with anomaly detection, capitalizing on the inherent temporal dynamics of sequential data for improved prediction accuracy.
- **Multi-Feature Anomaly Detection:** To differentiate from existing anomaly detection methods using optimal transport, our model efficiently addresses multi-feature scenarios, encompassing pressure, voltage, and vibration data.
- **Comparison with One-Class SVM:** The evaluation with an established model would definitely demonstrates its effectiveness in anomaly detection."'

This work delves into the historical evolution of **optimal transport** theory, tracing its path from Monge to Kantorovich. In **Section 3**, we shift our focus to **time series analysis and anomaly detection**, exploring insights from forecasting time series patterns to experimenting with state-of-the-art methods like One-Class SVM. This comprehensive exploration aims to provide a nuanced understanding of our research methodology and results in the realm of predictive maintenance using optimal transport theory and advanced time series analysis techniques.

2 Optimal transport and anomaly detection

2.1 Depecting the concept's evolution : traversing Monge to Kantorovitch [2]

The theory of Optimal Transport (OT) originated with the French geometer Gaspard Monge, a professor at the École du Génie Militaire de Mézières. In his seminal work, "Mémoire sur la théorie des déblais et des remblais" (Monge, 1781), Monge addressed the question: "How can one move a pile of earth (a natural resource) to a target location with the least effort or cost?"

Monge's Problem: Considering mass densities f and g as the densities of the source and target locations, respectively, Monge formulated the problem of finding a mapping $T : \mathbb{R}^d \to \mathbb{R}^d$ that pushes one density onto the other, while satisfying the condition

$$\int_A g(y)\,dy = \int_{T^{-1}(A)} f(x)\,dx$$

for any Borel subset $A \subset \mathbb{R}^d$. The objective is to minimize the quantity

$$M(T) = \int_{\mathbb{R}^d} |T(x) - x| f(x)\,dx$$

among all maps meeting this condition.

General Constraints: The image measure of a measure μ on X through a measurable map $T : X \to Y$ is denoted by $T\#\mu$ on Y, characterized by

$$(T\#\mu)(A) = \mu(T^{-1}(A)) \quad \text{for every measurable set } A,$$

or

$$\int_Y \phi\, d(T\#\mu) = \int_X \phi \circ T\, d\mu \quad \text{for every measurable function } \phi.$$

Problems Encountered in Monge Definition:

- Existence of a Solution [3]: The transport problem may not have a solution due to differences in total masses or the presence of atoms in the measures. The existence of atoms in μ implies atoms in ν, leading to a solutionless problem.
- Existence of an Optimal Transport Plan : In the context of optimal transport, Filippo raised a critical issue regarding the existence of an optimal transport plan in the Euclidean setting. When considering two measures μ and ν induced by densities f and g, respectively, the equality $\mu = T_{\#}\nu$ can be expressed in the Partial Differential Equation (PDE):

  $$g(T(x))\det(DT(x)) = f(x) \quad \text{for } x \text{ in the domain } \Omega \quad [2]$$

  This equation, obtained through a change-of-variables formula, represents the constraint for the optimal transport plan. However, the non-linearity of this constraint in T raises challenges.
  The non-linearity of the derived PDE poses challenges in guaranteeing the existence of an optimal transport plan. Strategies involve using minimizing sequences in weak topology, proving weak convergence for subsequences, demonstrating existence based on the lower semicontinuity of the integral criterion, and establishing limits. However, a persistent issue arises: the optimal value often fails to satisfy the non-linear constraint of the problem.

Relaxation of the Problem by Kantorovitch: In the 1940s, Russian mathematician and economist Leonid Kantorovich revitalized the theory of transport by broadening Monge's restrictive problem, thus overcoming the encountered issues. Instead of minimizing the transport cost with the transport map T, Kantorovich shifted the focus towards measures. This innovative approach earned him the Nobel

Prize in Economics. Given two probability measures μ and ν on Ω and a cost function $c : \Omega \times \Omega \to [0, +\infty]$, the problem aims to minimize

$$\min_{\gamma \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} c\, d\gamma,$$

where $\Pi(\mu, \nu)$ denotes the set of transport plans between μ and ν.

This relaxation by Kantorovich paved the way for a more versatile approach to optimal transport, embracing the broader context of measure-based solutions.

Wasserstein's distance : **Wasserstein Distance:**

The Wasserstein metric is a powerful metric deriving from Monge and Kantorovich transport problem for comparing probability distributions. It can be used to measure the similarity between distributions, to compare the performance of different algorithms, and to solve optimization problems. Its general expression is as follows:

$$W_p(\mu, \nu) = \inf_{(X,Y)\in\Pi(\mu,\nu)} \mathbb{E}[d(X, Y)^p] \tag{1}$$

‘‘‘

where

- $W_p(\mu, \nu)$ is the Wasserstein metric between μ and ν
- μ and ν are probability distributions
- $\Pi(\mu, \nu)$ is the set of all couplings of μ and ν
- $d(x, y)$ is a distance function

Considering $d(x, y)$ to be the euclidean norm :

$$W_p(\mu, \nu) = \inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^d} \|x - y\|^p \pi(dx, dy)$$

For p = 1, the Wasserstein distance is equivalent to the Earth Mover's Distance (EMD) which shows a lot of interesting features, even if from the point of the existence of an optimal map T it is one of the most difficult. [4]

2.2 The sinkhorn's distance : a fabulous addition for anomaly detection

Regularized optimal transport : Another fondamental tool gaining popularity in optimal transport research papers is the regularized optimal transport. It is an extension of the classical optimal transport problem that incorporates a regularization term. The problem is then redefined as follow, where the Sinkhorn distance, denoted as $S_\epsilon(\mu, \nu)$, between two probability measures μ and ν on a metric space Ω with a cost function $c : \Omega \times \Omega \to [0, +\infty)$, is defined as:

$$S_\epsilon(\mu, \nu) = \min_{\gamma\in\Pi(\mu,\nu)} \langle\gamma, c\rangle + \epsilon H(\gamma) \tag{2}$$

Here, $\Pi(\mu, \nu)$ represents the set of all joint probability measures on $\Omega \times \Omega$ with marginals μ and ν, $\langle \gamma, c \rangle$ is the cost of the optimal transport plan γ, ϵ is the regularization parameter, and $H(\gamma)$ is the entropy term given by:

$$H(\gamma) = -\sum_{i,j} \gamma_{ij} \log(\gamma_{ij})$$

The entropy term regularizes the transportation plan, preventing it from becoming too sparse. It penalizes high-mass concentration and ensures a smooth and stable computation of the Sinkhorn distance. The solution to the regularized optimal transport takes the form:

$$\forall (i,j) \in [[1,n]] \times [[1,m]], P_{i,j} = u_i K_{i,j} v_j$$

where K is the Gibbs kernel, $K_{i,j} = \exp\left(-\frac{C_{i,j}}{\varepsilon}\right)$, and u and v are unknowns.

One can follow steps of Sinkhorn algorithm as it offers an iterative approach to efficiently compute the optimal regularized transport matrix

Algorithm 1 Sinkhorn Algorithm

procedure SINKHORN($a, b, M, \mathrm{reg}, \mathrm{num\_iters}$)
    Normalize a and b to ensure they sum to 1
    $a \leftarrow a / \sum a$
    $b \leftarrow b / \sum b$
    Compute exponentiated cost matrix $K \leftarrow \exp(-M/\mathrm{reg})$
    Initialize $u \leftarrow \mathbf{1}$
    for $i = 1$ to num_iters **do**
        Update $u \leftarrow a/(K \cdot (b/(K^T \cdot u)))$ ▷ Sinkhorn update for u
    end for
    Compute optimal transport matrix $P \leftarrow \mathrm{diag}(u) \cdot K \cdot \mathrm{diag}(b)$
    return P
end procedure

In [5], it is proven that the algorithm's complexity is $O(n^3 \log(n))$ with an empirical complexity of $O(n^2)$. The Sinkhorn algorithm, mainly involving matrix multiplications, is highly parallelizable. For instance, it can compute multiple solutions to problem 2 simultaneously, making it suitable for GPU execution. This feature significantly enhances its efficiency, especially in learning contexts dealing with large datasets. The convergence of the Sinkhorn algorithm is guaranteed, as demonstrated by Franklin in [6].

Sinkhorn's distance for anomaly detection : Our study looks forward to get use of Sinkhorn's distance previously stated for its application into industrial fields, mainly to detect abnormal observations in data. Indeed, as Anomaly detection involves identifying patterns or instances in data that deviate significantly from the norm, Sinkhorn's distance will serve as a robust metric for measuring those deviations from a standard probability distributions indicating normal behavior of dataset.

Our work relies on Amina Alaoui-Belghiti's work in[7]. As mentioned in the state of the art, their study aimed to efficient identification of abnormal spectral densities of sound recording signals. Distances representing their deviations from a reference signal were calculated using sinkhorn's distance. Then,using a threshold corresponding to a certain percentile of the distances distributions the behaviors were classified as normal or abnormal.

Hence, in the context of the optimal transport implementation in our project, the same methodology would be extended to encompass multiple sensor data types, like pressure, voltage, rotation. Here is the pseudocode for the classification of abnormalities using optimal transport:

Algorithm 2 Anomaly detection with OT

Require: Set of reference signals X, signal to evaluate X'
Ensure: Binary classification: 1 for normal signal, -1 for abnormal signal
procedure PREDICT_ANOMALY(X, X')
for $i = 1$ to k **do**
    $d_i \leftarrow$ ROT_distance$(F(X), F(X'))$
end for
$\hat{\mu} \leftarrow$ lognormal_fit$(\{d_i\}_{i=1}^{k})$
$d \leftarrow$ ROT_distance$(F(X), F(X'))$
if $d >$ threshold$(\hat{\mu})$ **then**
    return -1 ▷ Abnormal
else
    return 1 ▷ Normal
end if

- The algorithm takes two sets of signals as input: a set of reference signals X and a signal to evaluate X'.
- It iterates k times, each time calculating the Regularized Optimal Transport (ROT) distance between a feature representation of the reference signals $F(X)$ and the feature representation of the signal to evaluate $F(X')$. This distance is stored in d_i.
- The algorithm then fits a log-normal distribution to the k ROT distances, obtaining an estimate of the distribution of ROT distances for normal signals.
- It then calculates the ROT distance between $F(X)$ and $F(X')$ again and compares it to a threshold derived from the estimated log-normal distribution.
- If the ROT distance is greater than the threshold, the algorithm returns -1, indicating that the signal is abnormal. Otherwise, it returns 1, indicating that the signal is normal.

3 Time series : from data analysis to anomaly detection

3.1 Time series, a data analysis concept

Time series analysis primarly seen as a data analysis concept, specificaly for the modeling of data points collected and ordered over time. It mainly focuses on uncovering temporal patterns, trends and dependencies within datasets. The data points are recorded on a consistent interval basis over a set period of time rather than just randomly.

Distinguishing time series data from other datasets lies in its unique attribute of capturing how variables evolve over time. This temporal dimension introduces an additional layer of information, enhancing the depth and context of the analytical insights derived from such data.

In this paper, our focus is directed towards the examination of a particular data structure. We delve into the intricacies of anomaly detection within a given time series, employing a comparative framework against a reference set comprised of normal time series. This problem formulation finds widespread applicability across various domains, including industrial machinery diagnostics, shape anomaly detection, and identification of aberrant light curves in astronomical data, among others.

Time series has indeed become a domain of attraction for researchers for past few decades.

3.2 Time series forecasting

Through time series modeling, essential part remains in rigorously studying the past observations of the series meticulously collected beforehand. Upon thorough analysis, future data are generated, commonly called forecasts. At the time of the work, the future outcome is completely unavailable and can only be estimated through careful analysis and evidence-based priors.

Time series forecasting properly starts by defining the problem meaning a clear clarification of the forecasting problem, specifying the target variable and desired prediction timeframe. The process must also go through data assessment implying the evaluation of the relevance, completeness, and quality of available data to ensure its suitability for forecasting. Next steps are Trend and Seasonality analysis involving the exploration of the time series utilizing statistical methods or visualization techniques. The model can be eventually get to its development and prediction phases afterwards.

Optimal transport and Time series forecasting for predictive maintenance : Our project emphasizes the symbiotic relationship between Sinkhorn optimal transport (ROT) and time series data. The marriage of ROT and time series is imperative, amplifying the efficacy of our anomaly detection approach, where the fusion of multi-sensor data types becomes pivotal for comprehensive insights and accurate predictions.

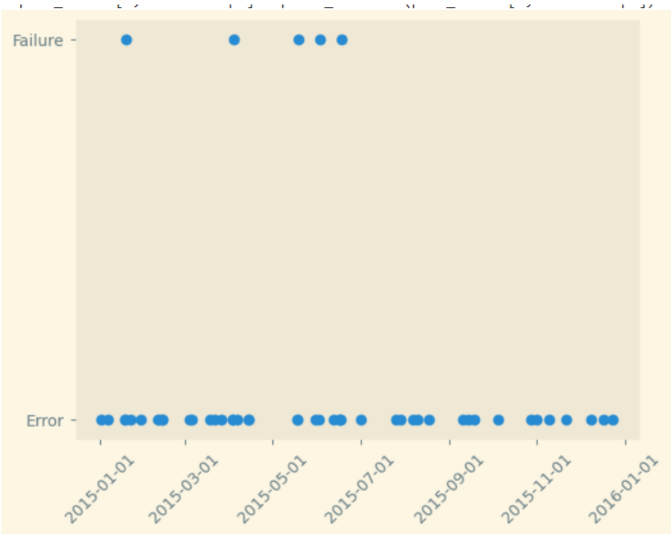

Fig. 1 Correlation between failure and error

4 Experimentation

Predictive Maintenance, a cornerstone in enhancing operational efficiency, faces a significant challenge in acquiring real-life industry data. Access to authentic datasets, reflective of actual industrial conditions, remains a formidable task due to confidentiality concerns and the sensitive nature of operational data. This challenging context hinders the implementation of models on genuine datasets from industrial areas. However, the example dataset discussed herein serves as a critical stepping stone, drawing inspiration from real industry area data to construct a comprehensive AI-generated dataset suitable for developing and refining predictive maintenance models.

4.1 Datasets description

In the realm of Predictive Maintenance Model Building, the choice of a comprehensive and representative dataset plays a pivotal role. An exemplary source that encapsulates diverse facets crucial for model development is at our disposal. This dataset meticulously documents various dimensions, including machine conditions, failure history, maintenance records, and distinct features of machines. With a focus on optimizing predictive maintenance strategies, this dataset emerges as a valuable asset.

Data Details: The **Telemetry Time Series Data** (PdM-telemetry.csv) provides hourly averages of voltage, rotation, pressure, and vibration from 100 machines in 2015. The **Error dataset** (PdM-errors.csv) records errors during machine operation, contributing to a comprehensive understanding of system behavior. Lastly, the **Failures dataset** (PdM-failures.csv) zeroes in on component replacements resulting from failures, offering crucial insights into reliability concerns.

Supplementary datasets on maintenance and machine metadata have been included alongside the core telemetry and failure records. While these datasets provide insights

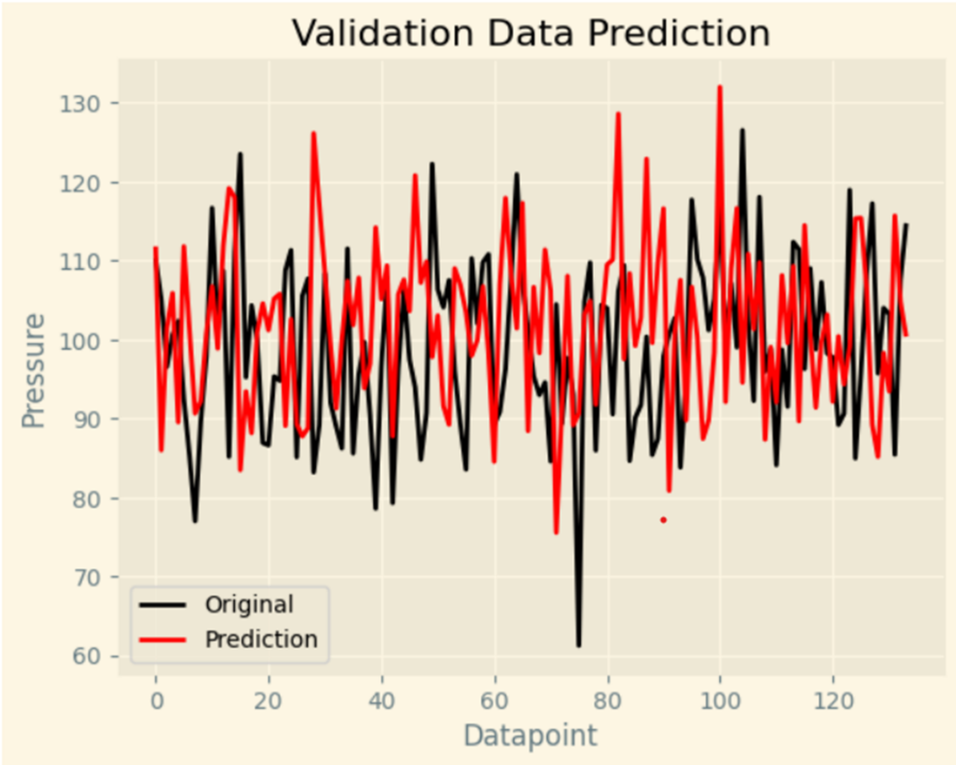

Fig. 2 Best visualization of the forecast for 150 pressure readings.

into maintenance scenarios and machine details, their relevance is comparatively lower for the primary objectives of predictive maintenance and time series forecasting.

Afterwards, We investigate the **correlation between errors and failures** by analyzing their history for a specific machine over a year in figure 1. The depicted correlation suggests that failures often follow errors, though not all errors result in immediate failures. This prompts us to prioritize failure data to discern the features impacted by machine failures.

4.2 Forecasting Time Series Patterns in the Provided Dataset

The time series forecasting model used in our approach belongs to the **family of machine learning models**, specifically a **deep learning model** implemented using the Keras library with TensorFlow backend. The model utilizes a Sequential architecture with **Long Short-Term Memory** (LSTM) layers, which are recurrent neural network (RNN) variants suitable for sequence prediction tasks. The LSTM layers are followed by densely connected layers, incorporating non-linear activation functions. 4 to 8 layers are used to enhance the capacity of the model so as to handle complexe and abstract features and temporal structure. The model is trained using the Adam optimizer with a mean squared error loss function for 500 to 1000 epochs.

To evaluate the model's robustness, it was tested on a dataset comprising over **600 hourly pressure measurements** within a specific month. This chosen month, selected for having the **highest number of failure records** during the year, was divided into 80% training data and 20% validation data.

The sequence of input data, denoted as X, is shaped using a function `shape_sequence` with a step size of 5 used to predict the sequence of output data shaped with a step size of 1. This means that Observation of past 5 hours will be used to predict the sensor reading for the Next 1 Hour and so on for all sequences of the data.

For fitting the LSTM model, the observations are rescaled using Min-Max scaling. The results are presented in the following table with the optimal parameter highlighted in bold :

Leading rate	**RMSE**
0.1	0.02219640362558153
0.035	0.022664972988761415
0.01	0.018584039
0.00498888	**0.01824395**

Table 2 Validation error with respect to learning rates

4.3 Detecting anomalies in forcasted series using Optimal transport

Feature	**Threshold**
Voltage (volt)	0.0035898579873170425
Pressure (pressure)	0.004133416797878981
Vibration (vibration)	0.0051033791177524015
Rotation (rotate)	0.0046177235130853275

Table 3 Thresholds for Different Features

Barycenters, representing the mean values of telemetry features ('volt', 'rotate', 'pressure', 'vibration') for each hour across all machines, are computed as the means of the features data **for each hour on one day** across all machines, resulting in a barycenter vector of **24 consecutively measurements** for each feature. Each barycentre modelizes the **normal behavior** in one specific day of the associated feature.

Leveraging the computed barycenters, the Sinkhorn distance is subsequently calculated between these barycenters and the daily machine data for each telemetry feature.. From those calculated distances, a **distribution of distances** following a **normal distribution** as shown in figure 3 is obtained in each case which allowed us to set thresholds at the **95th percentile**. Those thresholds are crucially important

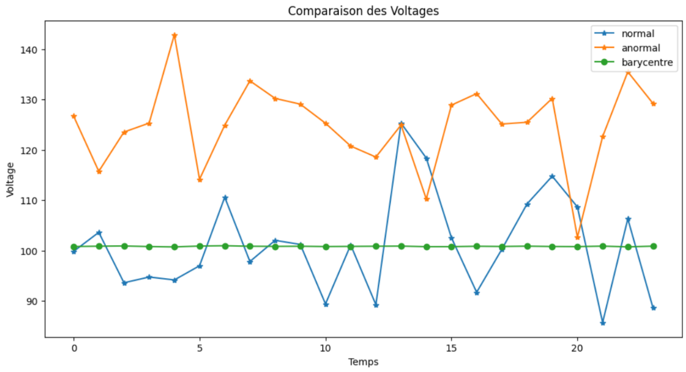

Fig. 3 Normal and abnormal performances around reference normal behavior of a machine over 24 hours

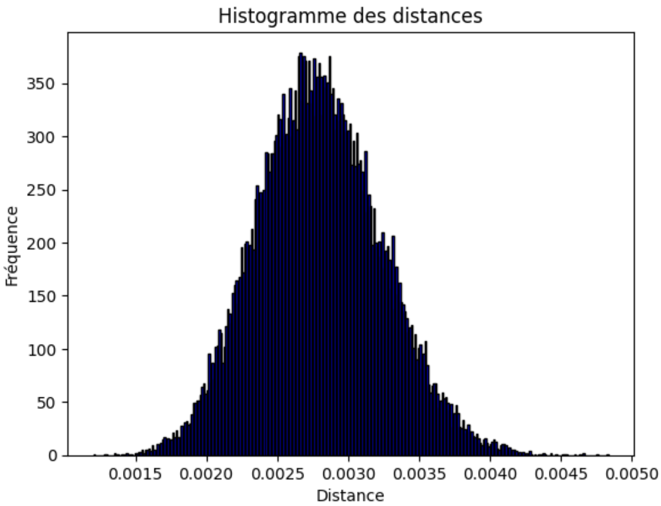

Fig. 4 histogram sinkhorn distances for pressure

as they help, associated with the sinkhorn distance, assess a behavior as normal or abnormal. The table gives the different obtained thresholds values.

Initially, the model is trained using the **complete telemetry dataset.** The dataframe is merged with the failure data and grouped by date and ID. It iterates through the groups, creating a list of DataFrames, where each DataFrame contains hourly measurements of features for specific day and machine ID . The resulting list `daily_dfs` contains these grouped DataFrames, facilitating further analysis or processing of data on **a daily and per-machine basis**.

In the context of multi-feature anomaly detection, we flag a machine as exhibiting abnormal behavior if any of its key features show unusual patterns surpassing predefined thresholds. The model is applied on the forcasted data gained from the time series forecasting model. Figure 4 shows the visual results of detected normal and abnormal behaviors of 24 hourly forcasted measurements in time.

4.4 Experimentation with the state of the art: One class SVM

The experimental methodology we adopt for One-Class SVM anomaly detection involves key steps. First, the DataFrame is divided into training and test sets, using the initial 700, 880 normal transaction observations for training. A One-Class SVM model with a 'linear' kernel, gamma of 0.001, and nu of 0.95 is trained exclusively on the normal transactions. Subsequently, the remaining observations are combined with anomalous ones to create a comprehensive test set. The methodology includes hyper-parameter tuning, exploring variations in kernel (linear, rbf, poly), gamma (0.001, 0.0001), and nu (0.25, 0.5, 0.75, 0.95). The model is trained and evaluated on the test set, and the chosen linear kernel and specific hyperparameters reflect a focus on linear separation for this anomaly detection task.

4.5 Comparison of results

Metric	Our Model	One-Class SVM Model
True Positive	9300	8812
False Negative	50	165955
True Negative	760	187
False Positive	250	3732
Sensitivity	$\frac{9300}{9350} \approx 0.9957$	$\frac{8812}{174767} \approx 0.0505$
Specificity	$\frac{760}{1010} \approx 0.7525$	$\frac{187}{3919} \approx 0.0477$
Accuracy	$\frac{10060}{10190} \approx 0.9873$	$\frac{8999}{180833} \approx 0.0498$

Table 4 Comparison of Detection Model Performance

In our evaluation, the robustness of our anomaly detection model is validated through three crucial criteria: **sensitivity, accuracy, and specificity**. Sensitivity measures the model's capability to identify anomalies accurately, accuracy reflects overall correctness, and specificity evaluates the model's effectiveness in recognizing non-anomalous instances. The results obtained are presented for the two detection models under consideration in table 4 proving the significant advantage of chosing the studied model.

4.6 Discussion and Conclusion

In conclusion, this work introduces a novel approach to predictive maintenance, integrating Sinkhorn Optimal Transport for anomaly detection. By efficiently addressing anomalies across various telemetry features, the model demonstrates promising potential for enhancing operational efficiency in diverse industrial scenarios. The integration of time series forecasting further strengthens anomaly detection, providing a comprehensive solution.

The challenges in acquiring authentic industrial datasets due to confidentiality concerns are acknowledged. Nevertheless, the innovative techniques presented in this work offer a valuable contribution to predictive maintenance strategies. The model's ability to align probability distributions using Sinkhorn Optimal Transport, combined with multi-feature detection, positions it as a noteworthy advancement.

For further exploration, the model's applicability across different industrial contexts and scalability for large datasets could be investigated. A more detailed discussion on limitations and potential improvements would enhance the overall understanding of the presented approach.

Overall, this work marks a significant step forward in leveraging advanced methods for predictive maintenance, showcasing the potential impact of innovative techniques in accurately detecting anomalies in industrial settings.

References

[1] Amina Alaoui-Belghiti, *Prediction of abnormal situations by machine learning in a predictive maintenance context: Optimal transport theory for anomaly detection*, PhD thesis, Université Paris-Saclay, 2021.

[2] Filippo Santambrogio, *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, Birkhäuser, 2015, Progress in Nonlinear Differential Equations and Their Applications, volume=87, doi=10.1007/978-3-319-14321-8.

[3] Cécile CARRÈRE, Didier LESESVRE, Paul PEGO, *École Normale Supérieure de Cachan Centre des Mathématiques et de leurs Applications Mémoire de stage Directeur de stage: Filippo SANTAMBROGIO Théorie générale du transport et applications*, École Normale Supérieure de Cachan, Year.

[4] Filippo Santambrogio, *Introduction to Optimal Transport Theory*, Grenoble, June 15th, 2009.

[5] Camille Mondon and Julien Malka, *Du transport optimal et de son application en apprentissage statistique*, DMA, ENS and DI, ENS, June 14, 2020.

[6] Joel Franklin, Jens Lorenz, *On the scaling of multidimensional matrices*, Linear Algebra and Its Applications, pages 717–735, 1989.

[7] Amina Alaoui-Belghiti, Sylvain Chevallier, Eric Monacelli, *Unsupervised anomaly detection using optimal transport for predictive maintenance*, Nexeya, Massy, France; UVSQ, Université Paris-Saclay, amina.alaouibelghiti@nexeya.com; (sylvain.chevallier, eric.monacelli)@uvsq.fr.