Trustworthiness Score for Echo State Networks by Analysis of the Reservoir Dynamics

Universidad de Oviedo

GSDPI grupo de supervisión, diagnóstico y descubrimiento del conocimiento en procesos de ingeniería

José M. Enguita[1], Diego García[1], Abel A. Cuadrado[1], Daniel Peña[1], José R. Rodríguez[2], and Ignacio Díaz[1]

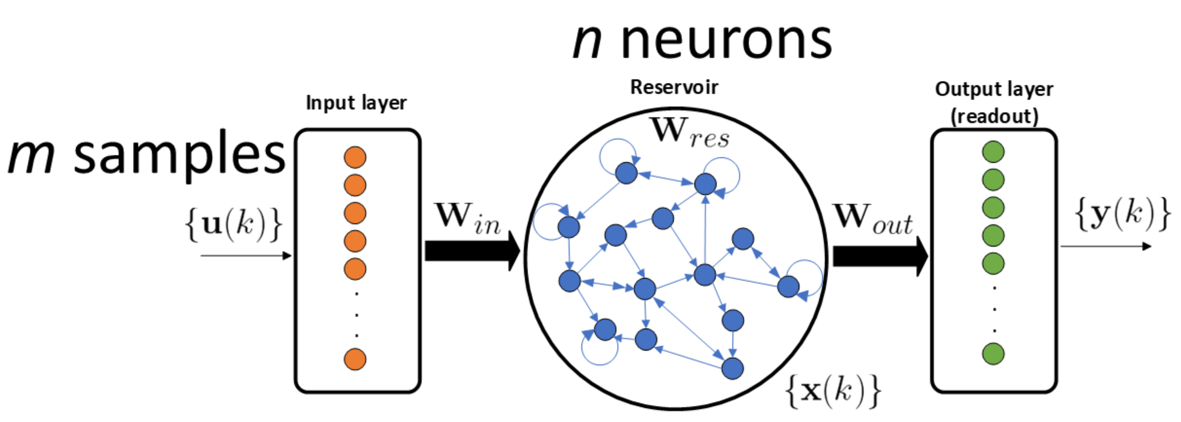
1– GSDPI Research Group, University of Oviedo, 2- SUPPRESS Research Group, University of León

Contact author: jmenguita@uniovi.es

Epistemic uncertainty arises from input data areas where models lack exposure during training and may result in significant performance degradation in deployment. Echo State Networks are often used as virtual sensors or digital twins processing temporal input data, so their robustness against this degradation is crucial. This paper addresses this challenge by **proposing a score comparing the similarity between the dynamic evolution of the reservoir in training and in inference**. This research aims to enhance model confidence and adaptability in evolving circumstances.

Methods

Theory



$$\mathbf{X} \in \mathbb{R}^{n \times m} = \begin{pmatrix} | & | & & | \\ \mathbf{x}(1) & \mathbf{x}(2) & \cdots & \mathbf{x}(m) \\ | & | & & | \end{pmatrix}$$

SVD

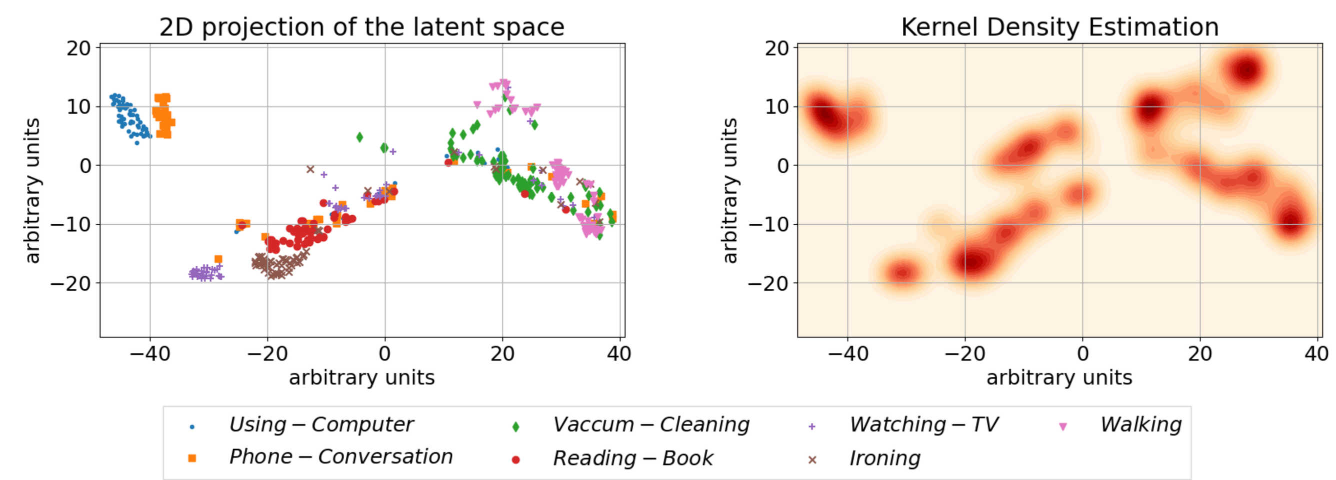
$$\boxed{\mathbf{U}\Sigma\mathbf{V}^T = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^T + \cdots + \sigma_n \mathbf{u}_n \mathbf{v}_n^T}$$

- Each state vector **x(k)** can be considered an expanded set of descriptors of the dynamics of the input signal as seen by the ESN at instant **k**.
- The SVD of **X** is used to analyse these dynamics.
- Singular values (**σ**) represent the weights of the principal modes of **X**, and may be used as descriptors of the dynamic evolution of the reservoir, when excited by the input signal.

Procedure

During Training:

1. Apply a **sliding window** algorithm with window size **m** and stride **s** to extract the matrices **X** and compute the SVD.
2. Select the **most significant r singular values**, which represent a point in a reduced latent space of dimension **r**.
3. Utilise a **Kernel Density Estimation** (KDE) with Gaussian kernel of an adequate bandwidth to approximate the Probability Density Function (PDF) of the reservoir dynamics covered during training.

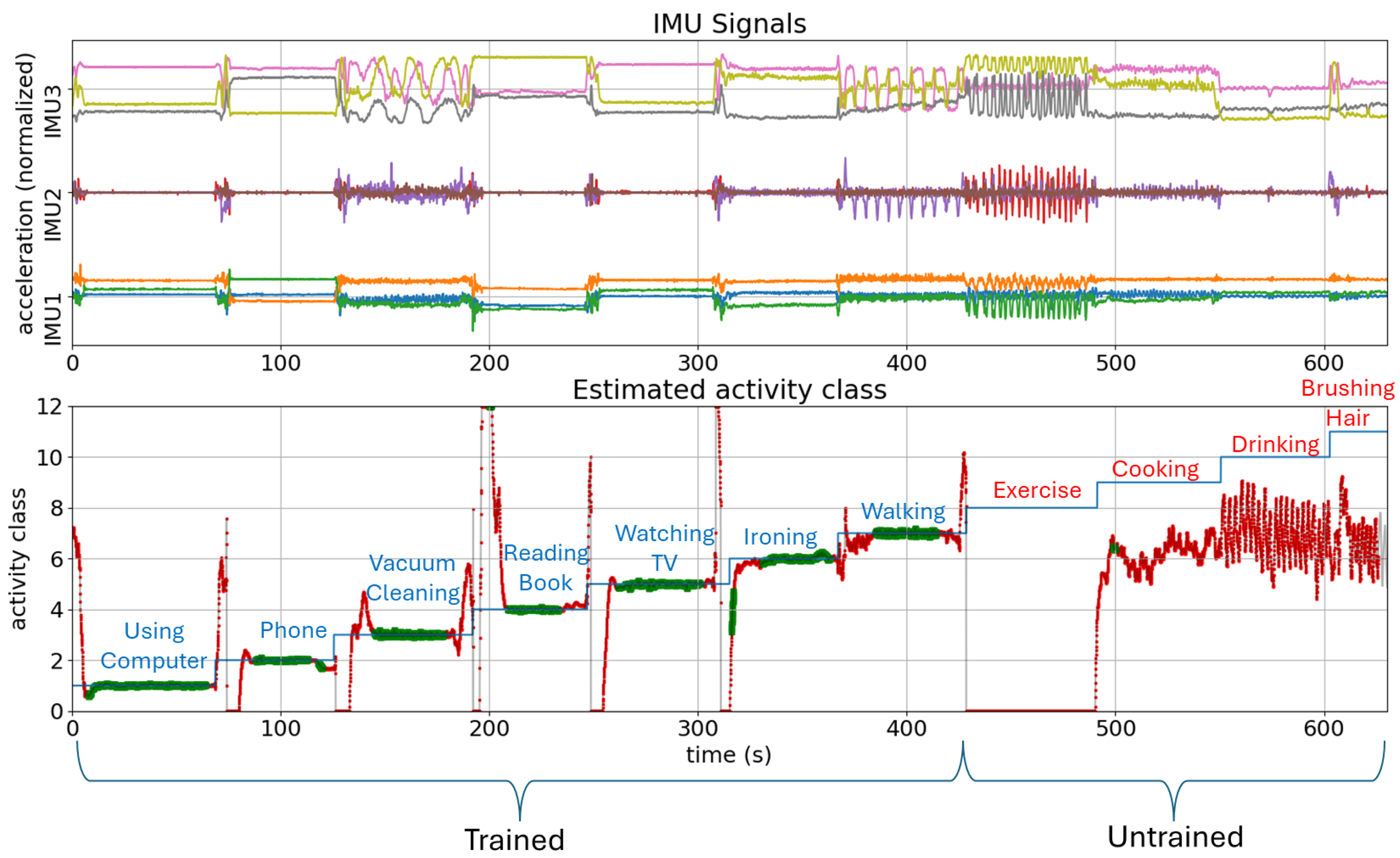


In Production:

1. Apply the **sliding window algorithm** again, capturing points in the latent space for the incoming signals.
2. The **trustworthiness score** of each point is calculated as its **log-likelihood under the PDF model.**
3. A **threshold** may be used to classify current predictions as **high or low confidence.**

Results

- **Dataset**: Intelligent Media Wearable Smart Home Activities dataset (Tahir 2020), available at https://portals.au.edu.pk/imc/Pages/Datasets.aspx, which comprises the acceleration signals of three triaxial IMU sensors attached to a set of subjects wrist, chest, and thigh region while performing 11 different activities
- The package **ReservoirPy** (Trouvain 2020) was used for the implementation.
- The model was **trained** to classify: 'using computer', 'phone conversation', 'vacuum cleaning', 'reading book', 'watching tv', 'ironing', and 'walking'. Each one is assigned a consecutive class number from 1 to 7. The first 300 data points (15 seconds) of each activity are omitted, as they contain basically noise, and the last 200 data points (10 seconds) are also omitted in order to have data of each class which the ESN model has not been trained with.
- The hyperparameters were manually selected as: n=300 neurons, spectral radius ρ = 0.95, sparsity = 0.01, $\mathbf{W_{in}}$ scale = 150, input scale = 1, warmup = 20, and bias = True.
- A reduced-dimension latent space formed by the first 5 singular values (**r** = 5) is used. The bandwidth for the KDE used in the estimation of the PDF is set to **bw** = 0.2.



Top, raw IMU signals. Bottom, result of prediction and score. Blue line: real class, green: ESN prediction, high confidence, red: ESN prediction, low confidence. ESN output has not been filtered or treated in any way, but it was cropped to the [0,12] interval for representation purposes (valid classes are 1-7).

Conclusions

- ✓ While the results are preliminary and numerous questions remain unanswered, the idea shows promise.
- ✓ The proposed scoring system effectively assess epistemic uncertainty, alerting users to potential inaccuracies in the model predictions.
- ✓ The method, simple and grounded in established techniques, requires no training and operates independently of the model's actual performance, relying solely on reservoir dynamics.
- ✓ It also unveils new avenues for exploration. This system could be the basis for detecting and responding to domain shifts or for analysing reservoir dynamics by projecting the latent space into a visualisation space.

This work is part of Grant PID2020-115401GB-I00 funded by MCIN/AEI/ 10.13039/501100011033
Code available at: https://github.com/gsdpi/trustworthiness-esn-ESANN2024

GOBIERNO DE ESPAÑA · MINISTERIO DE CIENCIA E INNOVACIÓN