ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

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

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

1- Dept. of Electrical Engineering, University of Oviedo, 33204 Gijón, Spain

2- SUPPRESS Research Group, University of León, 24007 León, Spain

Abstract. 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.

1 Introduction

As machine learning applications proliferate, new challenges beyond mere prediction accuracy appear. Among these challenges is the concept of *trustworthiness* (see, for instance, [1]), which refers to the model's ability to convey confidence in its own predictions. Trustworthiness spans various dimensions such as robustness, security, transparency, fairness, and safety [2]. We will focus in this paper on robustness against previously unseen input data.

Despite models showing good generalisation under specific conditions (sufficient data volume, appropriate biases, and consistency with the training distribution), they often struggle when faced with new scenarios in deployment. Even minor shifts in deployment conditions can lead to significant performance degradation. This uncertainty that arises from unfamiliarity with the input data is called *epistemic*[1], and indicates areas where the model lacks exposure during training. A good score must capture the nuances of epistemic uncertainty, enhancing the model's capacity to convey confidence and adapt to evolving circumstances.

An *Echo State Network* (ESN) is a computational model based on Reservoir Computing (RC) which is highly effective for processing sequential or temporal data, such as the behaviour of dynamic systems or time series, even when dealing with chaotic or spatio-temporally complex problems [3]. The input signals are fed into a set of interconnected neurons, referred to as the 'reservoir', which

*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`

[1]In contrast to *aleatoric* uncertainty, which stems from the inherent variability within the data.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

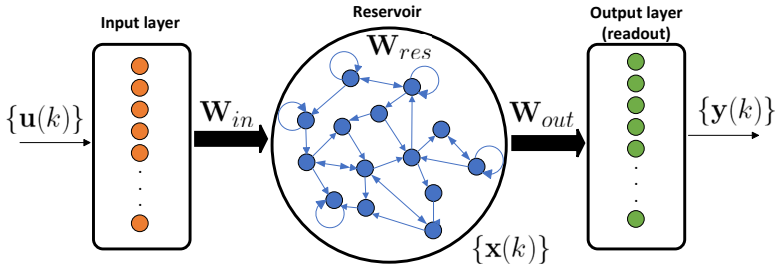

Fig. 1: Representation of an ESN.

generates a high-dimensional expansion [4]. The reservoir contains feedback (recurrent) loops that provide memory, and exhibits a complex and highly non-linear dynamic behaviour [5]. They also simplify the learning process, a simple readout system is trained to map the internal state of the reservoir to the desired output.

In this work we introduce a score aimed at quantifying the trustworthiness level in the predictions made by an ESN, leveraging the similarity between the evolution of the reservoir's dynamic response to the input signals and the reservoir's dynamics observed during the model-training process. Various methods exist for assessing this dynamic evolution; however, this paper specifically explores the application of Singular-Value Decomposition (SVD), which has already been used to characterize the dynamics of the reservoir [6, 7].

Evaluating the trustworthiness of an ESN model is an important task, as they are often used in control systems to predict outputs, as virtual sensors or as digital twins to detect unexpected or undesirable system behaviours [8, 9].

Although the methods presented in this paper are applied to ESN models, it should be simple to adapt them (with minor variations) to other reservoir computing systems.

2 Methods

2.1 Echo State Networks

ESNs were proposed in the beginning of the 21st century by Jaeger and Hass [5, 10, 11]. In its fundamental configuration, an ESN allows the modelling of non-linear systems through supervised learning. ESNs, depicted in Figure 1, are constructed on the principle of *non-linear expansion* [4]. This involves the conception of a high-dimensional state vector $\mathbf{x}(k) \in \mathbb{R}^n$ (being n the number of neurons in the reservoir), a non-linear formulation within the state equation, and a linear model governing the output derived from the state. The original rendition follows:

$$\begin{aligned} \mathbf{x}(k) &= \sigma(\mathbf{W}_{res}\,\mathbf{x}(k-1) + \mathbf{W}_{in}\,\mathbf{u}(k)) \\ \mathbf{y}(k) &= \mathbf{W}_{out}\,\mathbf{x}(k). \end{aligned} \tag{1}$$

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

The model represented by equation (1) includes the sets of input and output signals $\mathbf{u}(k)$ and $\mathbf{y}(k)$, the reservoir matrix $\mathbf{W}_{res} \in \mathbb{R}^{n \times n}$, the input matrix $\mathbf{W}_{in} \in \mathbb{R}^{n \times p}$, and the output matrix $\mathbf{W}_{out} \in \mathbb{R}^{q \times n}$ as parameters. A non-linear function σ, typically sigmoidal, is used in the state equation.

2.2 Analysis of the Reservoir Dynamics Using the SVD

By applying the model in equation (1) recursively, an input or excitation sequence $\mathbf{u}$ generates a sequence of state vectors containing the activation values of the reservoir neurons, $\mathbf{x}(k)$, from which the output of the ESN is calculated.

It is possible to apply a simple sliding window mechanism with size m to construct a matrix $\mathbf{X}$:

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

The rows of the matrix $\mathbf{X}$ contain the temporal evolution of the reservoir. Each state vector $\mathbf{x}(k)$ 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 the matrix $\mathbf{X}$ is used to analyse these dynamics: $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$. Singular values $\sigma_1, \ldots, \sigma_n$ (the diagonal of Σ) represent the weights of the principal modes of $\mathbf{X}$, and may be used as descriptors of the dynamic evolution of the reservoir, when excited by the input signal.

If we select only the r-first singular values ($r < n$), an approximation of $\mathbf{X}$ can be obtained as $\tilde{\mathbf{X}} = \mathbf{U}_r\Sigma_r\mathbf{V}_r^T$. In fact, this is the best rank-r approximation of $\mathbf{X}$ in the L_2-sense, and the r most significant singular values are particularly useful in characterising the system's operating point and define a reduced r-dimensional *latent-space* in which to study the reservoir dynamics.

2.3 Definition of the Trustworthiness Score

The proposed method works in two steps:

During the training phase, apply a sliding window algorithm with window size m and stride s to extract the matrices $\mathbf{X}$ and compute the SVD. Select the most significant r singular values $\sigma_1, \ldots \sigma_r$, which represent a point in a reduced latent space of dimension r. Once the training has completed, utilise a Kernel Density Estimation (KDE) with Gaussian kernel of an adequate bandwidth (bw) to approximate the Probability Density Function (PDF) of the reservoir dynamics covered during training.

In production, apply the sliding window algorithm again, capturing points in the latent space for the incoming signals. The trustworthiness score of each point is calculated as its log-likelihood under the PDF model.

There are, therefore, four free parameters that must be selected according to the problem: the window size m and stride s, the dimension of the latent space r, and the KDE bandwidth bw.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

2.4 The IM-WSHA Dataset

For testing, we have used the Intelligent Media Wearable Smart Home Activities dataset [12, 13], available at `https://portals.au.edu.pk/imc/Pages/Datasets.aspx`. The data comprises the acceleration signals of three triaxial IMU (inertial measurement unit) sensors attached to a set of subjects wrist, chest, and thigh region while performing 11 different activities: 'using computer', 'phone conversation', 'vacuum cleaning', 'reading book', 'watching tv', 'ironing', 'walking', 'exercise', 'cooking', 'drinking', 'brushing hair'. We used the data of just one subject for the experiments. Some class labels were slightly corrected, as they seemed to be out of sync at the beginning or end of each activity.

2.5 Training the ESN Model

The package *ReservoirPy* [14] was used for the implementation. The model[2] was trained to classify the following IM-WSHA activities: '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. 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.

3 Results and Discussion

A sliding window algorithm (m=100, s=20) is used to apply the method. This is first done over the training set. At each stride, the set of 300 singular values corresponds to a point in a 300-dimensional latent space representing the dynamics of the reservoir for the input signals. Fig. 2 shows a two-dimensional projection of this latent space (using a t-SNE [15]) for the different activities the model was trained with. This is just for illustrative (or visualisation) purposes. In the following results, 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 and, therefore, in the calculation of the trustworthiness score is set to $bw = 0.2$.

Then the method is repeated over the full signal and the score at each stride is calculated. Results shown in Figure 3 demonstrate how, by using a simple threshold, output data from the ESN can be classified either as 'high' (green) or 'low' (red) confidence, based solely in the similarity of the input signal with the training data as captured by the reservoir.

4 Conclusions

While the results are preliminary and numerous questions remain unanswered, the idea shows promise. The proposed scoring system effectively assess epistemic

[2]The hyperparameters were manually selected as: $n = 300$ neurons, spectral radius $\rho = 0.95$, sparsity $= 0.01$, $\mathbf{W}_{in}$ scale $= 150$, input scale $= 1$, warmup $= 20$, and bias $=$ True.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

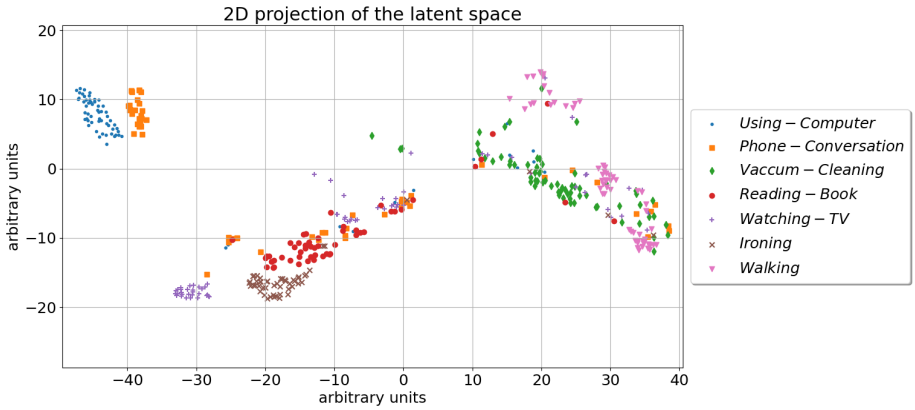

Fig. 2: 2D t-SNE projection of the latent space showing training data.

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.

References

[1] Bálint Mucsányi, Michael Kirchhof, Elisa Nguyen, Alexander Rubinstein, and Seong Joon Oh. Trustworthy machine learning. *arXiv preprint arXiv:2310.08215*, 2023.

[2] Bowen Zhou, Bo Li, Peng Qi, Bo Liu, Shuai Di, Jingen Liu, Jiquan Pei, and Jinfeng Yi. Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55:177, 2023.

[3] Erik Bollt. On explaining the surprising success of reservoir computing forecaster of chaos? the universal machine learning dynamical system with contrast to var and dmd. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 31:013108, 1 2021.

[4] Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

[5] Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304:78–80, 4 2004.

[6] Fanjun Li, Xiaohong Wang, and Ying Li. Effects of singular value spectrum on the performance of echo state network. *Neurocomputing*, 358:414–423, 9 2019.

[7] Claudio Gallicchio and Alessio Micheli. A markovian characterization of redundancy in echo state networks by PCA. In *Proceedings of the 18th European Symposium on ArtificialNeural Networks, Computational Intelligence and Machine Learning (ESANN-2010)*, pages 321–326. d-side, 4 2010.

[8] Allen G Hart. *Reservoir Computing With Dynamical Systems*. PhD thesis, University of Bath, 11 2021.

[9] José Ramón Rodríguez-Ossorio, Antonio Morán, Juan J. Fuertes, Miguel A. Prada, Ignacio Díaz, and Manuel Domínguez. Adaptive model for industrial systems using echo state networks. In *Communications in Computer and Information Science*, volume 1826 CCIS, pages 367–378. Springer Science and Business Media Deutschland GmbH, 2023.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

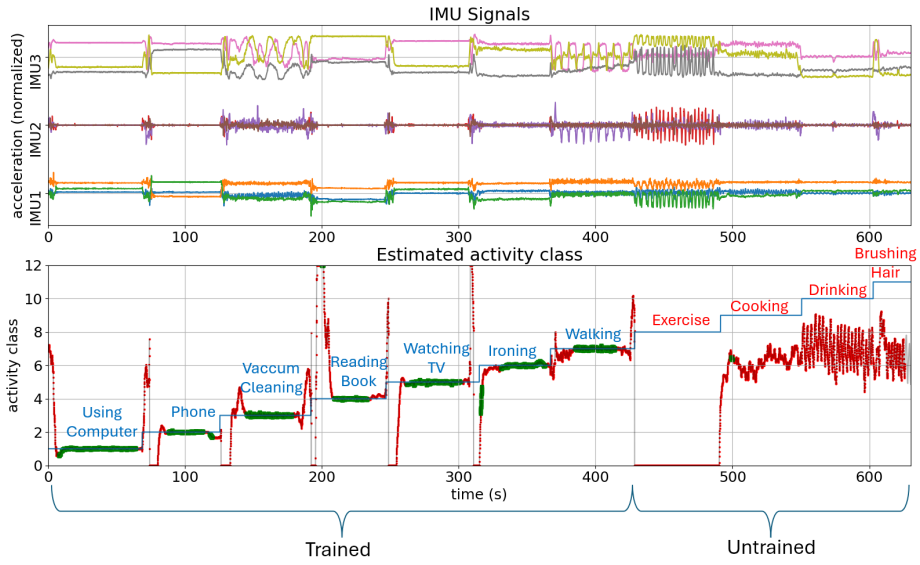

Fig. 3: Results over the full dataset for one individual. Top, raw IMU signals. Bottom, result of prediction and score. Blue line: real class, green: ESN prediction, high confidence, red: ESN prediction, low confidence. ESN output has not been filtered or treated in any way, but it was cropped to the [0,12] interval for representation purposes (valid classes are 1-7). The last 10 seconds of each activity are used for validation. There is a lot of noise in between activities which was omitted for the training process, but not during inference. A better model would have yielded better performance, but nevertheless the experiment is focused on the score system as confidence in the prediction.

[10] H. Jaeger. The "echo state" approach to analysing and training recurrent neural networks. GMD Report 148, GMD - German National Research Institute for Computer Science, 2001.

[11] Herbert Jaeger. Echo state network. *Scholarpedia*, 2(9):2330, 2007.

[12] Sheikh Badar ud din Tahir, Ahmad Jalal, and Mouazma Batool. Wearable sensors for activity analysis using smo-based random forest over smart home and sports datasets. In *2020 3rd International Conference on Advancements in Computational Sciences (ICACS)*, pages 1–6, 2020.

[13] Sheikh Badar Ud Din Tahir, Ahmad Jalal, and Kibum Kim. Wearable inertial sensors for daily activity analysis based on adam optimization and the maximum entropy markov model. *Entropy*, 22(5):579, 2020.

[14] Nathan Trouvain, Luca Pedrelli, Thanh Trung Dinh, and Xavier Hinaut. ReservoirPy: An efficient and user-friendly library to design echo state networks. In *Artificial Neural Networks and Machine Learning – ICANN 2020*, pages 494–505. Springer International Publishing, 2020.

[15] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.