A Global Post hoc XAI Method for Interpreting LSTM Using Deterministic Finite State Automata

Gargi Gupta[1,*,†], M.Atif Qureshi[2,†] and Luca Longo[3]

[1]School of Computer Science, Technological University Dublin, Ireland
[2]eXplainable Analytics Group, Faculty of Bussiness, Technological University Dublin, Ireland
[3]Artificial Intelligence and Cognitive Load Research Lab, Technological University Dublin, Dublin, Ireland

Abstract

We propose a global post-hoc XAI method to interpret Long Short-Term Memory (LSTM) models for univariate time series classification. Our approach integrates Symbolic Aggregate approXimation (SAX) to convert continuous time series into symbolic representations during preprocessing. We then apply k-means clustering to the activated hidden states of the LSTM, from which we extract Deterministic Finite Automata (DFA), which provides a transparent and interpretable explanation of the model's decision-making process. Experiments on synthetic and real-world datasets demonstrate high fidelity between DFA and LSTM, with enhanced interpretability for high-stakes domains like healthcare and power demand forecasting.

Keywords

RNN, interpretability, Explainable AI (XAI), LSTM, Deterministic Finite State Automata (DFA), k-means clustering

1. Introduction

Deep learning, particularly recurrent neural networks (RNNs), has revolutionized sequential data analysis in domains like speech and time series by effectively modelling temporal dependencies [1]. Among RNNs, Long Short-Term Memory (LSTM) networks are widely preferred due to their ability to address the vanishing gradient problem and retain long-term dependencies through memory cells. These features make LSTMs particularly suitable for time series classification tasks. However, their LSTMs categorizes them as "black-box" models, making their decision-making processes challenging to interpret. This lack of transparency hinders their adoption in critical fields such as healthcare [2] and finance, where regulations like GDPR mandate explainable AI (XAI) [3, 4, 5]. Addressing these challenges requires methods that enhance the interpretability of LSTMs without compromising predictive performance.

This study introduces a global post-hoc XAI method that leverages Deterministic Finite Automata (DFA) to provide interpretable insights into LSTM decision-making for univariate time series data. By clustering the hidden states of an LSTM, we extract finite states to represent interpretable transitions, offering a clear understanding of how the model processes sequential data. While prior studies have shown the effectiveness of DFA in explaining RNNs [6, 7, 8], our work extends this approach to LSTMs, focusing specifically on univariate time series. We employ Symbolic Aggregate approXimation (SAX) [9] to transform continuous time series data into symbolic sequences, reducing the complexity while preserving key temporal patterns. This integration of SAX preprocessing with DFA extraction offers a structured and interpretable approach to understanding LSTM behavior [10], particularly in applications like heart rate monitoring, stock price analysis, and power demand forecasting. Key contributions include:

- A novel method for extracting DFA from LSTM models for univariate time series classification.
- Integration of SAX preprocessing with DFA for interpretability, balancing complexity and temporal patterns.

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland
*Corresponding author.
†Science Foundation Ireland, Centre for Research Training in Machine Learning (ML-Labs).
D21125205@mytudublin.ie (G. Gupta); atif.qureshi@tudublin.ie (M.Atif Qureshi); luca.longo@tudublin.ie (L. Longo)

- Quantitative evaluation of fidelity to validate the alignment between DFA and LSTM.

2. Related Work

Explaining the decisions of deep learning models, particularly RNNs like LSTM models, has emerged as a prominent area of research. The challenge of interpreting these "black-box" models has driven the development of XAI methods to enhance transparency and trustworthiness [4]. Interpretability makes abstract model outputs meaningful, while explainability identifies key features influencing predictions in human-understandable terms [11]. Among many ways, XAI methods can be generally categorized into *ante-hoc* and *post-hoc* approaches:

- **Ante-hoc methods** integrate interpretability directly into the model's architecture.
- **Post-hoc methods** generate explanations after predictions and are further divided into:
  - **Local explanations**: Methods like saliency maps, LIME, and attention mechanisms explain individual predictions by highlighting input features affecting outputs [12, 13]. While effective for specific instances, these methods struggle to capture global decision-making.
  - **Global explanations**: These provide a holistic view of a model's behavior across datasets, making them essential for understanding state transitions and decision processes [14].

Although attention mechanisms and saliency maps identify critical features, they cannot explain how LSTMs process sequential data or transition between states. Symbolic representations like DFA address this gap by offering a structured, interpretable global explanation of state transitions [6]. DFA extraction visualizes LSTM decision-making as finite state machines, providing a broader perspective on model behavior [7, 8]. This study extends DFA-based interpretability to univariate time series, integrating SAX preprocessing [9] to reduce data complexity while preserving temporal patterns. The SAX-DFA combination provides symbolic, global post-hoc explanations for LSTMs, particularly in applications like power demand forecasting and heart rate monitoring [10]. Unlike local methods, DFA captures high-level transitions, offering a comprehensive understanding of LSTM behaviour. Table 1 compares common interpretability techniques for time series and LSTMs, highlighting their strengths and limitations. SAX-DFA addresses key gaps by offering structured, global explanations of state transitions. Besides qualitative metrics of explainability [15], more objective metrics such as *fidelity* and *robustness* are crucial for evaluating XAI methods. Fidelity ensures explanations align with the original model's predictions, while robustness examines consistency across data points [16, 17]. These metrics guide the development of reliable, interpretable models. Existing XAI techniques for time series have underexplored global interpretability, creating an opportunity for SAX-DFA to address these gaps.

Table 1
Comparison of Interpretability Techniques for Time Series and LSTMs

Method	Scope	Strengths	Limitations
Attention Mechanisms [13]	Local	Highlights important input regions for predictions.	Limited global insights into temporal dependencies.
Saliency Maps [13]	Local	Visualizes input-output sensitivity.	Noise-sensitive; lacks state-transition insights.
LIME [12]	Local	Simplifies predictions with surrogate models.	Explanations may not generalize globally.
SAX-DFA (Proposed)	Global	Structured, symbolic insights into transitions.	Requires careful SAX bin and clustering tuning.

This paper contributes to advancing global post-hoc interpretability for LSTMs by combining SAX with DFA. This approach generates state-transition explanations that accurately mirror the LSTM's decision-making process while maintaining high fidelity and robustness. Unlike local methods, SAX-DFA provides a dataset-wide perspective, enabling insights into temporal patterns and transitions.

3. Methodology

This section introduces the proposed *post-hoc* XAI method to explain the internal decision-making processes of LSTM models trained on univariate time series data. The methodology is organized into three phases, which align with the four major steps shown in the pipeline (Figure 1).

3.1. Phase 1: Preprocessing and Model Training

This phase involves transforming time series into symbolic sequences and training the LSTM model.

SAX Preprocessing SAX discretizes continuous time series into symbolic sequences, reducing complexity while preserving critical temporal patterns. We evaluated SAX bin sizes ($N_{\text{bin}} = 3, 5, 7$) and found a balance between interpretability and fidelity at $N_{\text{bin}} = 5$. The quantile strategy was chosen for its ability to ensure balanced symbol distribution, which improved generalization and DFA interpretability.

LSTM Training The SAX-encoded sequences were used to train an LSTM model. Hyperparameters, such as hidden size $(16, 32, 64)$ and layers (1 or 2), were tuned for optimal performance (Table 3). The Adam optimizer ($lr = 0.0001$) and cross-entropy loss were used, with early stopping to prevent overfitting. These settings ensured accurate classification while preserving the "black-box" nature of the LSTM for interpretability experiments.

3.2. Phase 2: DFA Extraction

This phase transforms the trained LSTM's hidden state dynamics into a DFA for interpretability. A *DFA* is formally defined as a 5-tuple: $\mathcal{A} = (Q, \Sigma, S, F, \delta)$ where:

- Q is the finite set of *states*.
- Σ is the *alphabet*, i.e., the set of symbols generated from the SAX preprocessing.
- $S \in Q$ is the *start state*, representing the initial state.
- $F \subseteq Q$ is the set of *accepting states* linked to classification outcomes.
- $\delta : Q \times \Sigma \to Q$ is the *transition function*, for state changes based on input symbols.

The LSTM's activated hidden states are extracted at each time step and visualized using t-SNE for dimensionality reduction. K-means clustering groups these states into K clusters, which form the finite states of the DFA. Transition frequencies between clusters, triggered by SAX symbols, are recorded in a transition matrix T, defined as: $T(i, j) = \arg\max_k \ Ns_j(i, k)$ where $T(i, j)$ represents the state transition from i upon input of symbol s_j. Clusters most frequently visited at the end of sequences are designated as the DFA's accepting states, providing an interpretable representation of LSTM decision-making.

3.3. Phase 3: Fidelity Evaluation

In the final phase, the fidelity of the DFA is quantitatively evaluated to assess its accuracy in replicating the LSTM's classifications. Fidelity measures the proportion of instances where the DFA and LSTM agree on their classification outcomes.

4. Experimental Settings

We use synthetic and real-world datasets to evaluate the proposed method, employing preprocessing, model training, and evaluation strategies.

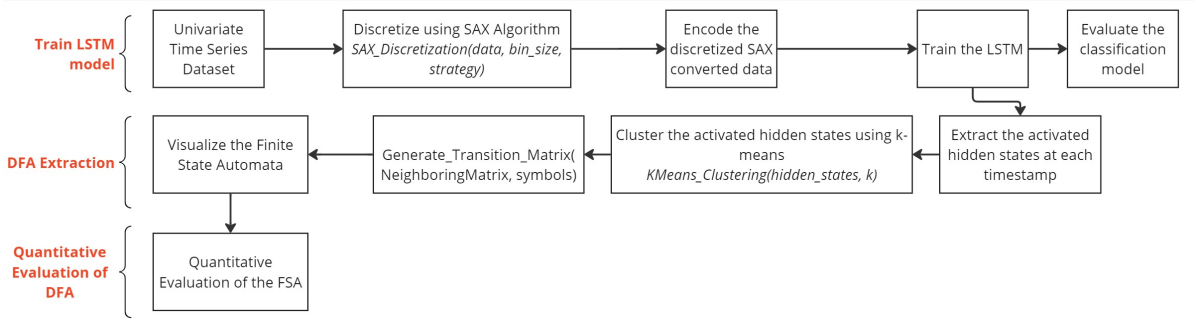


Figure 1: Pipeline of the proposed XAI method. The steps—(1) SAX preprocessing, (2) LSTM training, (3) hidden state clustering, and (4) DFA construction—are organized into three phases for clarity.

Table 2
Summary of Datasets Used in the Experiment

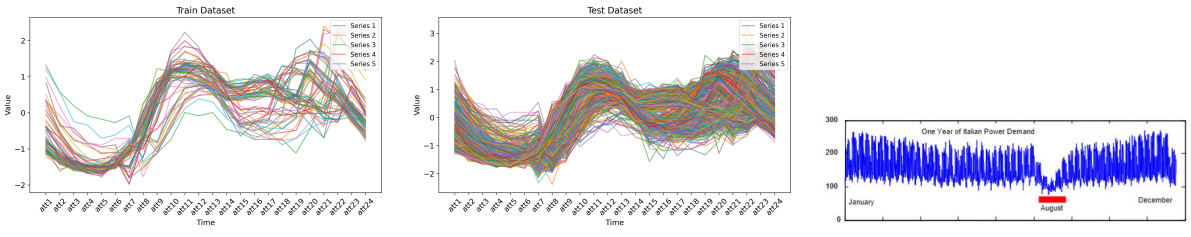
Name	Data Size	No. Classes	Length
Synthetic Dataset	1000	2	50
Italy Power Demand	1096	2	24

4.1. Datasets

We evaluated the proposed method on synthetic and real-world datasets, as described below, with a summary available in Table 2.

Synthetic Noisy Sine Wave Dataset A synthetic noisy sine wave dataset was created to simulate real-world time series complexity. It contains 1000 points divided into 20 sequences of 50 time steps each. Sine waves were generated with added Gaussian and low-frequency noise, labeled as Class 1 if the maximum amplitude exceeds 0.5 and Class 0 otherwise. This dataset offers a controlled environment for benchmarking DFA interpretability against ground-truth metrics [18].

Italy Power Demand Dataset The Italy Power Demand dataset [19] records daily electrical power demand for colder (October–March) and warmer (April–September) months. With 67 training and 1029 testing instances of 24 time steps each, it serves as a benchmark for time series classification. Figure 2 shows the dataset's distribution. A summary is provided in Table 2.



(a) Train Dataset (b) Test Dataset (c) Combined Dataset

Figure 2: Italy Power Demand dataset grouped by the training and testing sets and the combined data.

4.2. LSTM Training

LSTM models were implemented in PyTorch and trained using SAX-encoded sequences. The architecture was optimized through hyperparameter tuning, varying hidden units $(16, 32, 64)$, layers $(1, 2)$, and SAX bin sizes[1] $(3, 5, 7)$ for the quantile strategy. These parameters were selected based on empirical observations to balance interpretability and fidelity. Models were trained using a 70/15/15 split for

[1]The choice of SAX bin sizes was guided by empirical observations and the need to balance interpretability and fidelity

training, validation, and testing, with early stopping applied to avoid overfitting. Table 3 lists the key hyperparameters.

Table 3
A summary of the LSTM architecture's hyperparameters

Hyperparameter	Values
Bin Size (SAX)	3, 5, 7
Input Size	1 (Univariate)
Hidden Size	16, 32, 64
Output Size	2 (Binary)
Layers	1, 2
Optimizer	Adam (LR: 0.0001)
Epochs	50

4.3. Learning DFA

After training the LSTM model on the SAX-encoded time series, k-means clustering was applied to the activated hidden states at each time step. These clusters represent the finite set of states for constructing the DFA. In this study, the DFA was built using the validation set to ensure it captures the model's general behaviour and decision patterns, allowing for interpretable insights into the LSTM's performance on unseen data. The **Neighbouring Matrix** Ns tracks transition frequencies between clusters for each symbol in the sequence, where each entry $Ns(i,k)$ represents the number of transitions from state i to state k based on the input symbol.The transition matrix T is then derived as follows: $T(i,j) = \arg\max_k \ Ns_j(i,k)$ Here, $T(i,j)$ represents the state transition from state i when inputting the symbol s_j. The DFA's accepting states were determined by identifying the state most frequently visited at the end of each sequence during LSTM processing. The state with the highest occurrence for a given sequence was marked as the final accepting state.

4.4. Evaluation Metrics

The proposed method was evaluated using:

- **Accuracy:** Proportion of correctly classified instances.
- **Macro and Micro F1 Scores:** Macro F1 measures the unweighted average of class-specific F1 scores, while Micro F1 weights scores by class prevalence.
- **DFA-LSTM Fidelity:** Proportion of instances where the DFA matches LSTM predictions, measuring how well the DFA replicates LSTM behavior (Table 4).

Table 4
Quantitative evaluation metrics for DFA-based explanations

Metric	Definition	Formula
DFA-LSTM Fidelity	Ratio of instances where the DFA agrees with the LSTM model, divided by the total number of validation sequences.	F = $\frac{a}{N}$, where a is the number of instances where the DFA agrees with LSTM, and N is the total number of sequences.

5. Results and Discussions

The proposed global post hoc XAI method was evaluated on synthetic and real-world datasets (Italy Power Demand) to assess its ability to mimic the inferential process of LSTM models trained on univariate

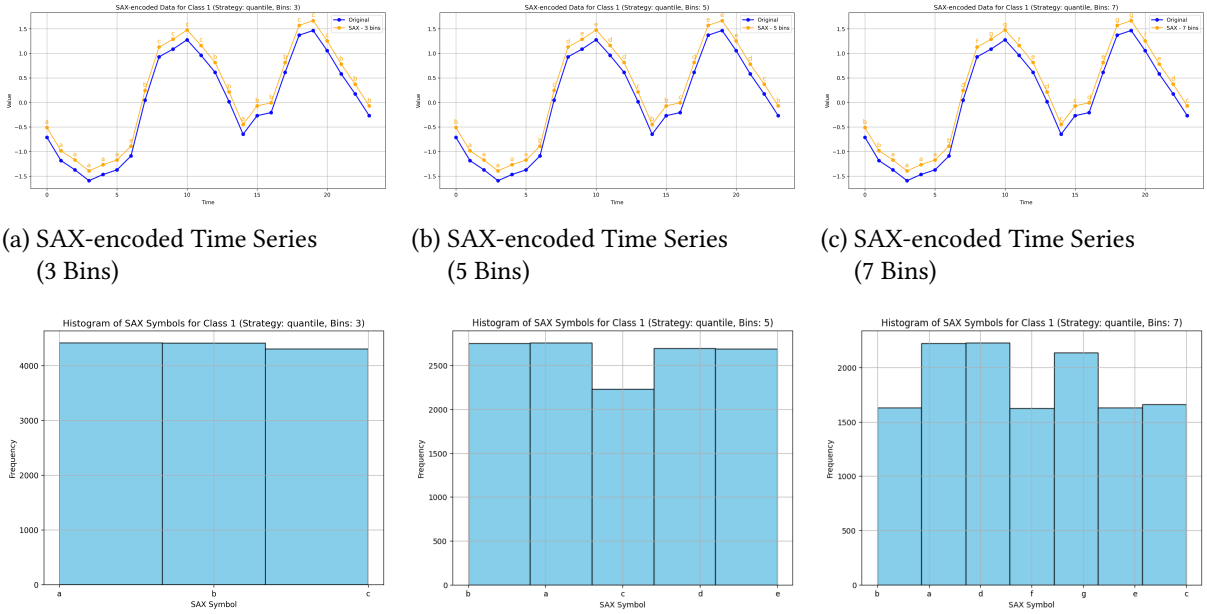
time series data. Key results focus on SAX preprocessing, LSTM performance, DFA extraction, and quantitative fidelity evaluation.

5.1. Impact of SAX Preprocessing

SAX algorithm transformed continuous time series into symbolic representations with varying bin sizes (3, 5, 7) and strategies (quantile, uniform, and normal).

The quantile strategy outperformed alternatives by ensuring balanced symbol distributions, particularly in datasets with skewed patterns or outliers. This balance improved LSTM classification accuracy and DFA interpretability by preserving key temporal features.

Figure 3 demonstrates how different bin sizes influence SAX-encoded sequences and symbol distributions. Bin size $N_{\text{bin}} = 5$ emerged as optimal, balancing fine-grained temporal representation with interpretability. Larger bin sizes ($N_{\text{bin}} = 7$) coarsened data patterns, while smaller bins ($N_{\text{bin}} = 3$) increased complexity without significant performance gains.



(a) SAX-encoded Time Series (3 Bins)

(b) SAX-encoded Time Series (5 Bins)

(c) SAX-encoded Time Series (7 Bins)

(d) Symbol Distribution (3 Bins)

(e) Symbol Distribution (5 Bins)

(f) Symbol Distribution (7 Bins)

Figure 3: Effect of SAX bin size on time series encoding and symbol distribution for class 1 of the Italy Power Demand dataset. The figure shows how different bin sizes (3, 5, 7) in the SAX algorithm, using the quantile strategy, affect the representation of time series data and the resulting symbol distributions. Subfigures (a), (c), and (e) depict the SAX-encoded time series, while subfigures (b), (d), and (f) show the corresponding symbol distribution histograms.

5.2. LSTM Performance Across Configurations

LSTM models were trained with varying configurations of hidden units (16, 32, 64), layers (1, 2), and SAX bin sizes (3, 5, 7). Performance was evaluated using accuracy and F1 scores (Tables 5, 6, 7).

Key findings include:

- A hidden size of 32 and one layer consistently achieved high accuracy across datasets.
- SAX bin size $N_{\text{bin}} = 5$ provided the best trade-off, with peak test accuracy of 96.36% on the Italy Power Demand dataset.
- Larger bin sizes ($N_{\text{bin}} = 7$) improved DFA-LSTM fidelity by capturing broader temporal patterns, while smaller bins ($N_{\text{bin}} = 3$) risked overfitting.

Table 5
Performance of LSTM Model with SAX Bin Size = 3; Dataset- Italy Power Demand

Hidden Size	Num of Layers	Test Accuracy	Test Macro F1 Score	Test Micro F1 Score
16	1	0.9636	0.9631	0.9636
16	2	0.9576	0.9568	0.9576
32	1	0.9576	0.9568	0.9576
32	2	0.9455	0.9446	0.9455
64	1	0.9394	0.9387	0.9394
64	2	0.9576	0.9568	0.9576

Table 6
Performance of LSTM Model with SAX Bin Size = 5 ; Dataset- Italy Power Demand

Hidden Size	Num of Layers	Test Accuracy	Test Macro F1 Score	Test Micro F1 Score
16	1	0.9394	0.9386	0.9394
16	2	0.9515	0.9507	0.9515
32	1	0.9636	0.9630	0.9636
32	2	0.9636	0.9630	0.9636
64	1	0.9515	0.9506	0.9515
64	2	0.9576	0.9568	0.9576

Table 7
Performance of LSTM Model with SAX Bin Size = 7; Dataset- Italy Power Demand

Hidden Size	Num of Layers	Test Accuracy	Test Macro F1 Score	Test Micro F1 Score
16	1	0.9515	0.9512	0.9515
16	2	0.9576	0.9574	0.9576
32	1	0.9576	0.9574	0.9576
32	2	0.9576	0.9574	0.9576
64	1	0.9636	0.9635	0.9636
64	2	0.9515	0.9512	0.9515

These results underscore the importance of hyperparameter tuning to balance interpretability and classification performance.

5.3. DFA Extraction and Fidelity Evaluation

The DFA extraction process leveraged k-means clustering to translate LSTM hidden states into interpretable states, with transitions defined by the SAX-encoded inputs. This subsection presents the results of clustering, transition matrix construction, and fidelity evaluation, focusing on the insights derived from the experiments.

5.3.1. Clustering and Transition Matrix Insights

The clustering of LSTM hidden states, visualized through t-SNE (Figure 4), revealed distinct clusters for the training and validation sets, reflecting effective learning of temporal dependencies. However, the test set clusters were more dispersed, indicating challenges in generalizing to unseen patterns. These observations underline the LSTM's ability to model temporal patterns while highlighting potential areas for improving its robustness.

The number of clusters (k) played a pivotal role in DFA complexity and interpretability. For the Italy Power Demand dataset, $k = 6$ provided an optimal balance between granularity and simplicity.

Transition probabilities were derived from SAX input sequences, and the most frequent transitions were mapped into a transition matrix. Larger k values captured finer-grained transitions but risked overfitting, while smaller k values offered simpler DFA representations but omitted subtle temporal dynamics.

5.3.2. Impact of SAX Bin Size on Fidelity

The SAX bin size (N_{bin}) significantly influenced the fidelity between the extracted DFA and the LSTM model. Table 8 summarizes fidelity scores across configurations:

- Larger bin sizes ($N_{\text{bin}} = 7$) achieved the highest fidelity of 0.5854, as the DFA could effectively capture broader temporal patterns while minimizing overfitting. These configurations were particularly effective for the Italy Power Demand dataset, demonstrating a robust approximation of LSTM behavior.
- Smaller bin sizes ($N_{\text{bin}} = 3$) yielded lower fidelity scores, such as 0.4756, due to coarser discretization. This reduction in detail limited the DFA's ability to distinguish between state transitions, particularly for datasets with complex temporal patterns.
- Misalignment between k and N_{bin}, where $k > N_{\text{bin}}$, introduced invalid state transitions and decreased fidelity. This mismatch was evident in configurations with $N_{\text{bin}} = 3$ and $k = 6$, where fidelity improved moderately (0.5305) but remained suboptimal.

5.3.3. Evaluation of DFA Visualizations

Figure 5 illustrates the DFA visualizations for different SAX bin sizes and k values. Configurations with $N_{\text{bin}} = 5$ and $k = 4$ provided interpretable representations, focusing on dominant state transitions. Larger bin sizes ($N_{\text{bin}} = 7$) resulted in more refined DFA structures, effectively capturing the primary decision patterns while minimizing redundancy. The absence of certain transitions in these configurations underscores the DFA's ability to generalize and avoid overfitting.

These visualizations demonstrate the utility of the SAX-DFA method for interpreting LSTM behavior. For example, DFA structures with larger bin sizes prioritized meaningful transitions, simplifying the explanation process without sacrificing fidelity. The reduction in noise and unnecessary state transitions improved the clarity of the extracted DFA, particularly for datasets with well-defined temporal patterns.

5.3.4. Comparison Across Datasets

The fidelity scores and visualizations reveal consistent trends across both datasets:

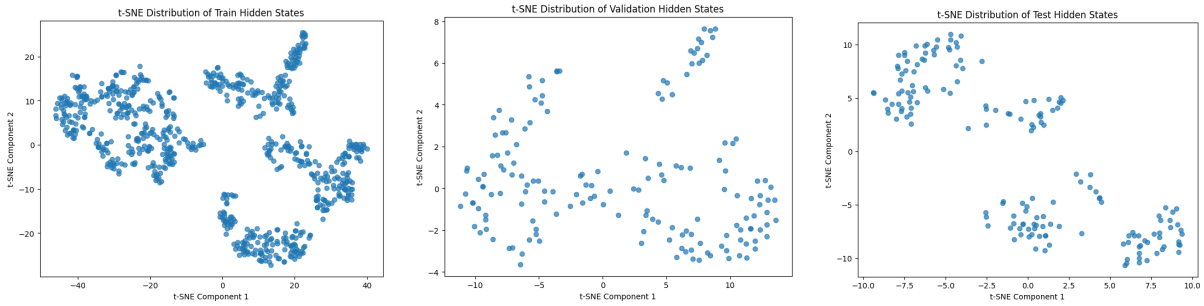
- For the Italy Power Demand dataset, $N_{\text{bin}} = 7$ and $k = 6$ achieved the highest fidelity, effectively capturing the LSTM's temporal dynamics. These configurations also provided the most interpretable DFA visualizations, balancing complexity and accuracy.
- The synthetic sine wave dataset, despite its smaller size, demonstrated similar trends. However, the reduced temporal complexity led to slightly lower fidelity scores for smaller bin sizes. For instance, configurations with $N_{\text{bin}} = 3$ and $k = 3$ achieved comparable fidelity (0.5305) but struggled to generalize transitions for noisier sequences.

5.3.5. Discussion and Limitations

The results highlight the strengths and limitations of the SAX-DFA method:

- **Strengths:** The method consistently achieved high fidelity with larger bin sizes and appropriate clustering. DFA visualizations offered interpretable insights into the LSTM's decision-making, particularly for datasets with well-defined temporal patterns.
- **Limitations:** Smaller bin sizes reduced fidelity, and misalignment between k and N_{bin} introduced invalid transitions. Addressing these challenges requires refining clustering techniques and balancing SAX parameters to improve DFA accuracy and scalability.

Future work will explore advanced clustering methods, such as k-means++, to improve alignment between clusters and SAX bins. Additionally, extending the methodology to multivariate datasets will test its scalability and robustness across more complex scenarios.



(a) Training set (b) Validation set (c) Test set

Figure 4: t-SNE visualizations of the activated hidden states extracted from an LSTM model with 2 layers and 64 hidden units for (a) the training set, (b) the validation set, and (c) the test set of the Italy Power Demand dataset. The plots show clustering patterns in 2D: well-defined clusters in the training set indicate effective learning, condensed clusters in the validation set suggest good generalization, and dispersed clusters in the test set highlight challenges with unseen data.

Conversely, smaller bin sizes, such as $N_{\text{bin}} = 3$, struggled to achieve high fidelity scores, with some configurations yielding values as low as 0.4756. This limitation likely stems from the coarse discretization of time series data, which reduces the DFA's capacity to distinguish between state transitions effectively.

Table 8
DFA-LSTM Fidelity Results Across Different Configurations

Bin Size (N_{bin})	Hidden Size	Num Layers	Cluster Size (K)	DFA-LSTM Fidelity
3	16	1	6	0.5305
5	16	2	6	0.5244
7	16	1	6	0.5854
3	64	2	6	0.4756

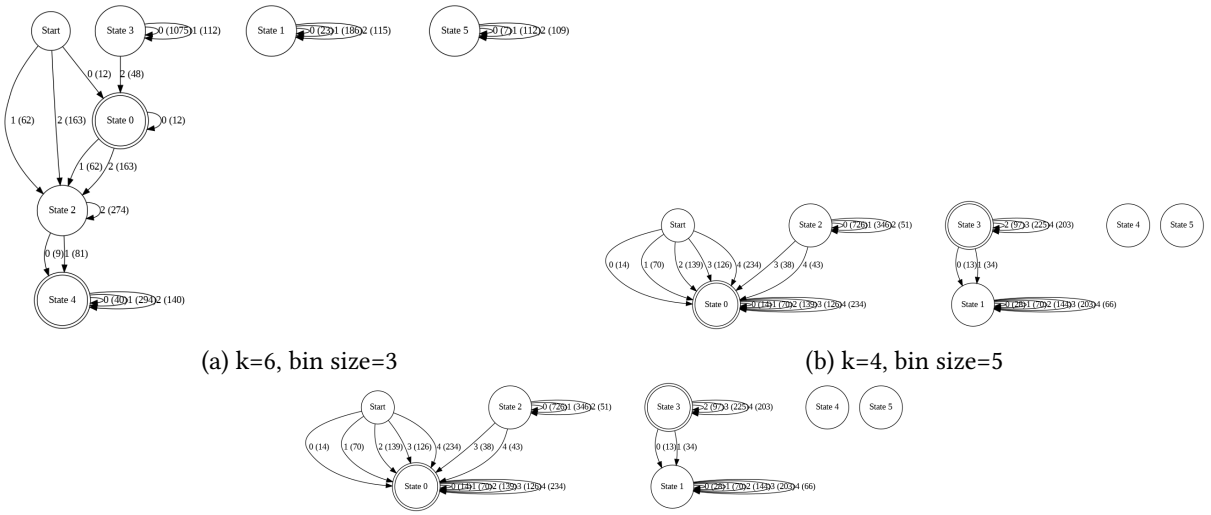
6. Conclusion and Future Work

The need for interpretability in complex machine learning models, particularly in high-stakes domains like healthcare and finance, has driven the development of XAI techniques. This paper presented a novel global post hoc XAI method that integrates SAX preprocessing with DFA extraction to provide interpretable insights into the decision-making processes of LSTM networks.

Our proposed method addresses key challenges in explaining LSTM behavior by transforming high-dimensional hidden state representations into a symbolic, interpretable structure. The DFA, constructed through clustering and state transition analysis, offers a clear visualization of temporal dependencies and decision patterns within LSTM models. This work contributes to the growing body of research in XAI by emphasizing global interpretability, a crucial aspect for understanding the overall behavior of models applied to univariate time series data.

6.1. Summary of Findings

Experiments on both synthetic and real-world datasets demonstrated the effectiveness of the SAX-DFA method:



(a) k=6, bin size=3

(b) k=4, bin size=5

(c) k=4, bin size=7

Figure 5: (a) DFA extracted from an LSTM trained on the Italy Power Demand dataset with SAX-encoded data, using a bin size of 3 and $k = 6$ for k-means clustering, consists of 6 states and transitions based on SAX input symbols. The double-circled state represents the accepting state, indicating the final classification decision. (b) DFA extracted using a bin size of 5 and $k = 4$, showing 4 states and structural changes due to the different bin size. (c) DFA extracted using a bin size of 7 and $k = 4$, also with 4 states. The reduced transitions highlight the model's focus on key state transitions, preventing overfitting.

- **High Fidelity:** The method achieved a maximum DFA-LSTM fidelity score of 0.5854 for the Italy Power Demand dataset with $N_{\text{bin}} = 7$, $k = 6$, and hidden size = 16. This indicates a strong alignment between the DFA and LSTM decision-making processes.
- **Interpretable Visualizations:** DFA structures provided clear insights into state transitions, capturing dominant decision patterns while minimizing noise. Larger SAX bin sizes produced refined DFA representations, balancing complexity and interpretability.
- **Generalizability Across Datasets:** The method effectively captured LSTM decision patterns for both datasets, demonstrating its robustness across different types of univariate time series data.
- **Parameter Sensitivity:** SAX bin sizes and cluster counts significantly influenced fidelity and interpretability. Larger bin sizes captured broader patterns but required careful tuning to avoid over-simplification.

These findings highlight the SAX-DFA method's potential to enhance transparency in LSTM models while maintaining high performance in time series classification tasks.

6.2. Key Contributions

This study advances the field of explainable AI through:

1. Development of a novel pipeline combining SAX preprocessing and DFA extraction for LSTM interpretability.
2. Quantitative evaluation of DFA fidelity as a metric to assess the alignment between DFA and LSTM decision-making.
3. Comprehensive analysis of SAX and clustering parameters, offering insights into their impact on fidelity and interpretability.
4. Application to both synthetic and real-world datasets, demonstrating the method's generalizability and scalability.

6.3. Limitations and Challenges

While the proposed method showed promising results, certain limitations warrant further attention:

- **SAX Parameter Sensitivity**: The choice of bin size and clustering parameters heavily influenced fidelity and interpretability. Suboptimal configurations reduced the DFA's ability to capture essential temporal patterns.
- **Handling of Invalid Transitions**: Mismatches between SAX bin sizes and cluster counts introduced invalid state transitions, impacting classification accuracy and fidelity.
- **Scalability**: The method was validated on univariate datasets. Its applicability to multivariate time series and datasets with longer sequences remains unexplored.
- **Local Interpretability Gap**: While the DFA provides global explanations, it lacks mechanisms for understanding specific individual predictions, which are crucial in certain domains.

6.4. Future Work

Building on the foundation of this study, future work will address these limitations and explore new directions:

1. **Parameter Optimization**: Advanced clustering techniques, such as k-means++ or hierarchical clustering, will be explored to better align SAX bin sizes and cluster counts, improving fidelity and handling invalid transitions.
2. **Scalability to Multivariate Datasets**: The methodology will be extended to handle multivariate time series, involving adaptations to SAX preprocessing and DFA construction to accommodate higher-dimensional data.
3. **Integration of Local Explanations**: Hybrid approaches combining SAX-DFA with local interpretability methods, such as SHAP, saliency maps, or attention mechanisms, will provide a comprehensive framework for both global and local explanations.
4. **Application to High-Stakes Domains**: The SAX-DFA method will be applied to real-world datasets in domains like healthcare (e.g., ECG analysis) and finance (e.g., stock price prediction) to evaluate its practical utility in critical decision-making scenarios [2].
5. **Enhanced DFA Structures**: The DFA framework will be refined to incorporate probabilistic transitions or weighted state connections, enabling a more nuanced representation of LSTM behavior.
6. **Evaluation Metrics**: Additional evaluation metrics, such as robustness and completeness, will be incorporated to assess the quality and reliability of explanations under varying conditions.
7. **User-Centric Evaluation**: Future research will focus on usability studies involving domain experts to evaluate the interpretability and utility of DFA visualizations in real-world applications.

6.5. Concluding Remarks

In conclusion, the SAX-DFA method represents a significant step toward achieving global interpretability for LSTM models in time series classification. By combining symbolic representations with finite automata, this method offers a structured, interpretable view of complex decision-making processes. While challenges remain, the findings of this study provide a robust foundation for advancing explainable AI methodologies and fostering trust in machine learning models across diverse applications.

Acknowledgments

This work was funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning(18/CRT/6183).

References

[1] T. Rojat, R. Puget, D. Filliat, J. Del Ser, R. Gelin, N. Díaz-Rodríguez, Explainable artificial intelligence (xai) on timeseries data: A survey, arXiv preprint arXiv:2104.00950 (2021).

[2] N. A. Aziz, A. Manzoor, M. D. Mazhar Qureshi, M. A. Qureshi, W. Rashwan, Explainable ai in healthcare: Systematic review of clinical decision support systems, medRxiv (2024). URL: https://www.medrxiv.org/content/early/2024/08/10/2024.08.10.24311735. doi:10.1101/2024.08.10.24311735.

[3] G. Vilone, L. Longo, A quantitative evaluation of global, rule-based explanations of post-hoc, model agnostic methods, Frontiers in Artificial Intelligence 4 (2021) 160. URL: https://www.frontiersin.org/article/10.3389/frai.2021.717899. doi:10.3389/frai.2021.717899.

[4] P. Hacker, J. Cordes, J. Rochon, Regulating gatekeeper artificial intelligence and data: Transparency, access and fairness under the digital markets act, the general data protection regulation and beyond, European Journal of Risk Regulation 15 (2024) 49–86.

[5] S. Wang, M. A. Qureshi, L. Miralles-Pechuán, T. Huynh-The, T. R. Gadekallu, M. Liyanage, Explainable ai for 6g use cases: Technical aspects and research challenges, IEEE Open Journal of the Communications Society 5 (2024) 2490–2540. doi:10.1109/OJCOMS.2024.3386872.

[6] C. L. Giles, C. B. Miller, D. Chen, H.-H. Chen, G.-Z. Sun, Y.-C. Lee, Learning and extracting finite state automata with second-order recurrent neural networks, Neural Computation 4 (1992) 393–405.

[7] R. Marzouk, On computability, learnability and extractability of finite state machines from recurrent neural networks, arXiv preprint arXiv:2009.06398 (2020).

[8] I. C. Kaadoud, N. P. Rougier, F. Alexandre, Knowledge extraction from the learning of sequences in a long short term memory (lstm) architecture, Knowledge-Based Systems 235 (2022) 107657.

[9] J. Faouzi, Symbolic aggregate approximation (sax) example - pyts documentation, 2024. URL: https://pyts.readthedocs.io/en/latest/auto_examples/approximation/plot_sax.html, accessed September 20, 2024.

[10] N. Tabassum, S. Menon, A. Jastrzębska, Time-series classification with safe: Simple and fast segmented word embedding-based neural time series classifier, Information Processing & Management 59 (2022) 103044.

[11] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, S. Stumpf, Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions, Information Fusion 106 (2024) 102301. URL: https://www.sciencedirect.com/science/article/pii/S1566253524000794. doi:https://doi.org/10.1016/j.inffus.2024.102301.

[12] B. Shickel, P. Rashidi, Sequential interpretability: methods, applications, and future direction for understanding deep learning models in the context of sequential data, arXiv preprint arXiv:2004.12524 (2020).

[13] Y.-J. Jung, S.-H. Han, H.-J. Choi, Explaining cnn and rnn using selective layer-wise relevance propagation, IEEE Access 9 (2021) 18670–18681.

[14] B.-J. Hou, Z.-H. Zhou, Learning with interpretable structure from gated rnn, IEEE transactions on neural networks and learning systems 31 (2020) 2267–2279.

[15] G. Vilone, L. Longo, Development of a human-centred psychometric test for the evaluation of explanations produced by xai methods, in: L. Longo (Ed.), Explainable Artificial Intelligence, Springer Nature Switzerland, Cham, 2023, pp. 205–232.

[16] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, ACM Computing Surveys 55 (2023) 1–42.

[17] G. Vilone, L. Longo, Notions of explainability and evaluation approaches for explainable artificial intelligence, Information Fusion 76 (2021) 89–106.

[18] Y. Liu, S. Khandagale, C. White, W. Neiswanger, Synthetic benchmarks for scientific research in explainable machine learning, arXiv preprint arXiv:2106.12543 (2021).

[19] E. Keogh, the UCR Time Series Classification Archive, Ucr time series classification archive, 2018. URL: https://www.cs.ucr.edu/~eamonn/time_series_data_2018/, accessed: 2024-09-12.