

# Post hoc explanations for RNNs using state transition representations for time series data

Gargi Gupta<sup>1,2</sup>

<sup>1</sup>*School of Computer Science, Technological University Dublin, Ireland*

<sup>2</sup>*Artificial Intelligence and Cognitive Load Research Lab, Technological University Dublin, Ireland*

## Abstract

The interpretability of deep learning models has gained the attention of many researchers and organisations in recent years. The doctoral research will include the interpretability and explainability of Recurrent Neural Networks (RNN) and its variant architectures. RNNs find extensive application in speech recognition and healthcare, yet their complex architecture often renders them as "black box models". FSA and RNN both handle sequential data, with FSA transitioning between states and RNN updating hidden states at each step. This similarity enables comparing FSA's transitions to RNN's hidden state updates. FSA's graphical representation enhances human understanding. The findings of this study will be used to lead the development of post hoc explanation methods (XAI) for time series data employing deterministic finite state machines as an explainability method.

## Keywords

RNN, LSTM, Explainability, XAI, DFA, state transition representations, post-hoc

## 1. Introduction and Motivation

With the rapid advancement of AI technologies, artificial intelligence (AI) has seamlessly integrated into our daily lives. The demand for transparent explanations in AI, deep learning, and machine learning is growing. AI's reach extends to healthcare, finance, computer vision, cybersecurity, education, and the judiciary. In this landscape, recurrent neural networks (RNNs) are essential, powering various applications across natural language processing, speech recognition, and image analysis [1]. Despite their high accuracy, RNNs often remain "black boxes" due to their complex internal processes, hindering comprehension of their predictions [2]. The lack of interpretability is a significant barrier to RNN adoption in practical domains, aligning with the growing need for transparency. The 'right to explanation' concept was introduced in the General Data Protection Regulation (GDPR) by the European Union in 2016 (effective in 2018). This provision grants individuals the right to seek explanations for decisions that impact their lives. The primary goal of this right is to promote transparency in automated processes [3, 4]. Interpretability methods vary by data type, global/local focus, and user expertise. To enhance the interpretability and explainability of Neural Networks and generate something that is indeed understandable to humans, some researchers have used the notion of 'rule' [5?


---

*Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal*

✉ D21125205@mytudublin.ie (G. Gupta)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

] or the notion of ‘arguments’ and their conflictual nature [6] or a mix of them [7]. Other researchers propose approaches like finite-state automata (FSA) to enhance RNN transparency, while recent advances highlight limited attention on time series models compared to vision or language domains [8, 9]. In my doctoral research, deterministic finite state automata (FSA) will be employed as an XAI method for post hoc explanations of RNNs and their variants such as LSTM, MGU, and GRU. This approach is especially valuable when applied to time series data due to the inherent interpretability of automata and their enhanced comprehensibility through graphical representations. A key goal of this research is to explore the evaluation of FSA as an XAI technique. This proposed methodology employs the k-means clustering algorithm to learn finite state automata from diverse RNN architectures. This paper is structured as follows. Section 2 introduces the finite state automata as XAI methods and background study. The proposed research plan, the approach used and the progress achieved to date are discussed in Section 3. Future directions and identifying research gaps for this research project are discussed in Section 4.

## 2. Related Work

State transition representations such as finite state automata have been studied in depth for a long and are often seen as interpretable models [10, 11]. Authors in their study [12] discuss that extracting knowledge from the network implies a representation form. These forms can be finite state automata, fuzzy automata, Markov chains and differential equations. The relationship between recurrent neural networks  $R$  and finite state machine  $M$  and its ability to mimic the inner workings of RNNs have been under study since recent years [13, 14]. The study [15] highlights the concept of rule extraction from RNNs ( $R$ ), which refers to finding models of the underlying RNN, typically in the form of the finite state machine, that mimics the network to a satisfactory degree while having the advantage of being more transparent. A study by [16] states that the activation’s of state units represent past histories, and clusters of these activation’s can represent the states of the generating automaton. The aspects of automata that make them more interpretable are discussed in the study [11], which states that automata can be easily represented graphically as directed, cyclic, labelled graphs that provide a hierarchical view of sequential data. Additionally, it is transparent to compute, and manual verification’s are also possible. Automata are generative. Also, automata are well studied in theory and practice, including composition and closure properties, sub-classes, and related equally expressive formalisms. This makes it easy for humans to transfer their knowledge onto it. The model is frequently used in system design to describe system logic and is accessible to a wide audience. In [17], binary classification experiments on artificial and movie review datasets revealed FSA’s superiority over the learned RNN for prediction explanations. The study emphasized the crucial role of RNN gates in FSA learning, advocating FSA’s potential to enhance transparency and trust, an important and old computational concept [18]. It was also discussed that the gates in RNN play an important role in learning an FSA. The study by [19] investigates the link between first-order recurrent neural networks and deterministic finite state automata. The authors introduce a neural network architecture capable of approximating discrete-time, time-invariant dynamic systems. They train these networks to classify strings based on grammar membership. The study reveals

the alignment between finite state automata and neural networks, highlighting two important learning stages. It also notes that, given extended training, neural networks can emulate finite state automata. Upon conducting a literature review, it appears that research predominantly focuses on binary classification tasks involving text, numeric, image, and tabular datasets. Little attention has been given to multi-class classification for time series data [9, 20, 21].

### 3. Research Objective and Research Questions

This section outlines our research objectives, research questions, and proposed approach. Our goal is to explore the utility of finite state automata (FSA) as a post hoc explainability technique for recurrent neural networks (RNNs) and their variants (LSTM, GRU, MGU) in multi-class time series classification. Our study assesses explanation effectiveness, interpretability, and user trust through diverse clustering algorithms, evaluation criteria, and human perception [22]. Given the above, the following key research questions have been identified:

- **RQ 1** How do RNNs and different variants of RNN architectures impact the extraction and learning of finite state automata as a post hoc explainability method?
- **RQ 2** How do different clustering algorithms used to cluster the activated hidden states of RNNs impact the learning and extraction of finite state automata from the RNNs and their different architectures?
- **RQ 3** What evaluation parameters can be used to evaluate the quality of clusters formed which represent the states of finite state automata (FSA), and what standard methods can be used to evaluate the explanations created by FSA for RNNs for time series data?

#### 3.1. Proposed Approach

This section will discuss the proposed approach to address the research questions identified. This study aims to provide a general method to learn deterministic finite state automata from RNN and its variant architectures, such as LSTM, GRU and MGU, on time series data. The method of extraction and learning deterministic finite state automata proposed in this work is based on clustering the activated hidden states of RNNs and their variant architectures using k-means clustering and uses the algorithm developed by [17] for RNN, LSTM, MGU and GRU. The proposed approach is divided into the following phases:

- **Intuitive approach** Intuitively, the RNN model ( $R$ ) is trained on training and validation data. Employing early stopping enhances generalization by preventing overfitting. The test dataset validates model performance. Each hidden state is regarded as a vector or a point. Thus, many hidden states accumulate when several sequences are input to RNNs. In a study [17], authors observed that activated hidden states in neural networks, including gated RNNs, tend to cluster. It is assumed that the different clusters represent the different states of the finite state automata. State transitions occur when input sequence symbols are read. Thus, it can be inferred that the network behaves like a state automaton machine. Additionally, it is assumed that the states obtained are finite, allowing the learning of deterministic finite state automata from the RNNs and their variant gated architectures, including LSTM, MGU, and GRU.

- **Hidden states clustering** It is contemplated to employ a k-means clustering algorithm to cluster the activated hidden states ( $hi$ ) generated by the RNN when processing symbols from the input sequence during network training.
- **Extraction and learning of FSA** FSA  $M$  is a 5-tuple  $M = \langle \Sigma, Q, S, F, \delta \rangle$  where  $\Sigma$  is the alphabet, meaning the set of the elements appearing in the input sequences,  $Q$  is a set of states,  $S \in Q$  is the start state,  $F \subseteq Q$  is a set of accepting states and  $\delta : Q \times \Sigma \rightarrow Q$  defines state transitions in  $M$  [17]. Two matrices are generated to learn the transitions between the states: the neighbouring matrix (Ns) and the transition matrix(Ts). The neighbouring matrix provides a way to visualize the relationships between different states of FSA. Each state is represented by a row and column in the matrix, and the elements in the matrix indicate the strength of the connection between different pairs of states. By analyzing the neighbouring matrix, one can identify patterns and clusters of related states. Matrix  $Ns$  captures state transitions' frequency ( $i$  to  $k$ ) with the symbol  $s$ . The transition matrix  $T$  is formed by identifying the most frequent transitions ( $i$  to  $k$ ) for each row  $i$ , given input symbol  $j$ . In matrix  $T$ , each row  $i$  corresponds to the state  $k$  with the highest  $Ns(i, k)$  value, representing the most frequent transition from state  $i$  for input symbol  $j$ .

### 3.2. Results and Contribution till date

Our research is in its early stages, focusing on comprehensively reviewing key research papers to validate our objectives. Most progress has been made in establishing a conceptual foundation and creating initial prototype feeder experiments. These experiments aim to learn deterministic finite state automata from RNN and LSTM, aligning with our stated objectives. Understanding state transitions within finite state automata, formed by clustering-activated hidden states of RNN and LSTM during sequence symbol reading, is crucial for concept clarity. A prototype experiment was created for multi-class classification, predicting the next sequence of finite lengths from multiple classes. The details are discussed as follows. A synthetic temporal dataset of 100 sine waves of 100 points each was chosen to perform this experiment. his dataset was chosen due to its shared properties with time series data, such as temporal nature, sequential dependence, cyclic patterns, and potential noise. The process was initiated with exploratory data analysis and preprocessing. Following that, the numerical signals underwent symbolic abstraction for state machine learning, utilizing Symbolic Aggregate Approximation (SAX).<sup>1</sup> to discretize the data. Employing this approach, time series data is transformed into sTheancing interpr exploredtability. Through the SAX algorithm, various strategies and bin sizes were explored for symbol transformation. Subsequently, the obtained symbols were encoded as part of the process. Various encoding methods were investigated while experimenting, and it was observed that integer encoding was suitable for our proposed approach. Further, the dataset was split into training, testing, and validation datasets. Our task is to predict the next in the sequence from multiple classes of symbols as generated above. The number of classes depends on the bin size while applying the SAX algorithm. The experiments were conducted for bin sizes 3, 4, and 5 during the prediction of the next item in the sequence using multi-class classification.

---

<sup>1</sup><https://www.cs.ucr.edu/~eamonn/SAX.html>

For example, if bin size=3, an input sequence would look like *aabbbca*, formed from 3 symbols (a, b, c). Our task for this experiment is to predict the next symbol in the sequence *aabbbca* from a class set of a, b, c. Subsequently, RNN and LSTM models were employed, implementing an early stopping technique and optimizing hyper parameters. Furthermore, clustering was performed on the activated hidden states of the models using the k-means algorithm. One can observe which states tend to activate together and which transitions are more frequent by clustering the hidden states. This analysis provides insights into how the LSTM processes information over time. For instance, if specific clusters frequently activate together and certain transitions dominate, it suggests the LSTM's detection of distinctive data patterns. These clusters represent the finite states of the FSA. From these clusters, the Transition matrix **Ts** and the neighbouring matrix **Ns** were constructed. These matrices define the frequency count of transitions between the different states for each symbol. The deterministic finite state automata obtained can be visualized using the open-source graph visualization software Graphviz. In summary, clustering hidden states simplifies transition matrix creation and enhances interpretability. By grouping similar feature representations, clustering aids in understanding. Once states are clustered and assigned to FSA, the neighbouring matrix **Ns** showcases relationships and transitions. Various methods for optimal k-value in k-means clustering for activated RNN and LSTM hidden states were explored while experimenting Neighboring (**Ns**) and transition (**Ts**) matrices are generated and used to visualize deterministic finite state automata. Visual representations can elucidate the intricate relationships between different components of the model, aiding in the identification of influential factors that drive the predictions. Graphical explanations empower experts and users, fostering transparency and trust in decision-making process.

#### **4. Expected next steps and final contribution to knowledge**

Our upcoming phase involves assessing cluster quality, evaluating FSA states using machine learning-based criteria, and testing RNN and LSTM-extracted FSA accuracy. In parallel, this approach is also being applied to real-world datasets, including an electroencephalogram dataset from the AI and Cognitive Load Research Lab (Technological University Dublin). Our future focus is understanding how well finite state automata capture RNN inner workings for varied sequence lengths, including infinite ones for time series data. Different clustering algorithms and additional parameters to enhance cluster quality evaluation will also be explored. This study will also address the lack of evaluation parameters for assessing DFA quality from RNNs, exploring FSA limitations for post hoc explanations and user trust. The primary contribution will be understanding RNN behaviour on time series data, enhancing their interpretability, explainability, and trustworthiness in various domains.

#### **Acknowledgments**

This publication has emanated from research supported in part by a grant from Science Foundation Ireland under Grant number 18/CRT/6183.

## References

- [1] M. Nagahisarchoghaei, N. Nur, L. Cummins, N. Nur, M. M. Karimi, S. Nandanwar, S. Bhattacharyya, S. Rahimi, An empirical survey on explainable ai technologies: Recent trends, use-cases, and categories from technical and application perspectives, *Electronics* 12 (2023) 1092.
- [2] Q. Wang, K. Zhang, A. G. Ororbia II, X. Xing, X. Liu, C. L. Giles, An empirical evaluation of rule extraction from recurrent neural networks, *Neural computation* 30 (2018) 2568–2591.
- [3] R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, P. De Hert, Bridging the gap between ai and explainability in the gdpr: towards trustworthiness-by-design in automated decision-making, *IEEE Computational Intelligence Magazine* 17 (2022) 72–85.
- [4] R. N. Zaeem, K. S. Barber, The effect of the gdpr on privacy policies: Recent progress and future promise, *ACM Transactions on Management Information Systems (TMIS)* 12 (2020) 1–20.
- [5] 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.
- [6] L. Rizzo, L. Longo, A qualitative investigation of the explainability of defeasible argumentation and non-monotonic fuzzy reasoning, in: *Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science Trinity College Dublin, Dublin, Ireland, December 6-7th, 2018.*, 2018, pp. 138–149.
- [7] L. Rizzo, L. Longo, An empirical evaluation of the inferential capacity of defeasible argumentation, non-monotonic fuzzy reasoning and expert systems, *Expert Systems with Applications* (2020) 113–220. URL: <http://www.sciencedirect.com/science/article/pii/S0957417420300464>. doi:<https://doi.org/10.1016/j.eswa.2020.113220>.
- [8] A. Ghods, Creating interpretable data-driven approaches for remote health monitoring, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2021, pp. 15712–15713.
- [9] 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).
- [10] G. Pellegrino, C. Hammerschmidt, Q. Lin, S. Verwer, Learning deterministic finite automata from infinite alphabets, in: *International Conference on Grammatical Inference*, PMLR, 2017, pp. 120–131.
- [11] C. A. Hammerschmidt, S. Verwer, Q. Lin, R. State, Interpreting finite automata for sequential data, *arXiv preprint arXiv:1611.07100* (2016).
- [12] A. L. Cechin, D. Regina, P. Simon, K. Stertz, State automata extraction from recurrent neural nets using k-means and fuzzy clustering, in: *23rd International Conference of the Chilean Computer Science Society, 2003. SCCC 2003. Proceedings.*, IEEE, 2003, pp. 73–78.
- [13] P. Tiño, B. G. Horne, C. L. Giles, P. C. Collingwood, Finite state machines and recurrent neural networks—automata and dynamical systems approaches, in: *Neural networks and pattern recognition*, Elsevier, 1998, pp. 171–219.
- [14] J. L. Elman, Finding structure in time, *Cognitive science* 14 (1990) 179–211.
- [15] H. Jacobsson, Rule extraction from recurrent neural networks: Ataxonomy and review,

- Neural Computation 17 (2005) 1223–1263.
- [16] C. L. Giles, C. B. Miller, D. Chen, G.-Z. Sun, H.-H. Chen, Y.-C. Lee, Extracting and learning an unknown grammar with recurrent neural networks, *Advances in neural information processing systems* 4 (1991).
  - [17] 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.
  - [18] P. Dondio, L. Longo, Trust-based techniques for collective intelligence in social search systems, in: *Next generation data technologies for collective computational intelligence*, Springer, 2011, pp. 113–135.
  - [19] P. Manolios, R. Fanelli, First-order recurrent neural networks and deterministic finite state automata, *Neural Computation* 6 (1994) 1155–1173.
  - [20] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D. A. Keim, Towards a rigorous evaluation of xai methods on time series, in: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, IEEE, 2019, pp. 4197–4201.
  - [21] A. Das, P. Rad, Opportunities and challenges in explainable artificial intelligence (xai): A survey, *arXiv preprint arXiv:2006.11371* (2020).
  - [22] G. Vilone, L. Longo, A novel human-centred evaluation approach and an argument-based method for explainable artificial intelligence, in: I. Maglogiannis, L. Iliadis, J. Macintyre, P. Cortez (Eds.), *Artificial Intelligence Applications and Innovations - 18th IFIP WG 12.5 International Conference, AIAI 2022, Hersonissos, Crete, Greece, June 17-20, 2022, Proceedings, Part I*, volume 646 of *IFIP Advances in Information and Communication Technology*, Springer, 2022, pp. 447–460. URL: [https://doi.org/10.1007/978-3-031-08333-4\\_36](https://doi.org/10.1007/978-3-031-08333-4_36). doi:10.1007/978-3-031-08333-4\_36.
  - [23] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, *Data Mining and Knowledge Discovery* (2023) 1–59.
  - [24] W. Saeed, C. Omlin, Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities, *Knowledge-Based Systems* 263 (2023) 110273.
  - [25] W. Freeborough, T. van Zyl, Investigating explainability methods in recurrent neural network architectures for financial time series data, *Applied Sciences* 12 (2022) 1427.
  - [26] F. Mujkanovic, V. Doskoč, M. Schirneck, P. Schäfer, T. Friedrich, timexplain—a framework for explaining the predictions of time series classifiers, *arXiv preprint arXiv:2007.07606* (2020).