Use of Explainable AI to Refine Artificial Immune System Algorithms*

Rachana R. Patel[1][0000−0003−2147−4080]

Heriot-Watt University, Edinburgh, EH14 4AS, UK

Abstract. Artificial Immune System (AIS) algorithms offer distributed Artificial Intelligence (AI) that can be effectively used to solve real-world data segregation problems. Current unmet needs include (i) refining the AIS algorithm using representative patterns based on deterministic features of the data and (ii) interpretability of the AIS based classifiers. Model agnostic eXplainable AI (XAI) approaches may cater to both of these unmet needs. Here, we propose building an AIS based classifier pipeline using XAI algorithms to identify deterministic feature patterns from data that can be used to improve the performance and interpretability of AIS classifiers.

Keywords: AIS· NSA· CSPRA· DCA· XAI· LIME· SHAP.

1 Introduction

Artificial Immune System (AIS) is a branch of biologically-inspired computation for developing computational models and methodologies based on the principles of the biological immune system (BIS). Self discrimination by distinguishing own body cells (self) from foreign or transformed entities (non-self) is one of the primary abilities of BIS [1]. AIS is relatively a young field of research where each molecular player of BIS can potentially inspire at least one AIS model. The molecular function of BIS, including self-discrimination, feature extraction, pattern recognition, learning, memory, and distributive yet cooperative nature, have been adapted in AIS to develop immune-inspired models. The AIS models have a wide range of applications such as anomaly detection, pattern recognition, clustering bases, data analyses, function optimisation, and computer security [1]. Basic AIS algorithms such as the Negative Selection Algorithm (NSA) that differentiates self (normal) from non-self (abnormal) are ideal for real-world problems where normal data is well defined, and any alterations can be deemed as anomalies. For example, cancer detection or labelling device performance as good or bad.

Although useful, the simple NSA algorithm suffers from a high false-positive rate [2]. The proposed refinements include extending NSA to include an additional check for pattern recognition. Briefly, in addition to the non-self detectors, NSA can be refined as Conserved Self Pattern Recognition Algorithm (CSPRA)

* Supported by Heriot-Watt University and the Data Lab

to include self pattern detectors that confirm the non-self classification. Such an approach has previously been shown to decrease the false positive rate and improve accuracy [2]. Similarly, the overall performance of AIS classifiers can be improved by using both self and non-self patterns during training. Dendritic Cell Algorithm (DCA) is one such algorithm that improves performance by including both self and non-self data during classifier training [3]. AI algorithms, including AIS algorithms, suffer from a lack of transparency. Hence, the eXplainable Artificial Intelligence (XAI) approaches are essential to improve the understanding and interpretability of AIS classifiers. Among the XAI approaches available, the model agnostic approaches are easier to incorporate into AI/ML pipeline. XAI algorithms such as Local Interpretable Model-agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) can be used for the identification of deterministic features [4] [5] [6]. Thus, LIME and SHAP based parallel approaches can be used to understand how AIS classifiers segregate data as self and non-self.

2 Proposed Idea

The current unmet need for AIS classifiers is two-fold- (i) refinement of algorithms to include appropriate self and non-self patterns and (ii) apply XAI to improve AIS interpretability. Based on the current limitation of AIS, we propose the use of XAI to identify self and non-self patterns to refine AIS algorithms. Briefly, the idea includes creating a generalisable classification pipeline for datasets with clearly anticipated (normal or self) data. The pipeline will include NSA, CSPRA and DCA as the three AIS algorithms. Each AIS based classifier is used to overcome the limitations of another classifier. For example, CSPRA will be used to reduce the anticipated false positive rate of NSA. Similarly, DCA will be used to overcome the limitation of using only one class to train NSA and CSPRA. Both CSPRA and DCA use data patterns as the deterministic factor for classification. We propose the use of LIME and SHAP to identify deterministic features governing the self and non-self classification. Once the deterministic features are identified, well-informed self and non-self patterns can be generated for CSPRA and DCA. Thus, the XAI directed self and non-self patterns will be used to refine AIS algorithms to improve the classifier performance and interpretability.

The two significant anticipated limitations for the proposed idea include (i) generalisability and (ii) coverage of the self and non-self patterns. Since LIME and SHAP uses only a subset of samples, it is possible that the pattern identified from the limited number of samples may not be representative of the entire dataset and therefore not generalisable. Similarly, the number of samples to be analysed by LIME and SHAP may not offer sufficient coverage for the entire dataset to identify appropriate self and non-self patterns. Thus, the number of samples analysed using LIME and SHAP with NSA to identify self and non-self patterns will be an anticipated challenge of the proposed approach, which is shown in Figure 1.

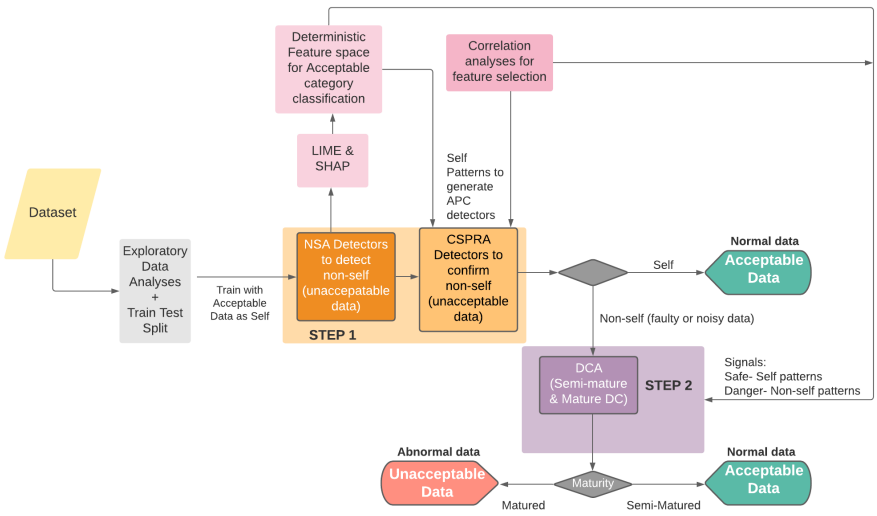

Fig. 1. Two step classification pipeline where LIME and SHAP algorithms are used to refine the AIS algorithms.

3 Conclusions

Overall, we propose the combination of XAI methods to refine AIS algorithm based classifiers as a new way to improve data segregation to tackle real-world problems. Further work on the proposed idea will enable us to identify best approach to select deterministic features for AIS refinements.

References

1. D. Dasgupta and N. L. Fernando, Immunological computation; theory and applications, First. CRC Press Taylor Fransis Group, 2009.
2. S. Yu and D. Dasgupta, "Conserved Self Pattern Recognition Algorithm with Novel Detection Strategy Applied to Breast Cancer Diagnosis," J. Artif. Evol. Appl., vol. 2009, pp. 1–12, 2009, doi: 10.1155/2009/130498.
3. J. Greensmith, U. Aickelin, and S. Cayzer, "Introducing Dendritic Cells as a Novel Immune-Inspired Algorithm for Anomaly Detection.," in Jacob C., Pilat M.L., Bentley P.J., Timmis J.I. (eds) Artificial Immune Systems. ICARIS 2005. Lecture Notes in Computer Science, C. Jacob, M. L. Pilat, P. J. Bentley, and J. I. Timmis, Eds. Berlin, Heidelberg: Springer., 2005.
4. A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.
5. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should i trust you?' Explaining the predictions of any classifier," Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., vol. 13-17-Augu, pp. 1135–1144, 2016, doi: 10.1145/2939672.2939778.

6. P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," Entropy, vol. 23, no. 1, pp. 1–45, 2021, doi: 10.3390/e23010018.