CAT in the Box: A CausalAI – Tsetlin Machine Duo Enabling explainable Stroke Diagnosis and Prevention

Jalpa Soni, Emelian Gurei, Jaime Lopez Sahuquillo, Sergio García Gomez, Victor M. Saenger
AI Innovation Lab, Capitole Consulting,
Balmes, 89, 08008 - Barcelona, Spain
email: jalpabensoni@capitole-consulting.com

Manuel Rodriguez Yañez, Francisco Campos Perez
Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS),
Clinical Hospital, D Building, 1st Floor,
Travesía da Choupana S/N, 15706 Santiago de Compostela, Spain

***Abstract*— In this paper, we propose an explainable framework to assess biomarker significance in brain stroke data by combining Causal Artificial Intelligence (AI), which models cause–effect relationships beyond simple correlations, with a Tsetlin Machine, a symbolic rule-based learning algorithm that generates human-readable logic clauses. In a first step, Causal AI is used to uncover complex interdependencies among biomarkers and to identify the most impactful ones, while the interpretable clauses of the Tsetlin Machine enhance understanding and support improved diagnosis, prognosis, and prevention in stroke patients. This methodological strategy sets a novel foundation for better understanding of complex brain diseases.**

Keywords - Brain stroke; Causal AI; Explainability; Interpretability; Tsetlin Machine.

I. INTRODUCTION

Stroke, caused by an alteration of the blood supply to the brain, is a medical emergency that requires immediate attention in urgent care departments and specialized stroke units. It is a leading cause of long-term disability and the second leading cause of death globally. In Spain, about 1 in 5 stroke patients are readmitted with a recurrent stroke [1][2]. These statistics highlight the importance of early and accurate diagnosis, as timely intervention can significantly reduce mortality and long-term disability. Despite notable advances in medical imaging and diagnostics, deciphering the intricate relationships among stroke-related biomarkers remains a significant challenge.

In recent years, Machine Learning (ML) has shown promise for detecting subtle patterns in biomedical data [3]. However, many ML models lack transparency, offering limited insight into how predictions are made. This opacity poses a major barrier to their adoption in clinical settings, where trust, accountability, and explainability are essential for informed decision-making.

In this paper, we propose a novel approach that integrates Causal AI [4] to model cause-effect relationships rather than simple correlations among stroke-related biomarkers with Tsetlin Machines [5][6][8][9], a symbolic, rule-based learning model that can uncover and help interpret how specific biomarkers influence stroke outcomes. Causal AI refers to machine learning methods that model cause–effect relationships, beyond mere correlations, whereas Tsetlin Machines are interpretable, rule-based learning models that construct human-readable logic clauses for classification tasks [6]. For example, a Tsetlin Machine might generate a rule such as: *"If LDL cholesterol is high **and** age is above 65, **and** prior use of antiplatelet drugs is absent, then the patient is more likely to suffer an ischemic stroke."* Such clauses are easily understandable by clinicians and can be directly compared with established medical knowledge. Together, these not only enhance predictive accuracy, but also provide a transparent, interpretable insight essential for clinical decision-making.

The rest of the paper is organized as follows. In **Section II**, we describe the methodology, including an overview of the dataset, pre-processing steps, the application of Causal AI, and the use of Tsetlin Machines for interpretable classification. In **Section III**, we present and discuss the results obtained from both the causal inference analysis and the Tsetlin Machine model, highlighting their clinical relevance. In **Section IV**, we conclude the paper by summarizing the key findings and outlining directions for future research and model improvements.

II. METHODOLOGY

In this section, we describe the methodology, with subsections on an overview of the dataset, pre-processing steps, Causal AI, and the Tsetlin Machines.

A. Overview

As mentioned in the introduction, we employ a hybrid methodology that combines Causal AI, a set of techniques designed to model cause–effect relationships rather than mere correlations, with Tsetlin Machines, symbolic rule-based learning algorithms capable of generating human-readable logic clauses. This integrated approach allows us to both identify the underlying causal relationships among

biomarkers that drive clinical outcomes in stroke diagnosis and prognosis, and to extract interpretable rules that clarify how specific biomarker patterns contribute to different stroke subtypes. By linking causal discovery with transparent classification, our method not only improves predictive power but also enhances clinical trust and explainability. The study has received the ethical approval of the Santiago/Lugo clinical ethical committee (code: 2025/221).

B. Dataset and pre-processing

The dataset consists of about 4000 data points with 62 features, containing relevant clinical, demographic and biochemical biomarkers. Standard pre-processing steps were applied, as listed below:

- Removal of non-relevant features using domain knowledge (e.g., multiple stroke determination tests at various times would dominate causal relations, suppressing the weight of other biomarkers).
- Missing value imputation using binary and iterative imputers, which estimate missing values by iteratively predicting them based on other available features. This is particularly useful in this data set as the relationships between medical features can provide valuable information for filling in missing data. This is done for binary and non-binary features respectively.

C. Causal AI

To identify potential causal relationships among biomarkers, we applied the PC algorithm (after its authors, Peter and Clark), a constraint-based causal discovery method, to the pre-processed dataset [7]. At this stage, the dataset contains approximately 50 features including the target (type of stroke – ischemic or haemorrhagic).

Since our objective is to isolate the most influential biomarkers, we employed two graph-theoretic measures to rank nodes (features) within the causal graph:

- *Degree Centrality:* Measures the number of direct connections for a node. High degree centrality suggests that a feature has broad influence.
- *Betweenness Centrality:* Quantifies how often a node appears on the shortest paths between other nodes. High betweenness centrality implies that a feature is a critical intermediary or bridge in the causal network.

To minimize selection bias to ensure that both direct and indirect influences are taken into account, we first created two separate ranked lists of features: one based on degree centrality and the other based on betweenness centrality. From each ranking, we extracted the top 25 features, representing those with the strongest influence according to the respective measures. Next, we introduced a composite centrality score, which assigns weights to features depending on their positions in the two rankings, thereby balancing the contribution of both centrality measures. Finally, by comparing the two lists and focusing on the features with the highest combined scores, we identified the 10 most influential biomarkers that consistently appeared as important across both centrality perspectives.

D. Tsetlin Machines

Following the identification of the top 10 biomarkers through causal inference, we applied a rule-based convergence Tsetlin Machine (TM) [8][9][10] to model their relationship with stroke subtypes. This model is a logic-based learning algorithm that constructs human-interpretable propositional logic clauses to perform classification. It operates by learning patterns expressed as conjunctive logical clauses, where each clause is essentially a combination of conditions that must be satisfied for a prediction to be made (for example, *if biomarker A is present and biomarker B is absent, then the case belongs to class X*). Rather than relying on a single clause, the Tsetlin Machine generates a large set of such clauses, each of which casts a "vote" for a particular class. These votes are then aggregated, and the overall prediction is determined by the balance of evidence provided by all the clauses together. This ensemble-like mechanism allows the model to capture subtle, complex patterns while still maintaining a form that remains human-interpretable.

We used the MultiClassTsetlinMachine from pyTsetlinMachine Python module and utilised the in-built *bit-per-feature binarization* to binarize the data [11]. This method discretizes continuous variables into a fixed number of bins, encoding each bin as a separate binary feature. This transformation ensures compatibility with TM's binary input format. The original bin values are stored separately to correctly identify the real values of the features corresponding to the clauses.

After binarization, an 80-20 train-test split was applied and the model was trained with appropriate hyper-parameters (i.e., the number of clauses, threshold, and specificity).

Our target variable represents stroke subtypes (a binary classification task) and the TM generated 50 clauses for each class. To identify the most influential clauses per class, we analysed their voting weights, which reflect how frequently a clause contributes to a particular class prediction. We selected the top clauses based on these weights to further enhance interpretability and explainability and to reduce redundancy, with two filters:

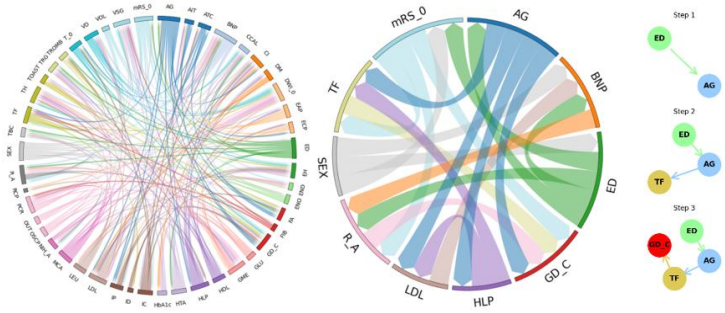

Figure 1. Causal graphs: (a) with all the features, (b) top 10 features using composite score of degree and betweenness centralities and (c) deconstructed specfic causal path

- *Bias Check*: We excluded clauses that were overwhelmingly positive or negative for a single class to avoid skewed interpretations.
- *Redundancy Check:* Clauses that appeared identically in both classes of the outputs were removed, as they introduce ambiguity in the interpretation of feature impact.

After filtering, we retained the distinct and unbiased clauses for each class with the highest voting weights. These clauses form the basis for interpreting how specific combinations of biomarker presence or absence influence the classification of stroke subtypes.

III. RESULTS AND DISCUSSION

Based on the process explained in the methodology section, our final goal is to obtain the top clauses for each of the output classes. To simplify further, we retrieve the most important features for each class as well as the information whether their absence or presence is important for either class.

In this section, we discuss the results of both the causal AI and the Tsetlin Machine.

A. Results of Causal inference

We extract the list of top nodes/features using the *composite centrality*, as defined in the methodology section. The causal Directed Acyclic Graph (DAG) connections comparing original features and the extracted top 10 features using causal inferences are shown in Figure 1.

The first graph (Figure 1a) presents the complete set of features and biomarkers included in the dataset. Because all variables and their interconnections are displayed at once, the result is a complex and visually dense network that makes it difficult to distinguish which biomarkers play the most critical roles. In contrast, the second graph (Figure 1b) focuses only on the top 10 most influential features, as identified through our causal inference procedure using the composite centrality score. This reduced network provides a much clearer picture of the variables that exert the strongest influence on stroke outcomes, allowing clinicians and researchers to focus on the most relevant biomarkers. To further illustrate how causal inference can assign importance to a feature, even when the connection to the target is indirect, the right-hand panel (Figure 1c) zooms in on a specific causal path. In this example, the feature *age (ED) in Figure 1b* does not connect directly to the target variable, GD-C, which represents the type of stroke. Instead, its influence is mediated through an intermediate biomarker, *AG* (prior use of antiplatelet drugs), which then affects *TF* (treatment to dissolve blood clots), and only at that point does the causal chain reach GD-C. This breakdown demonstrates how a variable can still be considered highly important when it contributes to the target outcome through a series of intermediate links, rather than through a direct relationship as well as to trace and understand how each node in the causal graph contributes to the target outcome, whether through direct or indirect pathways.

The top features/biomarkers identified by the causal model and their significance in the context of stroke related literature is summarized in Table 1 below.

TABLE I. MOST IMPORTANT BIOMARKERS AS PER CAUSAL MODEL

Feature	Description	Significance
BNP	Blood test to help diagnose heart failure	A strong indicator for cardiac stress, important for stroke diagnosis/prognosis
AG	Prior use of antiplatelet drugs	Aligns with existing clinical evidence that such medications reduce the risk of recurrent stroke
ED	Age of the patient	A critical determinant of stroke severity and recovery potential
HLP	Abnormally high levels of lipids (fats)	Associated with increased stroke risk; important for stroke prevention strategies
LDL	Bad cholesterol	Linked to atherosclerosis and subsequent cerebrovascular events; a key modifiable risk factor
R_A	Degree of disability after a stroke at discharge	Reflects the immediate functional outcome post-stroke; serves as a proxy for the effectiveness of acute care

mRS_0	Baseline disability in daily activities	Predictive of post-stroke recovery trajectories
SEX	Gender of the patient	Reflects gender effect in stroke prognosis and prevention
TF	Treatment to dissolve blood clots	Highlights a critical role of emergency treatments in improving stroke outcomes
GD_C	Category of the stroke type (target)	Classification of stroke types; target of this study

As can be seen from the *significance* column in Table 1, the causal model validates known clinical associations. Additionally, it also captures nuanced interdependencies among biomarkers by providing the strength of connections between them (i.e., node connection strengths calculated using *composite score* as described in the methodology section).

The model's ability to prioritize features with both statistical and clinical relevance strongly supports its potential application in decision support systems for stroke management.

B. Results of Tsetlin Machine

As previously mentioned, a TM produces human-readable clauses (e.g., *if A and not B, then class X*). After applying the model to the top features identified through causal inference, we derive such clauses for our target variable, the type of stroke.

Figure 2 provides a visual depiction of the clauses. In this illustration, pink cells indicate the absence of a feature for the corresponding class shown at the bottom, while light green cells represent its presence. Each feature's value range is displayed within its respective cell. The feature SEX is binarized, with 0 → female and 1 → male.

The clause for Ischemic stroke would then be:

> *If the modified Rankin Scale (mRS_0) score is greater than 2.67,* ***and*** *the LDL level is between 71 and 117 mg/dL,* ***and*** *the patient's age is not greater than 56 years,* ***and*** *the BNP level is not between 550 and 1123 pg/mL,* ***then*** *the predicted outcome is Ischemic stroke.*

Which in logic notation is:

$$IF\ (mRS_0 > 2.67) \wedge (71 < LDL > 117) \wedge (Age < 56) \wedge \neg(550 < BNP < 1123) \rightarrow ISCHEMIC$$

Such human-readable clauses, with well-defined value ranges for each feature or biomarker influencing the output classes, could become particularly valuable in clinical settings.

In terms of clinical research, they enhance model transparency, enabling researchers to validate findings against existing biomedical knowledge and uncover novel associations. This interpretability can help bridge the gap between data-driven models and domain expertise. Furthermore, such clauses can inform the design of prospective studies and contribute to the development of explainable clinical decision support tools.

Finally, having transparency in clinical decision-making would benefit effective patient communication, helping individuals understand prevention strategies and treatment options.

IV. Conclusion and Future Work

The findings presented here are preliminary and require further refinement. A key priority is to acquire additional

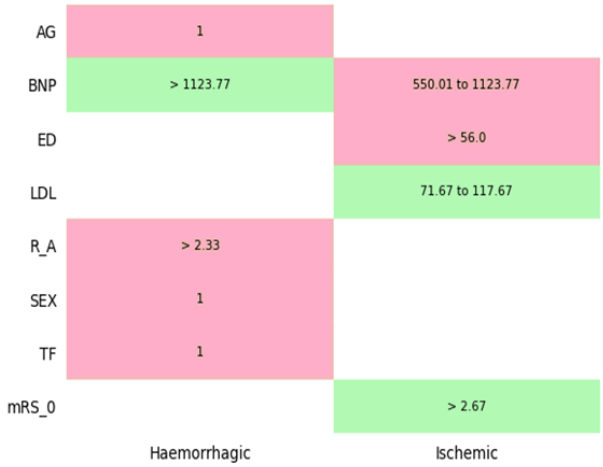

Figure 2. Visual representation of Tsetlin Machine clauses identified for the target with most important biomarkers.

data and repeat the analysis to ensure the robustness of the results. We are in the process of obtaining a more comprehensive dataset, which will include recent records of stroke patients.

To further strengthen the robustness of the results, the next steps are broadly categorized into two areas: one focusing on Causal AI and the other on rule extraction using the Tsetlin Machine.

A. Causal AI

To ensure the accuracy of the causal graphs, it is essential to correctly capture the directionality of the relationships. Achieving this will require deeper domain expertise and a thorough analysis of how various biomarkers interact.

Additionally, it is vital to conduct *what-if* scenario simulations based on the discovered causal relationships within the feature space. These *in-silico* experiments will

enable us to explore how changes in feature values, whether hypothetical or novel, might influence stroke prognosis, without the need for new empirical data.

B. Tsetlin machine

While our current model achieves an overall accuracy of approximately 80%, a closer examination of its performance metrics reveals a notable imbalance. Specifically, the F1-score for Class 0 (the majority class) reaches 0.88, whereas the F1-score for Class 1 (the minority class) drops sharply to just 0.15. This large disparity highlights that, although the model performs well in predicting the dominant class, it struggles to correctly identify cases that belong to the less frequent class. In practice, this means that the model fails to capture a substantial proportion of minority class instances, which may correspond to clinically critical or rare conditions. The root cause of this problem is the class imbalance present in the dataset, where examples of one stroke subtype greatly outnumber the other. We anticipate that the inclusion of additional patient records in our forthcoming dataset will help mitigate this imbalance by providing a more even distribution of classes.

It is also important to emphasize that a Tsetlin Machine (TM) differs fundamentally from many classical machine learning models. Instead of optimizing a global error function, the TM relies on a frequency-driven clause learning mechanism in which the prevalence of certain patterns directly affects the clauses it learns. While this makes the model efficient and interpretable, it also means that it tends to favor patterns associated with the majority class, often at the expense of learning sufficient rules for the minority class. This characteristic can amplify the effects of class imbalance, as seen in our results.

Nevertheless, in the context of biomedical datasets (where imbalanced class distributions are common) this bias does not necessarily negate the model's clinical utility. Optimizing for the majority class can still yield valuable insights, as the most prevalent stroke subtype remains a major focus of clinical diagnosis and treatment. However, achieving reliable detection of minority cases is equally critical, as these often represent the most challenging and high-risk scenarios. Addressing this imbalance in future work will therefore be essential, ensuring that the TM captures meaningful patterns for both majority and minority classes without sacrificing interpretability.

These facts also do not diminish the importance of accurately identifying minority class instances, which often represent critical or rare conditions. To address this, we are actively exploring various strategies (e.g., resampling, decision threshold tuning, etc.) to improve the model's ability to generalize and perform equitably across both classes. These efforts are guided by domain expertise to ensure that learned patterns are meaningful and to prevent the model from learning artifacts of the data rather than true signals.

Additionally, binarization must be approached with greater care. It is important to ensure that the binning of biomarkers identified as significant by the Tsetlin Machine aligns with domain knowledge and statistical distribution. For example, consider serum Vitamin D levels, which typically range from 0 to 100 ng/mL. Clinical guidelines define severe deficiency as levels below 10 ng/mL, deficiency as below 20 ng/mL, insufficiency between 20–30 ng/mL, and sufficiency as levels above 30 ng/mL. If all values below 30 ng/mL were grouped into a single bin (e.g., bin 0), this would obscure critical clinical distinctions between mild insufficiency and severe deficiency. Such coarse binning could reduce the model's ability to detect meaningful health risks associated with different deficiency levels.

ACKNOWLEDGMENT

The authors would like to thank the Instituto de Salud Carlos III_ICIII RICORS-ICTUS network (grant number RD24/0009/0017) and Xunta de Galicia (grant number: IN607A2022/02) for providing the resources to carry out this work.

REFERENCES

[1] Ministry of Health, "Annual Report of the National Health System 2023," Government of Spain, Aug. 2024. [Online]. Available: https://www.sanidad.gob.es [retrieved: Sep., 2025].

[2] Eurostat, "Causes of death statistics," *Statistics Explained*, European Commission, Mar. 2025. [Online]. Available: https://ec.europa.eu/eurostat/statistics-explained [retrieved: Jun., 2025].

[3] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, "Machine learning and integrative analysis of biomedical big data," *Genes*, vol. 10, no. 2, pp. 87, 2019. [Online]. Available: https://doi.org/10.3390/genes10020087 [retrieved: Aug., 2025].

[4] P. Sanchez, J. P. Voisey, T. Xia, H. I. Watson, A. Q. O'Neil, and S. A. Tsaftaris, "Causal machine learning for healthcare and precision medicine," *Royal Society Open Science*, vol. 9, no. 7, pp. 220638, 2022. [Online]. Available: https://doi.org/10.1098/rsos.220638 [retrieved: Jul., 2025].

[5] O. C. Granmo, "The Tsetlin Machine – A game theoretic bandit driven approach to optimal pattern recognition with propositional logic," *arXiv preprint arXiv:1804.01508*, 2018. [Online]. Available: https://arxiv.org/abs/1804.01508 [retrieved: Sep., 2025].

[6] G. T. Berge, O. C. Granmo, T. O. Tveit, B. E. Munkvold, A. L. Ruthjersen, and J. Sharma, "Machine learning-driven clinical decision support system for

concept-based searching: A field trial in a Norwegian hospital," *BMC Medical Informatics and Decision Making*, vol. 23, no. 5, pp. 1–12, 2023. [Online]. Available: https://doi.org/10.1186/s12911-023-02101-x [retrieved: Jun., 2025].

[7] M. Kalisch and P. Bühlmann, "Estimating high-dimensional directed acyclic graphs with the PC-algorithm," *Journal of Machine Learning Research*, no. 8, pp. 613–636, 2007. [Online]. Available: https://jmlr.csail.mit.edu/papers/v8/kalisch07a.html [retrieved: Aug., 2025].

[8] A. Wheeldon, A. Yakovlev, and R. Shafik, "Self-timed reinforcement learning using Tsetlin Machine," in *Proceedings of the 27th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC 2021)*, IEEE, 2021. [Online]. Available: https://arxiv.org/abs/2109.00846 [retrieved: Jun., 2025].

[9] S. Glimsdal and O.-C. Granmo, "Coalesced multi-output Tsetlin Machines with clause sharing," *arXiv preprint arXiv:2108.07594*, 2021. [Online]. Available: https://arxiv.org/abs/2108.07594 [retrieved: Sep., 2025].

[10] K. D. Abeyrathna, O.-C. Granmo, and M. Goodwin, "Adaptive sparse representation of continuous input for Tsetlin Machines based on stochastic searching on the line," *Electronics*, vol. 10, no. 17, pp. 2107, Aug. 2021. [Online]. Available: http://dx.doi.org/10.3390/electronics10172107 [retrieved: Jul., 2025].

[11] O. C. Granmo, et al., "pyTsetlinMachine [Computer software]," GitHub, n.d. [Online]. Available: https://github.com/cair/pyTsetlinMachine [retrieved: Jun., 2025].