# A Novel Tropical Geometry-Based Interpretable Machine Learning Method: Pilot Application to Delivery of Advanced Heart Failure Therapies

Heming Yao, Harm Derksen, Jessica R. Golbus, Justin Zhang, Keith D. Aaronson, Jonathan Gryak, and Kayvan Najarian

***Abstract*—A model's interpretability is essential to many practical applications such as clinical decision support systems. In this article, a novel interpretable machine learning method is presented, which can model the relationship between input variables and responses in humanly understandable rules. The method is built by applying tropical geometry to fuzzy inference systems, wherein variable encoding functions and salient rules can be discovered by supervised learning. Experiments using synthetic datasets were conducted to demonstrate the performance and capacity of the proposed algorithm in classification and rule discovery. Furthermore, we present a pilot application in identifying heart failure patients that are eligible for advanced therapies as proof of principle. From our results on this particular application, the proposed network achieves the highest F1 score. The network is capable of learning rules that can be interpreted and used by clinical providers. In addition, existing fuzzy domain knowledge can be easily transferred into the network and facilitate model training. In our application, with the existing knowledge, the F1 score was improved by over 5%. The characteristics of the proposed network make it promising in applications requiring model reliability and justification.**

***Index Terms*—Artificial intelligence, explainable machine learning, interpretable machine learning.**

Manuscript received 19 November 2021; revised 3 July 2022 and 2 September 2022; accepted 29 September 2022. Date of publication 4 October 2022; date of current version 5 January 2023. This work was supported by the National Science Foundation under Grant 2014003. *(Corresponding author: Heming Yao.)*

Heming Yao is with the Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48103 USA (e-mail: hemingy@umich.edu).

Harm Derksen is with the Department of Mathematics, Northeastern University, Boston, MA 02115 USA (e-mail: ha.derksen@northeastern.edu).

Jessica R. Golbus and Keith D. Aaronson are with the Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI 48103 USA (e-mail: jgolbus@med.umich.edu; keith@med.umich.edu).

Justin Zhang is with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI 48103 USA (e-mail: zjustin@umich.edu).

Jonathan Gryak is with the Department of Computer Science, Queens College, City University of New York, New York, NY 10017 USA (e-mail: gryakj@med.umich.edu).

Kayvan Najarian is with the Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48103 USA, also with the Department of Electrical and Computer Engineering, University of Michigan, Ann Arbor, MI 48103 USA, also with the Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI 48103 USA, and also with the Max Harry Weil Institute for Critical Care Research and Innovation, University of Michigan, Ann Arbor, MI 48103 USA (e-mail: kayvan@umich.edu).

Digital Object Identifier 10.1109/JBHI.2022.3211765

## I. Introduction

ARTIFICIAL intelligence (AI) and machine learning (ML) have been increasingly applied to healthcare problems [1]. Previous studies investigated AI in disease diagnosis, treatment effectiveness prediction, and outcome prediction [2], [3], [4]. Several studies have shown that AI performs as well as or better than humans [5]. With a lower cost, AI-based decision support systems have the potential to improve patient management.

Despite tremendous progress in the field of AI-based clinical decision support systems, there are significant challenges that prevent the widespread use of these methods in sensitive applications. While traditional models such as linear models provide accessible reasoning, they are less capable of achieving high performance on complicated problems. In contrast, ML models with higher complexity, can yield good metrics on experimental datasets. However, these "black box" models lack transparency and justification of their recommendations, making them much less likely to be trusted in clinical applications. Moreover, many popular ML methods, such as deep learning, utilize a large number of parameters, thus requiring large training datasets to avoid overfitting the data. However, in many clinical applications, collecting large annotated training datasets may be costly or even impossible. As such, there is a clear need for an interpretable ML model that can reliably model data using relatively small training sets. In addition, in healthcare applications, there exist many invaluable heuristics derived from domain knowledge expertise, often in the form of approximate rules that are used by human experts. In the majority of existing AI/ML models, there is no clear mechanism to leverage such approximate knowledge for model formation or training.

The goal of this study is to solve the aforementioned limitations in the field of AI with an application as proof of principle. An interpretable ML algorithm is proposed to produce a transparent classification model and leverage existing domain knowledge to improve model reliability. The proposed network

is built upon tropical geometry and fuzzy inference systems [6], [7], [8], [9]. Tropical geometry is a piecewise-linear version of conventional algebraic geometry. In the proposed network, the encoding functions and the aggregation operators in classical fuzzy inference networks were reformulated by introducing tropical geometry, which enables adaptive fuzzy subspace division and rule discovery. Two synthetic datasets and one practical application in clinical decision support as a pilot evaluation were investigated to demonstrate the capabilities of the proposed model.

The pilot application used in this study is to identify heart failure (HF) patients that are eligible for advanced therapies. HF afflicts 6.5 million Americans 20 and older, with its prevalence projected to increase annually [10], [11]. Treatment of these patients remains limited by medical therapies and, for those with advanced HF, by organ availability. The appropriate delivery of advanced therapies, heart transplantation (HT) or mechanical circulatory support (MCS) implantation, to patients with end-stage HF is highly nuanced and requires expertise from advanced HF cardiologists. Due to the high prevalence of HF, the majority of patients are managed by primary care physicians or cardiologists, who lack training in the management of patients with advanced diseases, such as determining the appropriate time to deliver HF advanced therapies. There are some existing HF risk models by logistic regression but have limited accuracy for individual patients due to the limitation in capturing multidimensional relationships [12]. Thus, there is a need for AI-based tools that can systematically identify patients warranting referrals to an advanced HF cardiologist for consideration of HT or MCS implantation in ambulatory settings. In this application, we built a clinical decision-making model capable of differentiating patients eligible for HF advanced therapies from those too well, too sick, or otherwise ineligible for advanced therapies.

Our contributions in this study can be summarized as follows:

1) A novel interpretable ML algorithm was proposed, whose resulting recommendations are transparent to users such as clinicians and patients. The model can produce humanly understandable rules, enabling new clinical knowledge discovery. The proposed network was validated using synthetic data with ground truth reasoning and a dataset from patients with HF. The experimental results show that the network has the capability to extract hidden rules from datasets and achieved comparable performance with other ML models.
2) With the proposed algorithm, approximate domain knowledge can be directly incorporated into model training to improve the model's performance and reduce the need for a large training set. It makes the proposed algorithm particularly appropriate for clinical applications. From our results, initializing a network with existing approximate knowledge can improve the model's accuracy.
3) The proposed algorithm was successfully used to identify HF patients eligible for advanced therapies, a highly sensitive application in medicine. From our results, the proposed algorithm achieved the highest F1 value. The rules from the trained network were visualized and validated qualitatively by cardiologists. This pilot application is presented as proof of principle to demonstrate the capabilities of the proposed algorithm in solving real-world clinical problems.

## II. Related Work

### A. Interpretable ML Models

In this work, we define "interpretability" as being with the following two properties (A) the ability to explain predictions; and (B) the ability to explain how a model works (i.e., intelligence). The property (A) makes the model capable of providing justification for its decision. The justification is critical for high-stakes decision-making in sensitive applications such as medicine and also is the key to building trust. The property (B) is an addition to (A), which requires the mechanism by which the model works are understandable to humans. Property (B) makes it possible to directly integrate existing human knowledge into the model. It is also critical for trouble-shooting when a model does not work as expected. In addition, if the training data does not represent the distribution of data in the deployment environment, a model with property (B) allows the user's manual intervention [13].

Post-hoc interpretation methods are dedicated to explaining predictions from "black box" ML models (property A). For example, LIME [14] is a popular method that explains the individual predictions of any classifier by learning local surrogate models from the target "black box" model. SHAP [15] is another commonly used method that computes the contribution of each feature to individual predictions for interpretability. However, explanations from post-hoc methods may not be faithful [16] and they have limited capacity in elucidating how to further improve the model.

For property (B), we need to address how a model functions internally by its structure. The simplest examples are linear models, but these may fail whenever the relationships between features and responses are non-linear. Decision trees are another class of transparent models that can capture interactions among different features. However, the structure of the decision tree is highly dependent on feature selection for each split. Generalized additive models are extended linear models that can capture non-linear relationships between the individual or pairwise features and responses [17]. They have been successfully used in practical applications [18] but are less capable of modeling in high-dimensional feature interactions. Another type of transparent model is a fuzzy inference model, which models the relationship between features and responses by constructing compositional rules [6]. In fuzzy inference models, knowledge is represented in the format of fuzziness of antecedents, consequents, and relations. As rules closely approximate human logic in decision-making, and fuzziness often exists in practical applications and especially in healthcare, the proposed network in this study is designed to leverage fuzzy inference systems.

### B. Fuzzy Inference System

Previous studies have shown that fuzzy inference systems can be used for non-linear system approximation and rule identification [8], [9]. A wide spectrum of fuzzy inference systems utilizes the Takagi-Sugeno (TS) inference model [7], whereby a complete rough partition of the input space is generated and

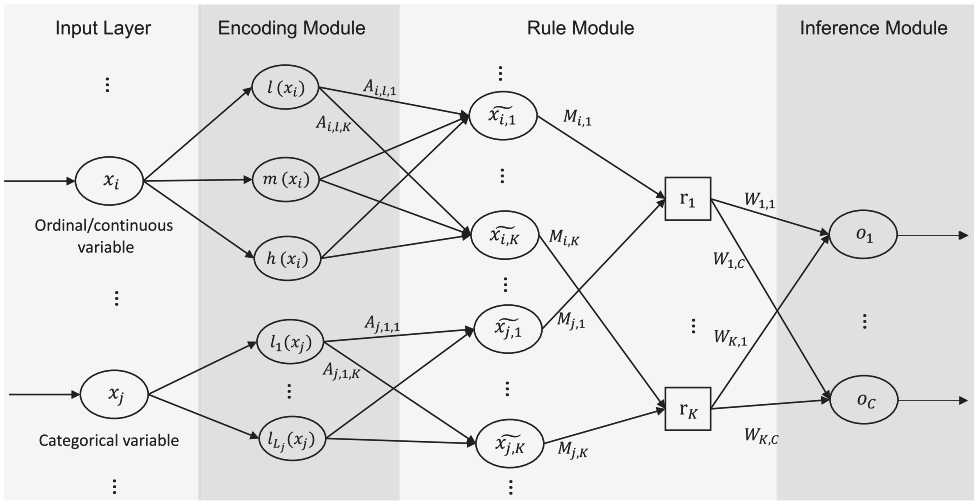


Fig. 1. An overview of the proposed network. The proposed network consists of an input layer, encoding module, rule module, and inference module. The nomenclatures we used in the diagram are described in Section III.

an input-output relation is formed for each subspace. Adaptive Network-based Fuzzy Inference System (ANFIS) [19] is a hybrid of a feed-forward neural network and fuzzy inference system with supervised learning capability that can be used to update the input-output relation in each subspace. ANFIS has been successfully applied in multiple applications [20], [21]. In our previous work [22], an adaptive fuzzy inference network was developed with a genetic algorithm to identify patients eligible for advanced therapies. From our results, the network achieved good classification performance and provided transparent rules.

However, the designs of the TS model and ANFIS pose challenges in practical complex applications where the number of input variables is relatively large as this results in exponential growth in the number of subspaces. To handle this problem, a flexible $k$-d tree [23] and quadtree [24] have been adopted for input space partition but with challenges in assigning understandable terms to membership functions using these methods. Unlike previous methods, we introduce tropical geometry into the fuzzy inference system, which allows the reformulation of the membership functions and aggregation operators. As a result, the shape of the membership functions and aggregation operators do not need to be pre-defined, and they can be optimized during the training process. In addition, instead of using a complete partition of the input space and modeling the relationship between every individual subspace and the output, we proposed a "network" structure. In this "network" structure, a fixed number of subspaces are constructed by combinations of concepts. More importantly, the construction of those subspaces can be updated by optimizing the connection weights. With such a design, an end-to-end network can adaptively and iteratively discover subspaces related to each class using gradient-based back-propagation.

## III. Method

### A. Overview of the Proposed Work

In this study, we designed an end-to-end interpretable classifier shown in Fig. 1. It takes tabular data as input and outputs the predicted class. The proposed network has three major components: an encoding module, a rule module, and an inference module. Firstly, every input variable is encoded into humanly understandable fuzzy concepts in the encoding module. Then a number of fuzzy subspaces are constructed as combinations of fuzzy concepts by attention and connection matrices $\mathbf{A}$ and $\mathbf{M}$. Given a specific data sample, the firing strength of each rule can be calculated. Finally, with the inference matrix $\mathbf{W}$ and the firing strength of each rule, the probabilities of the sample belonging to each class are calculated in the inference module. During the training process, parameters used for input encoding, matrices $\mathbf{A}$, $\mathbf{M}$, and $\mathbf{W}$ are optimized by gradient-based back-propagation. After the network is trained, those parameters can be analyzed to visualize the learned fuzzy subspaces. The fuzzy subspaces mimic human logic and can be presented as "rules". Those rules can be used to justify the model's prediction.

As the proposed network mimics human logic, not only can rules be extracted from the trained model, but also existing knowledge can be integrated/transferred into the model. We performed experiments to investigate whether initializing the network with existing domain knowledge improves the model's performance.

### B. Encoding Module

The input variables can be ordinal, continuous, or categorical. Ordinal and continuous variables are encoded into multiple fuzzy sets. Unlike with crisp sets, for which membership is binary, for fuzzy sets, a membership value in [0,1] will be assigned to a variable's observed value for a given fuzzy set, indicating the confidence of that value belonging to the set. Fuzzy set membership approximates the fuzzy concept used by human experts during decision-making. For example, given the heart rate of a patient, the clinician may describe it as a "low" / "medium" / "high" heart rate. "Low", "medium", and "high" are the fuzzy concepts used in clinical problems. In this study, we encoded ordinal/continuous variables into these three concepts. With an ordinal/continuous variable $x_i$, the membership

functions $l(x_i), m(x_i), h(x_i)$ for "low", "medium", and "high" concepts are defined as

$$f_{\epsilon_1}(x_i) = \epsilon_1 \log(1 + \exp(x_i/\epsilon_1)), \tag{1a}$$

$$l(x_i) = f_{\epsilon_1}\left(\frac{a_{i,2} - x_i}{a_{i,2} - a_{i,1}}\right) - f_{\epsilon_1}\left(\frac{a_{i,1} - x_i}{a_{i,2} - a_{i,1}}\right), \tag{1b}$$

$$m(x_i) = f_{\epsilon_1}\left(\frac{x_i - a_{i,1}}{a_{i,2} - a_{i,1}}\right) - f_{\epsilon_1}\left(\frac{x_i - a_{i,2}}{a_{i,2} - a_{i,1}}\right) - f_{\epsilon_1}\left(\frac{a_{i,3} - x_i}{a_{i,4} - a_{i,3}}\right) + f_{\epsilon_1}\left(\frac{a_{i,4} - x_i}{a_{i,4} - a_{i,3}}\right) - 1, \tag{1c}$$

$$h(x_i) = f_{\epsilon_1}\left(\frac{x_i - a_{i,3}}{a_{i,4} - a_{i,3}}\right) - f_{\epsilon_1}\left(\frac{x_i - a_{i,4}}{a_{i,4} - a_{i,3}}\right), \tag{1d}$$

where $a_{i,1} < a_{i,2} < a_{i,3} < a_{i,4}$ and are trainable. With $0 < \epsilon_1 < 1$, the membership functions are differentiable, with their smoothness modulated by $\epsilon_1$. As $\lim_{\epsilon_1 \to 0} f_{\epsilon_1}(x) = \max(0, x)$, when $\epsilon_1$ approaches 0, the membership functions in (1) are close to trapezoidal membership functions or triangular membership functions (if $a_{i,2}$ is close to $a_{i,3}$).

Using the defined membership functions, $x_i$ will be encoded as membership values in three fuzzy concepts: $l(x_i), m(x_i), h(x_i)$. In this study, we used three concepts - "low", "medium", and "high" - as they are commonly used in healthcare applications. The above formulations can be easily extended to a higher number of concepts.

Categorical variables are represented via a one-hot encoding directly and no fuzzy concepts are used. We denote $L_j$ as the number of levels of a categorical variable $x_j$. In this study, $x_j$ is encoded into $l_1(x_j), l_2(x_j), \ldots, l_{L_j}(x_j)$, where only one of them has a value of 1 and all others are 0.

### C. Rule Module

The rule module consists of two layers in the proposed architecture. In this module, the firing strength of a number of rules (fuzzy subspaces) are calculated for the classification task and denoted as $r_1, \ldots, r_K$ in Fig. 1, where $K$ is the total number of rules.

*1) The First Layer:* The first layer of the rule module selects the most relevant concept from each variable with respect to each rule using an attention matrix $\mathbf{A}$. $\mathbf{A}$ is the partitioned matrix formed by concatenating submatrices $\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_H$, where $\mathbf{A}_h$ is the attention submatrix for the input variable $x_h$ and $H = I + J$ is the total number of input variables, with $I$ and $J$ the total number of ordinal/continuous and categorical variables, respectively. For an ordinal/continuous variable $x_i$, the submatrix $\mathbf{A}_i$ with entries $A_{i,m,n}$ has dimension $3 \times K$, where 3 is the number of fuzzy concepts used in this study and $K$ is the number of rules in the network. For a categorical variable $x_j$, the submatrix $\mathbf{A}_j$ with entries $A_{j,m,n}$ has dimension $L_j \times K$. Thus, the attention matrix $A$ has dimension $(3I + \sum_j L_j) \times K$.

For an ordinal/continuous variable $x_i$, the entry $A_{i,1,k}$ in the attention matrix represents the contribution of $x_i$ being "low" to rule $k$ (and similarly, $A_{i,2,k}$ for $x_i$ being "medium" and $A_{i,3,k}$ for $x_i$ being "high"). Entries in the attention matrix are all trainable and constrained to $[0, 1]$ by the hyperbolic tangent activation function. A higher value in $\mathbf{A}$ indicates a higher contribution. As shown in Fig. 1, for an input variable $x_i$, the corresponding output from the first layer of the rule module is $\widetilde{x}_i$, a vector of length $K$. $\widetilde{x}_{i,k}$, the $k^{th}$ element of $\widetilde{x}_i$, is the firing strength of $x_i$ involved in $k^{th}$ rule.

For an ordinal/continuous variable $x_i$ and categorical variable $x_j$, $\widetilde{x}_{i,k}$, and $\widetilde{x}_{j,k}$ are calculated as:

$$\widetilde{x}_{i,k} = A_{i,1,k} l(x_i) + A_{i,2,k} m(x_i) + A_{i,3,k} h(x_i), \tag{2a}$$

$$\widetilde{x}_{j,k} = \sum_{d=1}^{L_j} A_{j,d,k} l_d(x_j) \tag{2b}$$

respectively.

*2) The Second Layer:* The second layer of the rule module calculates rule firing strength by a connection matrix $\mathbf{M}$ of dimension $H \times K$. The $k^{th}$ rule is constructed as a combination of $\widetilde{x}_{1,k}, \ldots, \widetilde{x}_{H,k}$ from the previous layer. An entry $M_{i,k}$ in the connection matrix $\mathbf{M}$ denotes the contribution of $x_i$ to the $k^{th}$ rule. Entries in the connection matrix are all trainable and constrained to $[0, 1]$ the hyperbolic tangent activation function, and a higher value indicates a higher contribution. In this layer, we define a parametrized T-norm to calculate $r_k$, the firing strength of the $k^{th}$ rule.

With $0 < \epsilon_2 < 1$, let $g_{\epsilon_2} : [0, \infty) \to [0, \infty)$ and its inverse function $g_{\epsilon_2}^{-1}$ be defined as

$$g_{\epsilon_2}(x) = \frac{\epsilon_2}{1 - \epsilon_2}\left(1 - x^{\frac{\epsilon_2 - 1}{\epsilon_2}}\right), \tag{3a}$$

$$g_{\epsilon_2}^{-1}(z) = \left(1 - \frac{1 - \epsilon_2}{\epsilon_2} z\right)^{\frac{\epsilon_2}{\epsilon_2 - 1}}. \tag{3b}$$

The parametrized T-norm on two inputs is defined as

$$\begin{aligned} T_{\epsilon_2}(x, y) &= g_{\epsilon_2}^{-1}(g_{\epsilon_2}(x) + g_{\epsilon_2}(y)) \\ &= \left(x^{\frac{\epsilon_2 - 1}{\epsilon_2}} + y^{\frac{\epsilon_2 - 1}{\epsilon_2}} - 1\right)^{\frac{\epsilon_2}{\epsilon_2 - 1}}, \end{aligned} \tag{4}$$

which has the following asymptotic behavior:

$$\lim_{\epsilon_2 \to 1} T_{\epsilon_2}(x, y) = xy, \tag{5a}$$

$$\lim_{\epsilon_2 \to 0} T_{\epsilon_2}(x, y) = \min(x, y), \tag{5b}$$

which means that the defined T-norm can be modulated between product and min by $\epsilon_2$.

Using this definition of the T-norm, $r_k$ is calculated by applying the T-norm to multiple inputs:

$$\begin{aligned} r_k &= T_{\epsilon_2}\left(\widetilde{x}_{1,k}^{M_{1,k}}, \widetilde{x}_{2,k}^{M_{2,k}}, \ldots, \widetilde{x}_{H,k}^{M_{H,k}}\right) \\ &= g_{\epsilon_2}^{-1}\left(\sum_{i=1}^{H} g_{\epsilon_2}(\widetilde{x}_{i,k}^{M_{i,k}})\right) \\ &= \left(\sum_{i=1}^{H} \widetilde{x}_{i,k}^{M_{i,k} \cdot \frac{\epsilon_2 - 1}{\epsilon_2}} - H + 1\right)^{\frac{\epsilon_2}{\epsilon_2 - 1}}. \end{aligned} \tag{6}$$

In (6), entries in the connection matrix $\mathbf{M}$ are used as exponents. Taking the example of $\widetilde{x}_{1,k}^{M_{1,k}}$, a lower $M_{1,k}$ (closer to 0) means $\widetilde{x}_{1,k}^{M_{1,k}}$ is closer to 1, consequently it contributes less to $r_k$ with the proposed T-norm. Thus, a lower value in $\mathbf{M}$ indicates a lower contribution to the rule firing strength, and vice versa.

With the rule module, the number of rule sub-spaces that can be encoded in the network is roughly $P(F)^N$, where $N$ is the number of variables, $F$ is the number of fuzzy concepts (in this study, $F = 3$), and $P(\cdot)$ denotes the number of permutations. The high complexity of the proposed method makes it capable of modeling complicated classification problems.

### D. Inference Module

Let $C$ denote the number of classes in the classification task. The inference layer has $C$ nodes, one for each class, that are fully connected to the rule layer nodes. The firing strength of each node $o_c$ is calculated using the rule firing strengths with an inference matrix $\mathbf{W}$ of dimension $K \times C$. An entry $W_{j,c}$ denotes the contribution of the $k^{th}$ rule to the $c^{th}$ class. Entries in the inference matrix are all trainable and positive. A higher value indicates a higher contribution. In this layer, we define a parametrized T-conorm to calculate $o_c$.

The parametrized T-conorm on two inputs is written as

$$Q_{\epsilon_3}(x, y) = \left(x^{\frac{1}{\epsilon_3}} + y^{\frac{1}{\epsilon_3}}\right)^{\epsilon_3}, \tag{7}$$

where $0 < \epsilon_3 < 1$. This T-conorm has the following asymptotic behavior:

$$\lim_{\epsilon_3 \to 1} Q_{\epsilon_3}(x, y) = x + y, \tag{8a}$$

$$\lim_{\epsilon_3 \to 0} Q_{\epsilon_3}(x, y) = \max(x, y), \tag{8b}$$

which means that the defined T-conorm can be modulated between addition and max by $\epsilon_3$.

Using this definition of the T-conorm, $o_c$ is calculated by applying the T-conorm to multiple inputs:

$$\begin{aligned} o_c &= Q_{\epsilon_3}\left(W_{1,c}r_1, W_{2,c}r_2, \ldots, W_{K,c}r_K\right) \\ &= \left(\sum_{k=1}^{K} (W_{k,c}r_k)^{\frac{1}{\epsilon_3}}\right)^{\epsilon_3}. \end{aligned} \tag{9}$$

After the calculation of $o_1, o_2, \ldots, o_C$, a softmax activation function is applied to generate probabilities $p_1, p_2, \ldots, p_C$ of being in each class, which are all in $[0, 1]$ with $\sum_{c=1}^{C} p_c = 1$.

As $\sum_{c=1}^{C} p_c = 1$, we can set the number of "valid" nodes in the inference module to $C - 1$ to avoid ambiguity in rule representation. For example, when performing binary classification $W_{:,0}$ can be set to 0 so that the model will only learn subspaces related to the positive class.

### E. Network Interpretation

The proposed network can both extract rules and inject rules in a way that humans can understand. The entries in the attention matrix $\mathbf{A}$ and connection matrix $\mathbf{M}$ represent the contribution of individual concepts and individual variables to each rule. The entries in the inference matrix $\mathbf{W}$ gives the contribution of individual rules to each class.

With $\mathbf{A}$ and $\mathbf{M}$, a contribution matrix $\mathbf{S}$ can be constructed that expresses the contribution of individual concepts to each rule in the model. The matrix $S$ is a partition matrix formed by concatenating submatrices $\mathbf{S}_1, \mathbf{S}_2, \ldots, \mathbf{S}_H$. For an ordinal/continuous variable $x_i$, the corresponding submatrix $\mathbf{S}_i$ has dimension $3 \times K$ and for a categorical variable $x_j$, $\mathbf{S}_j$ has dimension $L_j \times K$. The entries $S_{i,d,k}$ of $\mathbf{S}_i$ and $S_{j,d,k}$ of $\mathbf{S}_j$ are calculated as

$$S_{i,d,k} = A_{i,d,k} \times M_{i,k}, \quad d \in \{1, 2, 3\}, \tag{10a}$$

$$S_{j,d,k} = A_{j,d,k} \times M_{j,k}, \quad d \in \{1, \ldots, L_j\}, \tag{10b}$$

respectively, where $k \in \{1, \ldots, K\}$. The entry $S_{i,d,k}$ is the contribution of the $d^{th}$ concept of $x_i$ to the $k^{th}$ rule. $\mathbf{S}_{:,:,k}$ encodes the construction of the $k^{th}$ rule, while $W_{k,:}$ captures the relationship between classes and the $k^{th}$ rule.

The following is a toy example demonstrating how rules are represented in the network. Given a dataset with four continuous input variable $x_1, x_2, x_3, x_4$ and a binary response (negative/positive), $\mathbf{A}, \mathbf{M}, \mathbf{W}$ are trained and $\mathbf{S}$ can be calculated. Let us assume that in the contribution matrix $\mathbf{S}$, $S_{1,1,1}, S_{2,3,1}, S_{2,2,2}$, and $S_{3,1,2}$ are close to 1, with all other entries close to 0. In the inference matrix $\mathbf{W}$, $W_{1,2}$ and $W_{2,2}$ are close to 1 while $W_{1,1}$ and $W_{2,1}$ are close to 0. From the given $\mathbf{S}$ and $\mathbf{W}$, we can summarize two rules from the trained network as: (1) IF $x_1$ is low and $x_2$ is high, THEN the sample is positive; (2) IF $x_2$ is medium and $x_3$ is low, THEN the sample is positive.

The above two rules are represented in $(\mathbf{S}_{:,:,1}, W_{1,:})$ and $(\mathbf{S}_{:,:,2}, W_{2,:})$, respectively. The definitions of "low", "medium" and "high" concepts can be extracted from the parameters in the encoding module. The extracted rules mimic human logic. They can be used to justify the network's decisions and contribute to knowledge discovery.

In practice, the trained model may have some redundant rules, which means the representation of several rules are identical. For example, both Rule 1 and Rule 2 show that when $x_1$ is low, $x_2$ is high, then the sample is positive. From the current method formulation, this scenario can exist without harming the classifier's performance. However, in practice, a model that provides a small set of humanly understandable rules is favorable as it can be more easily used to provide guidance and reasoning to decision-makers. In this study, the correlations between each pair of rules are calculated. The correlation will be minimized during the training process. In addition, rules with high correlation and concepts with smaller contribution values are removed for rule visualization. The thresholds are chosen empirically.

### F. Model Training and Network Initialization

The proposed network is trained by back-propagation with an Adam optimizer. A regular cross-entropy loss $loss_{cs}$ is calculated to train the classification model. Additionally, an $\ell_1$ norm-based regularization term $loss_{\ell_1}$ is added to the loss function to favor rules with a smaller number of concepts, which are more feasible to use in practice. In addition, the correlation among encoded rules is calculated as a loss term

$loss_{corr}$ to avoid extracting redundant rules. The loss function can be written as:

$$loss_{total} = loss_{ce} + \lambda_1 loss_{\ell_1} + \lambda_2 loss_{corr}, \tag{11a}$$

$$loss_{l1} = \|vec(\mathbf{A})\|_1 + \|vec(\mathbf{M})\|_1, \tag{11b}$$

$$loss_{corr} = \sum_{i=1}^{H-1} \sum_{j=i+1}^{H} vec(\mathbf{S}_{:,:,i}) vec(\mathbf{S}_{:,:,j}) \tag{11c}$$

where $\lambda_1$ and $\lambda_2$ control the magnitude of the $\ell_1$ norm-based regularization term and correlation based regularization term, respectively. $vec(\cdot)$ denotes the vectorization of a matrix.

In this study, for simplicity, $\epsilon_1, \epsilon_2, \epsilon_3$ are constrained to be equal. They are initialized as 0.99 at the beginning of training and are gradually reduced with the number of training steps. The scheduling of the $\epsilon$ values can be written as

$$\epsilon = \max(\epsilon_{min}, \epsilon \cdot \gamma^{\text{training\_steps}}), \tag{12}$$

where $\gamma$ is the decay rate that can be tuned as a hyperparameter. From our preliminary analysis, $\gamma = 0.99$ usually is a good choice. $\epsilon_{\min}$ is another hyperparameter, whose optimal value varies with different applications. The hyperparameter tuning strategy will be discussed in the next section. Our experiments show that starting with $\epsilon = 0.99$ and reducing $\epsilon$ improves model optimization (as discussed in Section V-A).

Before model training, trainable parameters will be randomly initialized. To improve performance, especially when the size of the training dataset is small, practical rules from domain knowledge can be used to initialize the network. Revisiting the toy example in Section III-E, if the extracted rules were instead previously known within the application domain, the matrices $\mathbf{A}, \mathbf{M}$, and $\mathbf{W}$ in the network could then be initialized as:

- $\mathbf{A}$: $A_{1,1,1}, A_{2,3,1}, A_{2,2,2}, A_{3,1,2}$ have a higher value and other entries in $A_{:,:,1}$ and $A_{:,:,2}$ have a lower value;
- $\mathbf{M}$: $M_{1,1}, M_{2,1}, M_{2,2}, M_{3,2}$ have a higher value and other entries in $M_{:,1}$ and $M_{:,2}$ have a lower value;
- $\mathbf{W}$: $W_{1,2}, W_{2,2}$ have a high value and $W_{1,1}, W_{2,1}$ have a low value;
- Other entries in $\mathbf{A}$, $\mathbf{M}$, and $\mathbf{W}$ are randomly initialized.

## IV. Datasets and Experimental Settings

### A. Synthetic Datasets

Two synthetic datasets were built by simulating features with fixed distributions and rules to generate responses. The ground truth rules from the synthetic datasets can be used to assess a method's capability in extracting humanly understandable knowledge from the data and modeling the relationship between inputs and responses. In addition, with ground truth rules, synthetic datasets can be used to assess whether the proposed method can benefit from existing knowledge.

For each dataset, a 10-fold cross-validation was used for performance evaluation. In each iteration, the dataset was randomly split into the training set (64%), validation set (16%), and test set (20%).

*1) Synthetic Dataset 1:* Eight input variables were simulated as: $x_1 \sim \mathcal{N}(0,\ 2)$, $x_2 \sim \mathcal{N}(5,\ 3)$, $x_3 \sim \mathcal{N}(-1,\ 5)$, $x_4 \sim \mathcal{N}(1,\ 2)$, $x_5 \sim \mathcal{N}(-2,\ 1)$, $x_6 \sim$ Bernoulli$(0.5)$, $x_7 \sim \mathcal{N}(0,\ 1)$, $x_8 \sim \mathcal{N}(0,\ 1)$. If any of the following rules apply to one observation, then this observation is positive and otherwise negative:

- Rule A: $x_2 < 3.8$ and $x_3 > -2$ and $x_6 = 1$;
- Rule B: $x_2 > 6.3$ and $x_3 > -2$ and $x_6 = 1$;
- Rule C: $x_1 < 1$ and $x_4 > 2$ and $x_6 = 0$;
- Rule D: $x_3 > 0$ and $x_5 > -1$ and $x_6 = 0$;
- Rule E: $x_1 < 1$ and $x_5 > -1.5$ and $x_6 = 0$.

Additionally, random noise sampled from $\mathcal{N}(0,\ 0.01)$ are added to input variables. From the above rules we can readily observe that the response of one observation doesn't rely on $x_7$ and $x_8$. $x_7$ and $x_8$ are used as irrelevant variables to assess the model's resilience to redundant features.

*2) Synthetic Dataset 2:* Nine input variables were simulated as: $x_1 \sim \mathcal{N}(0,\ 2)$, $x_2 \sim \mathcal{N}(5,\ 3)$, $x_3 \sim \mathcal{N}(-1,\ 5)$, $x_4 \sim \mathcal{N}(1,\ 2)$, $x_5 \sim \mathcal{N}(-2,\ 1)$, $x_6 \sim \mathcal{N}(-1,\ 4.4)$, $x_7 \sim \mathcal{N}(0,\ 1.2)$, $x_8 \sim \mathcal{N}(0,\ 1)$, $x_9 \sim \mathcal{N}(0,\ 1)$. The sample is positive if $(x_1 + 0.5x_2 + x_3)^2/(1 + e^{x_6} + 2x_7) < 1$.

In this dataset, a highly non-linear function is used to assign the response. Though such a relationship between input variables and responses rarely exists for clinical applications, this dataset is used to determine if the proposed network can still achieve good performance by approximating the complicated relation as simple rules.

### B. Heart Failure Dataset

A HF dataset is created to train a classification model that identifies patients eligible for advanced therapies. Two cohorts were used in this study.

*1) REVIVAL Cohort:* The REVIVAL (Registry Evaluation of Vital Information for VADs in Ambulatory Life) registry contains information on 400 patients with advanced systolic HF from 21 US medical centers [25]. As part of the registry, patients were evaluated at up to 6 pre-specified time points over a 2-year period and underwent relevant examinations. At each time point, investigators were asked to record whether the participant had been evaluated for HT or LVAD and the result of that evaluation. For purposes of this analysis, study participants were labeled at each time point as appropriate (positive) or not appropriate (negative) for advanced therapies.

*2) INTERMACS Cohort:* The INTERMACS (Interagency Registry for Mechanically Assisted Circulatory Support) registry is a North American registry of adults who received an FDA approved durable MCS device for the management of advanced HF [26], [27], [28]. The registry includes clinical data on all adults $\geq$ 19 years of age who received a device at one of 170 active INTERMACS centers. The registry includes information on patient demographics, clinical data before and at the time of MCS implantation, and clinical outcomes up to one year post-MCS implantation or until HT. For this analysis, data was extracted at the time of LVAD implantation and patients classified as "appropriate for advanced therapies."

*3) Combined Dataset and Variable Selection:* Patients from the two cohorts were combined to form a larger dataset. HF clinicians selected 22 variables used in clinical practice which were in both datasets and which have strong associations with advanced

TABLE I
DATA SPLIT ON THE HEART FAILURE DATASET†. VALUES ARE PRESENTED AS AVERAGE NUMBER OF SAMPLES (AVERAGE NUMBER OF PATIENTS)

| | Training set | Validation set | Test set |
|---|---|---|---|
| REVIVAL with advanced therapy | 0 (0) | 46 (31) | 50 (31) |
| REVIVAL w/o advanced therapy | 782 (228) | 176 (52) | 181 (54) |
| INTERMACS | 7781 (7781) | 0 (0) | 0 (0) |

†The distribution of patient INTERMACS level and the the rationale of the proposed data split strategy are presented in Appendix B.

HF. These include heart rate, systolic blood pressure (SYSBP), sodium concentration, albumin concentration, uric acid concentration, total distance walked in 6 minutes (DISTWLK), gait speed during a 15 feet walk test, left ventricular dimension in diastole (LVDEM), left ventricular ejection fraction (EF), mitral regurgitation (MITRGRG), lymphocyte percentage (LYMPH), total cholesterol (TCH), hemoglobin (HGB), age, sex, comorbidity index, glomerular filtration rate (GFR), pulse pressure, treatment with cardiac resynchronization therapy (AR), need for temporary MCS device, treatment with guideline directed medical therapy (GDMT) for heart failure, and peak oxygen consumption during a maximal cardiopulmonary exercise test (pVO2). Appendix A provides more details of the clinical variables.

*4) Clinical Rules:* To facilitate model training, we assembled a panel of five HF and transplant cardiologists, all from different institutions. Two cardiologists were first asked to generate a set of clinical rules using the aforementioned variables. These were then collated and distributed to three additional cardiologists for review and additional rules were added as indicated, creating a final set of consensus rules. For this demonstration study, rules were simplified as follows:

- Rule A: EF is low, and pVO2 is low;
- Rule B: EF is low, and DISTWLK is low;
- Rule C: Age is high, EF is low, and SYSBP is low;
- Rule D: EF is low, and MITRGRG is high;
- Rule E: EF is low, and the GDMT is low;

### C. Experimental Settings

For synthetic datasets, 10-fold cross-validation was used to evaluate model performance. For the heart failure dataset, to better evaluate the model's generalizability on external dataset, the training set includes all samples from the INTERMACS registry and 80% of the negative samples from the REVIVAL registry. The remaining negative samples and all positive samples from the REVIVAL registry were equally and randomly split into the validation and test sets. The proposed data split was randomly repeated 10 times to evaluate the model. For one repetition, samples from the same patient will only exist in one set. The average number of patients and samples in subsets are presented in Table I. The rationale of the proposed data split strategy on the heart failure dataset is presented in Appendix B.

TABLE II
PERFORMANCE OF THE PROPOSED MODEL ON SYNTHETIC DATASET 1 WITH $N = 400$ USING 10-FOLD CROSS-VALIDATION

| Model | Accuracy | Recall | Precision | F1 | AUC |
|---|---|---|---|---|---|
| $\epsilon_{min}$ = 0.8 | 0.955 (0.025) | 0.911 (0.073) | 0.955 (0.038) | 0.883 (0.040) | 0.986 (0.016) |
| $\epsilon_{min}$ = 0.4 | 0.959 (0.030) | 0.904 (0.073) | **0.972 (0.035)** | 0.888 (0.048) | 0.991 (0.010) |
| $\epsilon_{min}$ = 0.2 | 0.961 (0.026) | **0.919 (0.087)** | 0.968 (0.039) | **0.892 (0.045)** | **0.992 (0.008)** |
| Fixed $\epsilon$ = 0.8 | **0.966 (0.023)** | 0.903 (0.083) | 0.964 (0.019) | 0.886 (0.037) | 0.978 (0.019) |
| Fixed $\epsilon$ = 0.4 | 0.939 (0.040) | 0.867 (0.086) | 0.948 (0.056) | 0.857 (0.064) | 0.964 (0.024) |
| Fixed $\epsilon$ = 0.2 | 0.786 (0.041) | 0.519 (0.190) | 0.803 (0.109) | 0.558 (0.132) | 0.819 (0.117) |

For the first three rows, $\epsilon$ was initialized to 0.99 and was gradually reduced to $\epsilon_{\min}$ during training. For the last three rows, the value of $\epsilon$ was fixed.

A random search algorithm was applied using the training set and validation set for hyperparameter tuning, including the number of rules $K$, learning rate, batch size, $\lambda_1$, $\lambda_2$, and $\epsilon_{min}$. The model trained with the optimal combinations of hyperparameters was then evaluated on the test set. The performance of the proposed network will be presented as the average and standard deviation (std) from 10 iterations.

For comparison, several popular "black box" machine learning algorithms were chosen, including random forest, SVM, and XGBoost. In addition, several interpretable models were chosen including logistic regression, fuzzy inference classifier [29], XGBoost-based decision tree [30], GAMI-Net [31], and Explainable Boosting Machine (EBM) [32]. Those models have the same hyper-parameter tuning process and model evaluation as the proposed algorithm. Class weights are used when the dataset is unbalanced.

Accuracy, recall, precision, F1, and area under the ROC curve (AUC) were calculated to evaluate the performance of the trained classifiers.

## V. RESULTS AND DISCUSSION

### A. Synthetic Dataset 1 ($N = 400$)

Let $N$ denote the number of observations in a given dataset. Several experiments were performed with differently sized simulated datasets. In this section, we discuss the performance of the proposed method on synthetic dataset 1 when $N = 400$.

The first experiment starts with $N = 400$. The percentage of positive samples is 34.25%, and the percentages of samples with Rule A, Rule B, Rule C, Rule D, Rule E are 8.25%, 7.50%, 9.00%, 2.00%, and 10.75%, respectively.

Table II depicts the performance of the proposed algorithm with different $\epsilon_{min}$ on the test sets from 10-fold cross-validation. We can observe that model training benefited from decreasing $\epsilon_{min}$ from 0.8 to 0.2, but the performance of the trained model decreased when $\epsilon_{min}$ was decreased to 0.1. We also evaluated the model with a fixed $\epsilon$, rather than gradually decreasing it from 0.99. While fixing $\epsilon$ at 0.8 leads to comparable performance with the model using $\epsilon_{min} = 0.8$, the performance of the models

TABLE III
PERFORMANCE OF ML METHODS ON SYNTHETIC DATASET 1 WITH $N = 400$ USING 10-FOLD CROSS-VALIDATION

| Model | Accuracy | Recall | Precision | F1 | AUC | Transparent |
|---|---|---|---|---|---|---|
| Proposed | 0.960 (0.023) | 0.933 (0.054) | 0.953 (0.060) | 0.893 (0.032) | 0.994 (0.005) | Yes |
| Logistic Regression | 0.724 (0.029) | 0.344 (0.078) | 0.692 (0.098) | 0.413 (0.070) | 0.701 (0.065) | Yes |
| Fuzzy Inference Classifier | 0.680 (0.036) | 0.456 (0.102) | 0.540 (0.076) | 0.441 (0.071) | 0.668 (0.056) | Yes |
| XGBoost-based Decision Tree | 0.904 (0.053) | 0.814 (0.145) | 0.894 (0.066) | 0.798 (0.094) | 0.956 (0.040) | Yes |
| EBM | 0.835 (0.027) | 0.678 (0.060) | 0.807 (0.060) | 0.688 (0.045) | 0.924 (0.018) | Yes |
| GAMI-Net | 0.754 (0.063) | 0.474 (0.193) | 0.637 (0.133) | 0.497 (0.123) | 0.748 (0.058) | Yes |
| Random Forest | 0.924 (0.015) | 0.826 (0.062) | 0.944 (0.037) | 0.832 (0.028) | 0.981 (0.006) | No |
| XGBoost | **0.977 (0.013)** | **0.959 (0.031)** | **0.975 (0.028)** | **0.919 (0.020)** | **0.996 (0.003)** | No |
| SVM | 0.821 (0.038) | 0.641 (0.076) | 0.796 (0.077) | 0.661 (0.061) | 0.897 (0.026) | No |

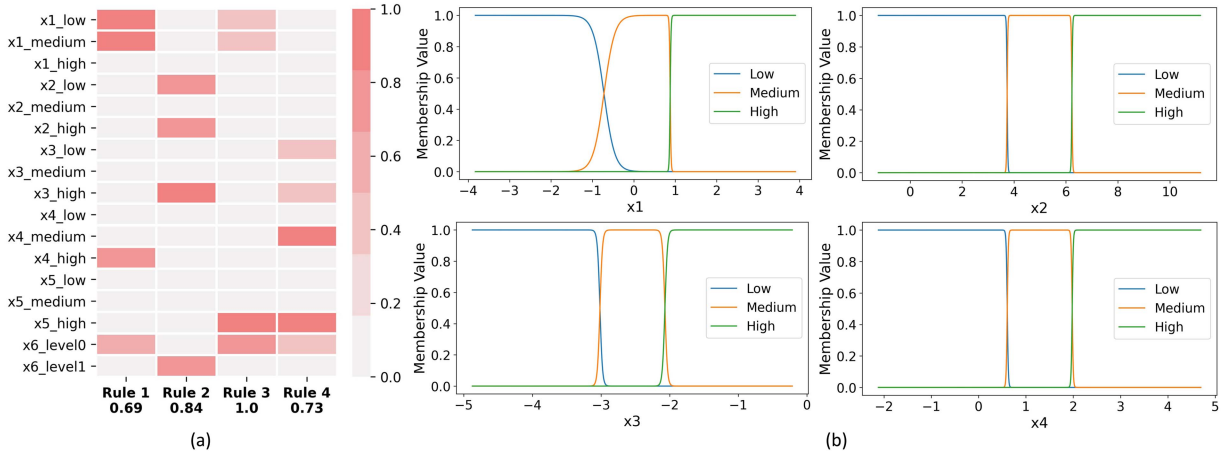


Fig. 2. Interpretation of a trained model on synthetic dataset 1 with $N = 400$. (a) Visualization of four rules contributing to the positive class, which are summarized from the trained model. Rules are visualized in individual columns with each row corresponding to concept. For example, "x1_low" means "the value of $x_1$ is low". The contribution of individual concepts to individual rules are shown in color. (b) Membership functions for "low", "medium", and "high" concepts of $x_1, x_2, x_3$, and $x_4$ in the encoding module, respectively.

with a smaller fixed $\epsilon$ value decreased significantly. Our results show the effectiveness of the algorithm that gradually decreases $\epsilon$ during the training. Using this dataset, the proposed network with a reasonable degree of piecewise linearity has a better performance.

Table III describes the performance of the proposed method where $\epsilon_{min}$ was tuned on the validation set in each iteration. The performance of the proposed network is compared with that of other machine learning algorithms. From Table III, we can see that the proposed network achieved significantly better performance than other interpretable models and had comparable performance to the XGBoost model, which is the best among the other established machine learning algorithms.

To examine the proposed network's ability to learn rules from the dataset, we summarized rules contributing to the positive class from a trained network. Those rules are visualized in Fig. 2(a). Comparing the learned rules with rules in Section IV-A1, we can observe that Rule 1 corresponds to Rule C; Rule 2 corresponds to a union of Rule A and Rule B; Rule 3 corresponds to Rule E; and Rule 4 is closest to Rule D. Membership functions of the variables involved in Rule 1 and Rule 2 are visualized in Fig. 2(b) and we can observe a great match. For example, the membership value of $x_2$ to the "low" concept is high when $x_2$ smaller than 3.7 and the membership value of $x_2$ to the "high" concept is high when $x_2$ is larger than 6.2. Simple thresholds were used to construct synthetic dataset 1, and for this reason, the fuzzy regions in the membership functions are very narrow. From the interpretation in Fig. 2, the trained model learned the majority of rules used to construct the dataset. Rule 4 is close to Rule D but with two additional concepts that are misidentified as related to the class. This may be due to only 2.00% of samples in the dataset being consistent with Rule D, making it more challenging to learn from the data. In addition, from Fig. 2(a), concepts from $x_7$ and $x_8$ are not shown because their significance to learned rules is too low. This demonstrates that the proposed network can identify and exclude irrelevant variables.

### B. Synthetic Dataset 1 ($N = 50$)

In the second experiment, we used synthetic dataset 1 with $N = 50$. The percentage of positive samples is 42.00%, and the percentages of samples with Rules A-E are 14.00%, 14.00%, 4.00%, 4.00%, and 12.00%, respectively. In this experiment, we investigated the performance of the proposed network with a small training set and if initiating the network with existing knowledge would enable the model to learn more accurate rules. Limited training data is a common issue in medical applications, which may result from the small patient population or tedious / expensive annotation collection. Considering that domain

TABLE IV
PERFORMANCE OF ML METHODS ON THE SYNTHETIC DATASET 1 WITH $N = 50$ USING 10-FOLD CROSS-VALIDATION

| Model | Accuracy | Recall | Precision | F1 | AUC | Transparent |
|---|---|---|---|---|---|---|
| Proposed (None) | 0.640 (0.143) | 0.550 (0.292) | 0.518 (0.249) | 0.473 (0.236) | 0.688 (0.213) | Yes |
| Proposed (Rule A) | 0.670 (0.110) | 0.575 (0.275) | 0.543 (0.238) | 0.504 (0.223) | 0.710 (0.188) | Yes |
| Proposed (Rule B) | 0.670 (0.135) | 0.600 (0.255) | 0.646 (0.211) | 0.535 (0.170) | 0.658 (0.183) | Yes |
| Proposed (Rule C) | 0.690 (0.104) | 0.625 (0.202) | 0.658 (0.197) | 0.566 (0.129) | 0.698 (0.158) | Yes |
| Proposed (Rule D) | 0.730 (0.142) | **0.675 (0.251)** | 0.658 (0.282) | **0.607 (0.225)** | 0.710 (0.194) | Yes |
| Proposed (Rule E) | 0.700 (0.190) | 0.600 (0.229) | 0.710 (0.259) | 0.573 (0.202) | **0.740 (0.191)** | Yes |
| Proposed (Rule F, partially correct) | 0.680 (0.183) | 0.600 (0.200) | 0.665 (0.278) | 0.565 (0.196) | 0.688 (0.206) | Yes |
| Proposed (Rule G, partially correct) | 0.700 (0.210) | 0.625 (0.280) | 0.605 (0.308) | 0.566 (0.276) | 0.652 (0.213) | Yes |
| Proposed (Rule H, partially correct) | **0.750 (0.112)** | 0.575 (0.195) | **0.775 (0.197)** | 0.593 (0.176) | **0.740 (0.152)** | Yes |
| Logistic Regression | 0.610 (0.145) | 0.425 (0.275) | 0.512 (0.339) | 0.395 (0.236) | 0.583 (0.181) | Yes |
| Fuzzy Inference Classifier | 0.520 (0.117) | 0.525 (0.208) | 0.416 (0.120) | 0.413 (0.146) | 0.550 (0.103) | Yes |
| XGBoost-based Decision Tree | 0.530 (0.174) | 0.375 (0.256) | 0.343 (0.247) | 0.318 (0.230) | 0.548 (0.183) | Yes |
| EBM | 0.650 (0.120) | 0.500 (0.224) | 0.562 (0.260) | 0.469 (0.192) | 0.670 (0.151) | Yes |
| GAMI-Net | 0.610 (0.145) | 0.300 (0.245) | 0.525 (0.202) | 0.315 (0.223) | 0.595 (0.110) | Yes |
| Random Forest | 0.650 (0.081) | 0.475 (0.236) | 0.580 (0.275) | 0.450 (0.176) | 0.619 (0.168) | No |
| XGBoost | 0.650 (0.186) | 0.600 (0.300) | 0.591 (0.275) | 0.521 (0.238) | 0.675 (0.187) | No |
| SVM | 0.580 (0.075) | 0.125 (0.230) | 0.250 (0.403) | 0.130 (0.204) | 0.521 (0.173) | No |

knowledge usually exists in the medical field, this experiment is to demonstrate that the proposed method can do well when the training set is small.

Table IV has three blocks, presenting the performance of the proposed networks, established interpretable ML methods, and established black-box ML methods on synthetic dataset 1 ($N = 50$), respectively. The first block shows the performance of the proposed network without and with existing knowledge. The performance of the proposed network with random initialization is shown in the first row of the first block, followed by the performance of the proposed network initialized with existing knowledge (rules). Rules A through E are fully correct as described in Section IV-A1 while Rules F through H are partially correct. In practical applications, it is very rare that the ground truth rule is available. As such, in this experiment, we only initialized $\mathbf{A}$, $\mathbf{M}$, and $\mathbf{W}$, while the parameters in the membership functions were randomly initialized. In addition, to investigate whether inexact domain knowledge can facilitate model training, we proposed the following three rules and assumed they lead to a positive class:

- Rule F: $x_2$ is "low" and $x_6 = 1$;
- Rule G: $x_1$ is "low" and $x_5$ is "low" and $x_6 = 0$;
- Rule H: $x_1$ is "low" and $x_5$ is "high" and $x_6 = 0$ and $x_7$ is "high";

Rule F, G, and H are only partially correct. Compared with ground truth Rule A, the "high" concept of $x_3$ is missing in Rule F. In Rule G, $x_5$ should be "high" rather than "low" as in Rule E. In Rule H, "high" concept of $x_7$ is actually irrelevant to the class.

From Table IV, we first observe that because of the reduction in the size of the training set, performance decreased. Still, XGBoost achieves the best performance, and the proposed network with random initialization has a comparable performance to XGBoost. Second, we observe that the improvement can be achieved when the network was initialized with Rules A through E. Third, the model's performance increased when it was initialized with partially correct rules. This indicates that existing domain knowledge can help with model training even when the rules are vague and/or inexact.

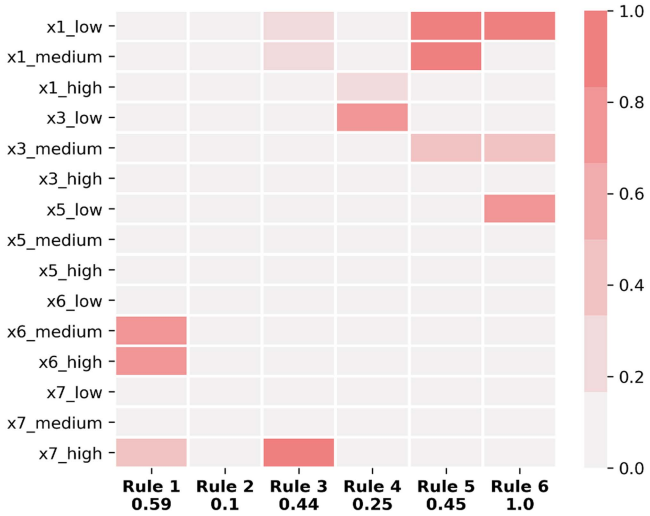


Fig. 3. Interpretation of a trained model on synthetic dataset 2 with $N = 400$.

### C. Synthetic Dataset 2 ($N = 400$)

The responses in synthetic dataset 1 were constructed by rules, where a rule-based or tree-based machine learning algorithm may be more favorable. Therefore, responses in synthetic dataset 2 were built from a non-linear function to further explore the capacity of the proposed network in function approximation. A performance comparison of different ML models is presented in Table V. From the table, we can see that SVM achieved the best performance. The performance of the proposed network is lower than SVM but comparable with other machine learning algorithms.

Rules extracted from the trained proposed network are presented in Fig. 3. We see that these rules capture meaningful information. Observations in this dataset were annotated as positive if $(x_1 + 0.5x_2 + x_3)^2/(1 + e^{x_6} + 2x_7) < 1$. Rule 1 shows that "high" levels of $x_6$ and $x_7$ lead to the positive class. In this dataset, $x_1$, $x_2$, and $x_3$ were simulated as: $x_1 \sim \mathcal{N}(0, 2)$, $x_2 \sim \mathcal{N}(5, 3)$, and $x_3 \sim \mathcal{N}(-1, 5)$. As such, a "high" $x_1$ and

TABLE V
PERFORMANCE OF ML METHODS ON THE SYNTHETIC DATASET 2 WITH $N = 400$ USING 10-FOLD CROSS-VALIDATION

| Model | Accuracy | Recall | Precision | F1 | AUC | Transparent |
|---|---|---|---|---|---|---|
| Proposed | 0.714 (0.041) | 0.738 (0.067) | 0.693 (0.062) | 0.657 (0.045) | 0.801 (0.040) | Yes |
| Logistic Regression | 0.746 (0.046) | 0.703 (0.084) | 0.738 (0.053) | 0.671 (0.058) | 0.774 (0.073) | Yes |
| Fuzzy Inference Classifier | 0.654 (0.048) | 0.408 (0.090) | 0.721 (0.076) | 0.475 (0.084) | 0.761 (0.037) | Yes |
| XGBoost-based Decision Tree | 0.722 (0.096) | 0.617 (0.108) | 0.493 (0.105) | 0.521 (0.124) | 0.745 (0.057) | Yes |
| EBM | 0.736 (0.028) | 0.686 (0.047) | 0.731 (0.044) | 0.660 (0.028) | 0.826 (0.042) | Yes |
| GAMI-Net | 0.749 (0.039) | 0.697 (0.056) | 0.747 (0.058) | 0.673 (0.016) | 0.805 (0.016) | Yes |
| Random Forest | 0.734 (0.040) | 0.692 (0.030) | 0.726 (0.058) | 0.660 (0.034) | 0.827 (0.035) | No |
| XGBoost | 0.734 (0.043) | 0.705 (0.072) | 0.714 (0.043) | 0.662 (0.054) | 0.837 (0.033) | No |
| SVM | **0.781 (0.074)** | **0.741 (0.077)** | **0.780 (0.094)** | **0.712 (0.079)** | **0.871 (0.066)** | No |

TABLE VI
PERFORMANCE OF ML METHODS ON THE TEST SET OF THE HEART FAILURE DATASET FROM 10 REPETITIONS

| Model | Accuracy | Recall | Precision | F1 | AUC | Transparent |
|---|---|---|---|---|---|---|
| Proposed (None) | 0.735 (0.047) | 0.500 (0.069) | 0.384 (0.059) | 0.386 (0.047) | 0.730 (0.042) | Yes |
| Proposed (with existing rules) | 0.718 (0.035) | 0.645 (0.125) | 0.410 (0.045) | **0.452 (0.043)** | 0.753 (0.025) | Yes |
| Logistic Regression | 0.773 (0.022) | 0.285 (0.084) | 0.459 (0.089) | 0.297 (0.079) | 0.719 (0.049) | Yes |
| Fuzzy Inference Classifier | 0.506 (0.124) | 0.788 (0.151) | 0.298 (0.080) | 0.358 (0.054) | 0.707 (0.071) | Yes |
| XGBoost-based Decision Tree | 0.719 (0.018) | 0.430 (0.055) | 0.395 (0.032) | 0.369 (0.031) | 0.715 (0.039) | Yes |
| EBM | 0.752 (0.010) | 0.444 (0.071) | 0.455 (0.033) | 0.402 (0.042) | 0.737 (0.036) | Yes |
| GAMI-Net | 0.719 (0.020) | 0.347 (0.093) | 0.490 (0.045) | 0.355 (0.059) | 0.701 (0.014) | Yes |
| Random Forest | 0.759 (0.041) | 0.590 (0.088) | 0.458 (0.047) | 0.448 (0.051) | **0.801 (0.042)** | No |
| XGBoost | 0.764 (0.021) | 0.444 (0.080) | 0.455 (0.065) | 0.402 (0.065) | 0.756 (0.048) | No |
| SVM | 0.746 (0.040) | 0.447 (0.086) | 0.438 (0.080) | 0.381 (0.029) | 0.710 (0.048) | No |

"low" $x_3$ can lead $(x_1 + 0.5x_2 + x_3)^2$ to a small value. A "low" or "medium" $x_1$ and "medium" $x_3$ is another combination that can lead $(x_1 + 0.5x_2 + x_3)^2$ to a small value. As expected, Rules 4 and 5 unite concepts from $x_1$ and $x_3$. From this analysis, we observe that the proposed network can learn simple rules in a format that humans can understand from a dataset that was constructed with a complicated non-linear function.

### D. Heart Failure Dataset

We applied the proposed network to identify patients that are eligible for advanced therapies. Table VI presents the performance of the proposed method and other techniques. In this particular application, we want to have a model that is less likely to miss patients that are eligible for advanced therapies, yet provides a reasonably high probability that referred patients will subsequently be deemed appropriate for an advanced therapy. F1 score is the best evaluation metric because the balance between recall and precision is important, and the dataset is unbalanced.

From Table VI, initializing the network with existing knowledge can greatly improve the model's performance. The proposed method had the highest AUC and F1 score compared with other interpretable learning approaches. Compared with "black-box" methods, the proposed method without existing knowledge achieved a comparable F1 score and AUC. With existing knowledge, the proposed method had the highest F1 score and a comparable AUC. We also found that "black-box" methods have higher generalization errors (more than 20% in F1 score) between the validation set and test set. In contrast, the proposed method had a significantly smaller generalization error (less than 5% in F1 score). Notably, integrating existing domain knowledge can not only improve the classification performance but also further reduce the generalization error.

Fig. 4(a) shows the learned rules of the trained model initialized with existing knowledge. The learnt membership functions of the continuous / ordinal variables with high contribution are shown in Fig. 4(b). The learned rules were compared with manually curated rules and presented to clinicians for a qualitative review. Concepts that exist in the manually curated rules such as "low" EF, "low" pVO2, "low" DISTWLK, "high" MITRGRG were also captured by the proposed method. These learned rules approximated those provided by heart failure cardiologists though in unique combinations and with additional learned features. All of the rules from heart failure cardiologists included a reduced ejection fraction and an objective marker of significant functional limitations, most often by cardiopulmonary exercise testing. As seen in Fig. 4, almost all rules learned by the model included ejection fraction as well as a second variable denoting a patient's functional tolerance, either by cardiopulmonary exercise testing, 6-minute walk distance, or by gait speed. Notably, while gait speed is an objective and valid measure of functional capacity, it was not included in any of the provided rules and thus represents learned knowledge.

## VI. CONCLUSION

In this study, we proposed a novel machine learning model that is transparent and interpretable. The proposed network was tested on both synthetic datasets and a real-world dataset. Our experimental results show that (1) the model can learn hidden rules from the dataset and represent them in a way that humans understand; and (2) initializing the network with approximate

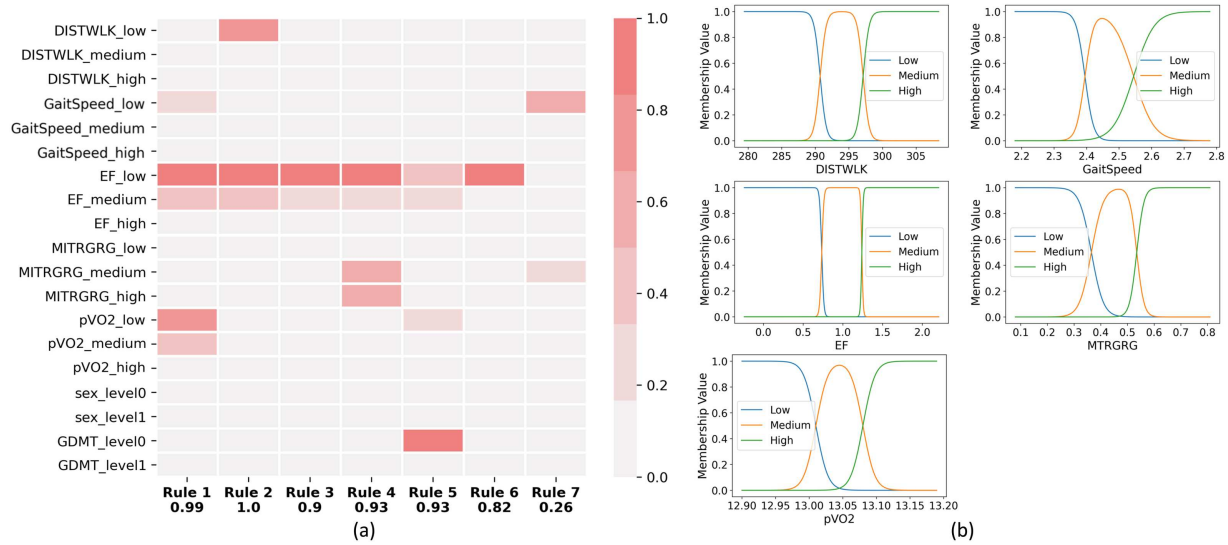


Fig. 4. Interpretation of a trained model on the heart failure dataset. (a) Rule visualization; (b) Trained membership functions for continuous/ordinal variables involved in the rules shown in (a).

domain knowledge can effectively improve model performance, especially when the size of the training set is limited. Notably, the proposed network shows significantly improved generalizability when identifying patients with heart failure who would benefit from advanced therapies. The proposed algorithm is promising in building multiple other clinical (and non-clinical) decision-making applications.

The proposed algorithm is optimized by stochastic gradient descent and the $\epsilon$s are reduced gradually during the training process. As a result, it takes a longer time in model training. For general applications where the interpretability is not critical, we still think existing machine learning algorithms such as XGBoost, random forest, and SVM are good choices as they can achieve good classification accuracy and are computationally efficient. However, for sensitive applications, e.g., medicine, the proposed method has its unique strengths as discussed above. Its capability in rule extraction and representation improves the model's transparency and transferability. It also has the potential to help the discovery of knowledge.

For the heart failure application, limitations exist in dataset size and patient population distribution. In the INTERMACS dataset, half of the data samples are from patients in critical conditions, and we don't have information on medication intolerance. In this study, the heart failure application is presented as proof of principle. The proposed method has multiple potential uses in other important and sensitive clinical applications outside of HF care, such as patient classification, outcome prediction, treatment efficiency estimation, and disease grade classification.

## Appendix A
## Variables in the Heart Failure Dataset

Definition of GDMT: Yes, if the patient has been on >2 categories of the appropriate evidenced-based heart failure medications: (1) ACE inhibitor or Angiotensin receptor blocker or sacubitril/valsartan (LCZ); (2) Beta-blocker; (3) Aldosterone antagonist. In this study, patients not on these therapies were assumed to have contraindication or intolerance.

Definition of the ordinal EF score: 1: 20≥EF ≥29; 2: 30≥ EF≥39; 3: 40≥EF≥49; 4: EF≥50.

The distribution of continuous/ordinal variables in the heart failure dataset is shown in Table VII.

## Appendix B
## Data Split on the Heart Failure Application

Two registries were used in the HF application. While the INTERMACS registry has more severe heart failure cases, the REVIVAL registry was specifically designed to evaluate an ambulatory population. In the INTERMACS dataset, the distribution of INTERMACS levels 1–2 (critical), 3 (stable), 4–7 (ambulatory) are 45.0%, 37.4%, 17.6%, respectively; in REVIVAL dataset, the distribution of INTERMACS levels 1–2 (critical), 3 (stable), 4–7 (ambulatory) are 0.1%, 1.3%, 98.6%, respectively. The REVIVAL registry was specifically designed to evaluate an ambulatory population [25].

Thus, the REVIVAL dataset serves as a more challenging dataset for distinguishing the positive samples from negative samples. And it can be used to validate the possibility of using this tool to streamline referrals from primary and secondary care to specialized HF centers. Since the REVIVAL dataset has a limited number of positive samples, we introduced the INTERMACS databases in the method development - it enriches the severe heart failure cases and includes more patient variability.

To better train and validate the ML algorithms, we proposed a data split strategy that only included data samples from the REVIVAL registry in the validation set and test set (shown in Table I).

TABLE VII
DISTRIBUTION OF THE CONTINUOUS/ORDINAL VARIABLES

| | REVIVAL (postive) | REVIVAL (negative) | INTERMACS |
|---|---|---|---|
| HR (beats/min) | 76.0 (64.0-86.0) | 74.0 (66.0-83.0) | 87.0 (75.0-100.0) |
| SYSBP (mmHg) | 98.0 (92.0-108.0) | 110.0 (100.0-120.0) | 105.0 (96.0-115.0) |
| Sodium (mEq/L) | 138.0 (136.0-140.0) | 139.9 (137.0-141.0) | 136.0 (133.0-138.0) |
| Albumin (g/dL) | 4.0 (3.7-4.3) | 4.2 (3.9-4.4) | 3.5 (3.1-3.9) |
| DISTWLK (feet) | 272.0 (185.9-320.3) | 357.0 (300.0-418.7) | 243.8 (163.4-321.0) |
| Gait speed (feet/second) | 3.0 (2.5-3.8) | 3.8 (3.1-4.4) | 2.5 (1.8-3.3) |
| LVDEM (mm) | 69.0 (63.0-75.4) | 66.0 (60.0-73.8) | 68 (61.0-75.0) |
| EF score (range [1-4]) | 1.0 (1.0-2.0) | 2.0 (1.0-2.0) | 1.0 (1.0-2.0) |
| MITRGRG (range [0-3]) | 2.0 (1.0-3.0) | 1.0 (0.0-2.0) | 2.0 (1.0-2.0) |
| pVO2 (mL/kg/min) | 10.6 (8.6 - 11.2) | 13.9 (11.4-16.5) | 11.0 (9.0-13.0) |
| Uric acid (mg/dL) | 8.1 (6.5 - 10.4) | 7.8 (6.4-9.6) | 8.1 (6.1-10.1) |
| LYMPH (%) | 18.7 (13.9-23.9) | 22.0 (16.1-28.1) | 17.0 (11.0 - 23.7) |
| TCH (mg/dL) | 141.0 (110.0-172.0) | 154.0 (128.0-189.0) | 125.0 (100.0-154.0) |
| HGB (g/dL) | 12.5 (11.5-13.8) | 13.5 (12.4-14.5) | 11.4 (10.0-12.9) |
| Pulse pressure (mmHg) | 35.0 (33.0 - 39.3) | 41.0 (35.0-51.0) | 39.0 (31.0-49.0) |
| GFR (mL/min/1.73m2) | 44.2 (33.4-56.7) | 54.3 (41.5-70.4) | 59.1 (44.2-76.5) |
| Comorbidity (range [0-14]) | 3.0 (2.0 - 4.3) | 3.0 (2.0-4.0) | 3.0 (2.0-4.0) |
| Age (year) | 62.0 (57.0-68.3) | 61.0 (53.0-68) | 59.0 (50.0-66.0) |

Values are presented in median (25th-75th).

## REFERENCES


[1] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, "Artificial intelligence transforms the future of health care," *Amer. J. Med.*, vol. 132, no. 7, pp. 795–801, 2019.

[2] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, "Artificial intelligence transforms the future of health care," *Amer. J. Med.*, vol. 132, no. 7, pp. 795–801, Jul. 2019.

[3] M. A. Myszczynska et al., "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," *Nature Rev. Neurol.*, vol. 16, no. 8, pp. 440–456, 2020.

[4] H. Yao, C. Williamson, J. Gryak, and K. Najarian, "Automated hematoma segmentation and outcome prediction for patients with traumatic brain injury," *Artif. Intell. Med.*, vol. 107, 2020, Art. no. 101910.

[5] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthcare J.*, vol. 6, no. 2, pp. 94–98, 2019.

[6] L. A. Zadeh, "Fuzzy logic and approximate reasoning," *Synthese*, vol. 30, no. 3, pp. 407–428, 1975.

[7] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 1, pp. 116–132, Jan./Feb. 1985.

[8] K. Y. Chan, S.-H. Ling, T. S. Dillon, and H. T. Nguyen, "Diagnosis of hypoglycemic episodes using a neural network based rule discovery system," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9799–9808, 2011.

[9] R. Zhang and J. Tao, "A nonlinear fuzzy neural network modeling approach using an improved genetic algorithm," *IEEE Trans. Ind. Electron.*, vol. 65, no. 7, pp. 5882–5892, Jul. 2018.

[10] K. S. Parikh et al., "Heart failure with preserved ejection fraction expert panel report: Current controversies and implications for clinical trials," *JACC: Heart Failure*, vol. 6, no. 8, pp. 619–632, 2018.

[11] E. J. Benjamin et al., "Heart disease and stroke statistics—2018 update: A report from the American Heart Association," *Circulation*, vol. 137, no. 12, pp. e67–e492, 2018.

[12] M. Canepa et al., "Performance of prognostic risk scores in chronic heart failure patients enrolled in the European society of cardiology heart failure long-term registry," *JACC: Heart Failure*, vol. 6, no. 6, pp. 452–462, 2018.

[13] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[14] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau, "Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 1096–1106, Jan. 2020.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 4768–4777.

[16] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[17] Y. Lou, R. Caruana, and J. Gehrke, "Intelligible models for classification and regression," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2012, pp. 150–158.

[18] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," in *Proc. 18th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2015, pp. 1721–1730.

[19] J.-S. R. Jang, "ANFIS: Adaptive-network-based fuzzy inference system," *IEEE Trans. Syst., Man, Cybern.*, vol. 23, no. 3, pp. 665–685, May/Jun. 1993.

[20] A. F. Cabalar, A. Cevik, and C. Gokceoglu, "Some applications of adaptive neuro-fuzzy inference system (ANFIS) in geotechnical engineering," *Comput. Geotechnics*, vol. 40, pp. 14–33, 2012.

[21] M. Al-Mahasneh, M. Aljarrah, T. Rababah, and M. Alu'datt, "Application of hybrid neural fuzzy system (ANFIS) in food processing and technology," *Food Eng. Rev.*, vol. 8, no. 3, pp. 351–366, 2016.

[22] H. Yao, K. D. Aaronson, L. Lu, J. Gryak, K. Najarian, and J. R. Golbus, "Using a fuzzy neural network in clinical decision support for patients with advanced heart failure," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2019, pp. 995–999.

[23] M. Sugeno and K. Tanaka, "Successive identification of a fuzzy model and its applications to prediction of a complex system," *Fuzzy Sets Syst.*, vol. 42, no. 3, pp. 315–334, 1991.

[24] C.-T. Sun, "Rule-base structure identification in an adaptive-network-based fuzzy inference system," *IEEE Trans. Fuzzy Syst.*, vol. 2, no. 1, pp. 64–73, Feb. 1994.

[25] K. D. Aaronson et al., "Registry evaluation of vital information for VADs in ambulatory life (REVIVAL): Rationale, design, baseline characteristics, and inclusion criteria performance," *J. Heart Lung Transplant.*, vol. 39, no. 1, pp. 7–15, 2020.

[26] J. K. Kirklin et al., "INTERMACS database for durable devices for circulatory support: First annual report," *J. Heart Lung Transplant.*, vol. 27, no. 10, pp. 1065–1072, 2008.

[27] M. A. Miller, K. Ulisney, and J. T. Baldwin, "INTERMACS (interagency registry for mechanically assisted circulatory support): A new paradigm for translating registry data into clinical practice," *J. Amer. College Cardiol.*, vol. 56, no. 9, pp. 738–740, 2010.

[28] E. J. Molina et al., "The society of thoracic surgeons intermacs 2020 annual report," *Ann. Thoracic Surg.*, vol. 111, no. 3, pp. 778–792, 2021.

[29] S. K. Meher, "A new fuzzy supervised classification method based on aggregation operator," in *Proc. 3rd Int. IEEE Conf. Signal-Image Technol. Internet-Based Syst.*, 2007, pp. 876–882.

[30] O. Sagi and L. Rokach, "Approximating XGBoost with an interpretable decision tree," *Inf. Sci.*, vol. 572, pp. 522–542, 2021.

[31] Z. Yang, A. Zhang, and A. Sudjianto, "GAMI-Net: An explainable neural network based on generalized additive models with structured interactions," *Pattern Recognit.*, vol. 120, 2021, Art. no. 108192.

[32] H. Nori, S. Jenkins, P. Koch, and R. Caruana, "InterpretML: A unified framework for machine learning interpretability," 2019, *arXiv:1909.09223*.