

Inherently Interpretable Machine Learning Models for Time-Series: A Systematic Literature Review

Ruqayyah Nabage

Malmö University

Zahra Asghari Varzaneh

Lund University

Reza Malekian

Malmö University

Azra Abtahi

`azra.abtahi-fahliani@mau.se`

Malmö University

Research Article

Keywords: Time-series Data, Explainable AI, Interpretability, Interpretable Machine Learning, Transparent Models, Self-Explainable Models

Posted Date: September 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10985595/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Inherently Interpretable Machine Learning Models for Time-Series: A Systematic Literature Review

Ruqayyah Nabage^{1*}, Zahra Asghari Varzaneh,³
Reza Malekian,^{1,2} Azra Abtahi,^{1,2*}

^{1*}Department of Computer Science and Media Technology, Malmö
University, 205 06, Malmö, Sweden.

²Sustainable Digitalisation Research Centre, Malmö University, Malmö,
205 06, Sweden.

³Department of Translational Medicine, Lund University, 205 02,
Malmö, Sweden.

*Corresponding author(s). E-mail(s): ruqayyah.nabage@mau.se;
azra.abtahi-fahliani@mau.se;

Contributing authors: zahra.asghari_varzaneh@med.lu.se;
reza.malekian@mau.se;

Abstract

Time-series data underpins critical decision-making in domains such as health-care, finance, industrial processes, and environmental forecasting. As [Machine Learning \(ML\)](#) models are increasingly deployed in these domains, understanding how their decisions are made is essential for aligning model behaviour with domain logic and supporting expert trust. Existing approaches to model transparency often rely on post-hoc [Explainable Artificial Intelligence \(XAI\)](#) methods that explain predictions after they are generated. In contrast, inherently interpretable (ante-hoc) models embed interpretability into the model itself, enabling their internal mechanisms to directly support understanding of the resulting decisions. Despite growing interest in interpretable [ML](#), the landscape of inherently interpretable approaches for time-series data remains fragmented. To address this gap, we conduct a systematic literature review following the [Preferred Reporting Items for Systematic Reviews and Meta-Analyses \(PRISMA\)](#) methodology, synthesising 77 peer-reviewed studies. We clarify foundational concepts and time-series-specific challenges of interpretability and propose a taxonomy

of existing approaches based on their architectures and interpretability mechanisms, encompassing linear, tree-based, interpretable neural, attention-based, case- and rule-based, fuzzy, and neuro-fuzzy models. We further categorise the datasets used in the reviewed studies according to temporal modality, application domain, and problem type. Finally, we identify key open challenges and outline research directions toward transparent and reliable time-series models.

Keywords: Time-series Data, Explainable AI, Interpretability, Interpretable Machine Learning, Transparent Models, Self-Explainable Models

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) systems are increasingly being deployed in domains where their predictions influence important decisions, including healthcare, finance, transportation, manufacturing, and scientific research (Rudin 2019; Adadi and Berrada 2018). Recent advances in learning algorithms, computational resources, and data availability have enabled these models to achieve state-of-the-art performance across a wide range of tasks. However, many of the most accurate models operate as highly complex systems whose internal decision-making processes are difficult for humans to understand (Samek et al. 2017; Adadi and Berrada 2018; Molnar 2020). The opacity of these models presents challenges in applications where understanding how a prediction was produced is as important as the prediction itself, as stakeholders need to assess the reliability of model outputs, identify potential sources of bias, justify decisions, or satisfy regulatory requirements (Rudin 2019).

Subsequently, increasing attention has been directed towards approaches that improve the transparency and understandability of ML systems. Explainable Artificial Intelligence (XAI) as a field developed in response to this need for approaches that enable humans to understand algorithmic decisions better (Islam et al. 2022; Räuker et al. 2023; Felzmann et al. 2020). It encompasses a broad range of perspectives drawing from ML, human-computer interaction, psychology, philosophy, and ethics. As a result,

no single, universally accepted definition of [XAI](#) has been established, and related concepts are frequently interpreted differently throughout the literature. Nevertheless, a common core can be identified across most accounts: [XAI](#) is generally understood to encompass approaches capable of providing explanations for their decisions ([Theissler et al. 2022](#); [Molnar 2020](#)). The growing interest in [XAI](#) has also been influenced by ethical and regulatory considerations. For example, discussions surrounding the European Union’s ”right to explanation” have highlighted the need for transparent and accountable algorithmic decision-making processes ([Vilone and Longo 2020](#)).

The nature of this interpretability, however, varies considerably across existing approaches. There are two main approaches within [XAI](#) that provide different levels of insight into a model’s decision-making process: post-hoc explanation methods and inherently interpretable models.

Post-hoc explanation methods or tools generate explanations for a model’s decisions after it has been trained. They can provide useful information about model behaviour, without affecting the model’s predictions. These explanations are often external approximations of the underlying decision process rather than a direct reflection of how the model reasons ([Samek et al. 2017](#); [Larsson and Heintz 2020](#)). These approximations may not always faithfully represent the decision-making process of the underlying model. Furthermore, these methods prioritise predictive performance and flexibility over trust and faithfulness ([Gilpin et al. 2018](#); [Graziani et al. 2023](#); [Krishnan 2020](#); [Larsson and Heintz 2020](#)). These methods are also referred to as [XAI](#) techniques, and it is these methods that are typically meant when the term ”explainability” is used on its own in the literature. In this work, we use the term post-hoc explainability in reference to these methods.

In contrast, inherently interpretable models, also referred to as ante-hoc or self-explainable models in the literature, are models that are designed such that the

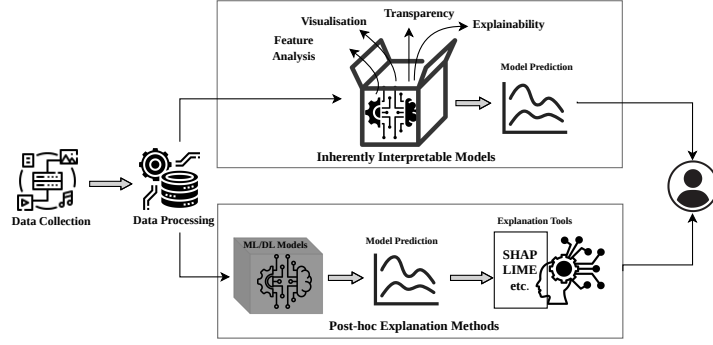


Fig. 1: A summary and comparison of inherently transparent models and post-hoc methods

reasoning process is incorporated into the model itself. These models prioritise transparency, trustworthiness, and faithful explanations from inception and treat them as main design variables. By making the decision-making mechanism transparent by design, such approaches can provide a stronger basis for understanding, validating, and trusting model predictions (Adadi and Berrada 2018; Arrieta et al. 2020). This distinction between inherently interpretable models and post-hoc explanation methods is further illustrated in Fig. 1.

These approaches, however, have not been developed equally across data modalities. Time-series data, in particular, underlies many safety-critical and high-stakes scenarios, like healthcare monitoring, industrial systems, or scientific research, where decisions depend not only on individual observations but also on patterns and dependencies that evolve over time. Unlike static data modalities such as images or tabular data, time-series data contains temporal relationships between observations. Consequently, the approaches used for image and tabular data cannot always be directly transferred to time-series (Theissler et al. 2022; Molnar 2020). This distinct temporal structure creates a specific need for interpretability approaches that account for the data’s sequential, time-dependent nature. Yet despite its prevalence in such high-stakes settings, interpretability for time-series data has received comparatively less attention in the XAI literature, particularly relative to the substantial body of work targeting

images and tabular data (Theissler et al. 2022). Motivated by this, the present survey focuses specifically on interpretability for time-series data.

While considerable progress has been made towards improving the interpretability of ML systems, existing literature often groups a wide range of approaches under the broad concept of explainability, despite substantial differences in how and when explanations are generated, the types of models they target and the data modalities they address (Theissler et al. 2022; Rojat et al. 2021; Jovic et al. 2025; Arsenault et al. 2025). This has resulted in a fragmented landscape where different types of interpretability approaches, model properties, and explanation objectives are not always clearly distinguished (Gilpin et al. 2018; Graziani et al. 2023). As a result, inherently interpretable models are often considered alongside fundamentally different post-hoc methods, making their unique characteristics and contributions less clear. The contributions of these existing surveys are valuable, and a detailed discussion of them can be found in Section 2; however, they remain limited in their treatment of interpretability-by-design approaches. There remains a need for a more systematic understanding of approaches that incorporate interpretability directly into the model design, rather than relying on generated post-hoc explanations. This work presents inherently interpretable models for time-series data, classified according to their structural principles and reasoning mechanisms, thereby addressing the gap by reviewing models that incorporate interpretability directly into the design rather than introducing it through post-hoc explanations.

We began by defining the terms related to XAI to establish a clear conceptual foundation for the discussions that follow. The key contributions of this work are summarised below:

1. **Systematic literature review:** Following Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology, this study identified

and reviewed 77 relevant studies applying inherently interpretable [ML](#) models to time-series data.

2. **Conceptual foundations of interpretability in time-series:** We clarify key concepts related to interpretability, the characteristics of time-series data and discuss the challenges associated with achieving interpretability in sequential domains.
3. **Taxonomy of inherently interpretable models:** We provide a structured classification of inherently interpretable models for time-series based on their underlying architectures and mechanisms, including linear models, tree-based models, interpretable neural architectures, attention-based models, case- and rule-based models, fuzzy, and neuro-fuzzy models.
4. **Dataset analysis:** We analyse the datasets used across the reviewed studies, categorising them based on time-series modality, application domain, and learning task. This provides an overview of the data sources commonly used in interpretable time-series [ML](#) research and highlights trends in their adoption.
5. **Identification of challenges and future directions:** We analyse existing research gaps and discuss open challenges and potential future research directions for developing more transparent and reliable interpretable time-series models.

The remainder of the paper is organised as follows: Section [2](#) reviews existing research relevant to our area of study, outlining key findings, challenges, and how our work addresses these challenges. Section [3](#) presents our methodology, detailing our search process and the selection criteria for the included models. Section [4](#) discusses time-series data and explores what interpretability means in the context of time-series models. Section [5](#) presents the taxonomy of the reviewed models in this survey, classified by their underlying architecture. Each subsection is prefaced with relevant foundational concepts from the model architectural family to facilitate understanding of the presented models. Then Section [6](#) presents an overview of data sets used in

the survey studies, providing a resource for commonly used data sets in the domain, categorised by application field and task type, with download links where applicable. Building on this, Section 7 provides a robust discussion of our findings, examining emerging trends and bibliometric analyses across the reviewed studies. Finally, Section 8 concludes the paper by summarising our key contributions, discussing limitations, and reflecting on future directions.

2 Existing Studies Review

The growing interest in XAI has led to the publication of numerous surveys, systematic reviews, and literature studies that examine the development of explainable and interpretable ML approaches. These studies generally fall into two categories: broad surveys that review XAI approaches across different data types, application domains, and model families (Vilone and Longo 2020; Minh et al. 2022; Arrieta et al. 2020; Angelov et al. 2021), and more specialised surveys that focus on a particular data modality, application area, or class of explainable approaches (Theissler et al. 2022; Rojat et al. 2021; Arsenault et al. 2025; Jovic et al. 2025). Several recent surveys have also explored XAI in the context of time-series data (Theissler et al. 2022; Rojat et al. 2021; Arsenault et al. 2025; Jovic et al. 2025). However, they differ considerably in their scope, objectives, and categorisation strategies, and none focus specifically on inherently interpretable models for time-series analysis in the manner considered in this work.

Theissler et al. (2022) provided one of the more thorough overviews of explainability specifically for time-series classification tasks. It collects and categorises XAI approaches, the majority of which are post-hoc explanation methods, with a handful of inherently interpretable models also included, though this distinction is not carried into the taxonomy itself. Their developed taxonomy instead groups methods into time-point-based, subsequence-based, and instance-based criteria and highlights open

challenges, especially in evaluation and user-centric explanation design. The authors noted that much of the existing literature is model-specific and lacks unified evaluation frameworks, and that many approaches adapted from computer vision have not been fully validated in temporal domains.

The earlier survey by [Rojat et al. \(2021\)](#) similarly reviewed explainability applied to time-series across several tasks (classification, forecasting, clustering). It considered both post-hoc explanation methods and inherently interpretable models, with a strong emphasis on explaining [Deep Learning \(DL\)](#) models that are otherwise opaque. This work also outlined the variety of explanation types (including attribution and attention methods) and reflected on their role in building trust and confidence in models. It highlights that interpretability has been less explored for temporal data than in vision and language domains, and underscores the need for better evaluation strategies for explanations.

[Arsenault et al. \(2025\)](#) focused specifically on forecasting in the financial domain. It distinguishes inherently interpretable models from post-hoc explanation methods and presents a taxonomy of methods used to make financial forecasting models more transparent for high-stakes decision-making. While this review offers a structured overview, its domain restriction to financial time-series means many broader methodological insights about model architecture and inherently interpretable designs are not addressed.

Recently, [Jovic et al. \(2025\)](#) presents another perspective: it surveys [ML](#) methods applied to biomedical time-series and assesses the balance between interpretability and predictive accuracy. It notes that many accurate models (especially [DL](#) architectures) remain uninterpretable, and that inherently interpretable methods are surprisingly scarce in the biomedical literature. Common interpretable models identified include decision trees and k-nearest neighbours, whereas deep architectures dominate accuracy but lack built-in interpretability.

[Park et al. \(2026\)](#) present a comprehensive cross-task analysis of explainable time-series methodologies, with a primary focus on decomposition and forecasting tasks. Their central contribution is a set of seven "Priority Properties" that distinguish different methods from one another to facilitate and guide model selection. They apply this property analysis to their established model taxonomy, which divides methods into four main categories: Traditional Statistical Methods, Probabilistic Process Methods, Frequency-based Decomposition Methods and Deep Learning-based Approaches. They also dedicate a section to evaluating post-hoc explanation methods and the different levels of transparency provided by these methods. However, their treatment of inherently interpretable models is largely confined to traditional statistical and probabilistic approaches; their first four categories as listed earlier, with other well-established families of inherently interpretable models, such as case-based and tree-based methods, left unaddressed.

Despite the value of these reviews, several limitations remain when contrasted with our study. In all of the aforementioned works, the bulk of the models reviewed use post-hoc explanation methods, and inherently interpretable models remain in the minority. The taxonomies used to classify models also fail to highlight the unique characteristics of inherently interpretable models, as they are not classified separately on that basis, whereas our work focuses strictly on inherently interpretable models. Existing taxonomies tend to classify how to interpret opaque models after training, but they do not systematically organize interpretable models by architecture or model family, as we present the models included in this work in Section 5. Furthermore, domain-specific reviews (finance, biomedical), as in [Arsenault et al. \(2025\)](#) and [Jovic et al. \(2025\)](#), emphasize the application context rather than model design principles that ensure interpretability by construction. Our work fills this gap by narrowing the scope to architectures that are interpretable by design, comparing them on structural principles.

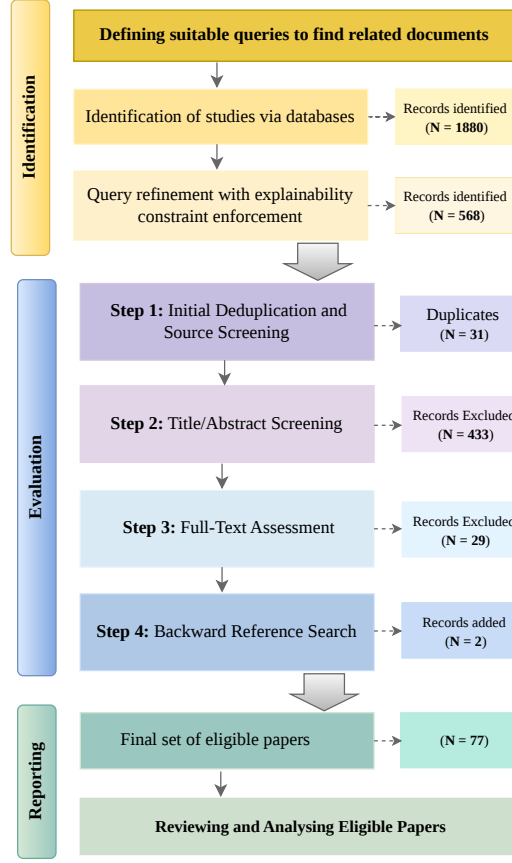


Fig. 2: Method employed in the study based on the PRISMA statement

3 Research Methodology

This section outlines the methodology employed in this systematic review to identify and analyse eligible studies focusing on inherently interpretable [ML](#) models applied to time-series. To ensure rigour, transparency, and reproducibility, we adopted the [PRISMA](#) framework ([Page et al. 2021](#); [Moher et al. 2009](#)). The process is organised into three consecutive stages: Identification, Evaluation, and Reporting. [Fig. 2](#) provides a visual schematic of this workflow, detailing the number of records identified, screened, assessed, and included at each stage.

3.1 Identification Phase

The initial phase was dedicated to constructing a comprehensive and precise search strategy to capture all relevant literature at the intersection of three core conceptual domains:

1. **Explainability and Interpretability:** Terms related to the core concepts of model transparency.
2. **Inherently Interpretable Model Families:** Specific names of model classes known for their inherent interpretability.
3. **Time-Series and Signal Data:** Application domains and data types of interest.

We developed a multi-part Boolean search query executed across the Title, Abstract, and Keywords in the Web of Science core database. Fig. 3 shows the Boolean query strings employed to identify studies and filter relevant studies for inclusion in the review. The search was limited to publications from January 2010 to the beginning of May 2026. An initial execution yielded 1,880 candidate records. To ensure strict relevance to our core research question, specifically focusing on inherently explainable models for temporal data, we performed a critical refinement of the query logic. This involved tightening proximity operators and adjusting term groupings to prioritise papers where explainability was a central contribution, not a peripheral mention. This precise calibration significantly sharpened the focus, refining the initial result set to a more targeted pool of 568 publications for the subsequent screening stages.

3.2 Evaluation Stage

This phase involved a rigorous, multi-step screening process to filter the identified records and select studies that met the pre-defined eligibility criteria (C1–C5). The pre-defined eligibility criteria are as follows:

- **C1:** Published between January 2010 and May 2026.

```

("explainable" OR "Interpretable" OR "explainability" OR "interpretability" OR
"transparency" OR "transparent") AND (("Interpretable" OR "Explainable" OR "self-
explainable" OR "self-explaining" OR "inherently explainable" OR "inherently
interpretable" OR "intrinsically interpretable" OR "transparent model" OR "Gaussian
Process Regression" OR ("bayesian network" NEAR/5 (explain* OR interpret*)) OR
("fuzzy" NEAR/5 (explain* OR interpret*)) OR ("decision tree" NEAR/5 (explain* OR
interpret*)) OR ("logistic regression" NEAR/5 (explain* OR interpret*)) OR ("attention
mechanism" NEAR/3 (explain* OR interpret*)) OR ("attention-based" NEAR/3 (explain*
OR interpret*)) OR ("attention" NEAR/3 ("explainability" OR "interpretability")) OR
"nearest neighbors" OR "rule-based" OR "linear model" OR "ridge regression" OR "lasso
regression" OR "explainable boosting machine" OR "sparse linear model" OR "elastic net"
OR "Generalized Linear model" OR "Generalised Linear model" OR "generalised additive
model" OR "decision rule" OR "scoring model" OR "scorecard" OR "score card" OR
"neural additive models" OR "case-based reasoning" OR "prototype-based" OR "concept-
based" OR "counterfactual explanation")
AND
("machine learning" OR "deep learning" OR "artificial intelligence" OR "neural
network")
AND
("time series" OR "sequential data" OR "temporal data" OR "longitudinal data" OR
"signal data" OR "sensor data" OR "streaming data" OR "EEG" OR
"electroencephalogram" OR "ECG" OR "electrocardiogram" OR "EMG" OR
"electromyogram" OR "PPG" OR "photoplethysmogram" OR "heart rate" OR "heart rate
variability" OR "physiological signal" OR "temperature" OR "rainfall" OR "seismic data"
OR "climate data" OR "stock prices" OR "financial data" OR "energy consumption" OR
"IoT data" OR "traffic flow"))

```

Fig. 3: Boolean query strings for identifying studies

- **C2:** Written in the English language.
- **C3:** Focuses on the development, application, or evaluation of one or more inherently interpretable [ML](#) models.
- **C4:** Explicitly applies the model(s) to time-series, sequential, or continuous signal data.
- **C5:** Published as a primary research article in a reputable, peer-reviewed journal.

These criteria were applied through a three-stage screening process:

- **Step 1: Initial Deduplication and Source Screening** - The 568 records were checked, and duplicate entries were automatically and manually removed. Subsequently, we excluded articles from non-academic sources, low-quality or disreputable

publications, non-English studies, review papers, and preprints. This left us with 537 papers eligible for further review.

- **Step 2: Title/Abstract Screening** - The titles and abstracts of the remaining records were reviewed to assess their relevance to the study’s objectives. Papers were excluded if they did not primarily involve: 1) an inherently interpretable ML model (post-hoc explanation methods were not included), 2) applications involving time-series or continuous signal data, and 3) a focus on model interpretability as a key research outcome. This process significantly narrowed down the pool of potential candidates. The number of papers remaining after this step was 104.
- **Step 3 – Full-Text Assessment and Inclusion Criteria** - The full texts of the remaining papers were examined to verify that they satisfied all eligibility criteria (C1–C5). Studies failing to meet any criterion were excluded from the final review.

Following this three-step evaluation, 77 studies were selected for final inclusion.

3.3 Reporting Stage

The included studies were categorised based on the computational models they developed. For each study, we systematically analysed key characteristics including the publication year, the learning paradigm and architecture used, the datasets, the specific tasks and applications, and the performance metrics evaluated.

4 Interpretability in Time-Series ML: Concepts and Foundations

Time-series data consist of ordered sequences of observations indexed by time, where each observation may be univariate or multivariate. Formally, a time-series can be represented as shown in Eq. 1.

$$\mathbf{X} = \{x_t\}_{t=1}^T \tag{1}$$

where $x_t \in \mathbb{R}^d$ denotes the observation at time step t , and d is the number of variables. Unlike static data, time-series are characterised by temporal dependencies, autocorrelation, nonstationarity, trends, seasonality, and other irregularities (Box et al. 2015).

Time-series models aim to capture these temporal dependencies to perform tasks such as forecasting, classification, anomaly detection, or segmentation. Classical approaches include linear autoregressive models, state-space models, and probabilistic graphical models. In contrast, more recent approaches rely on ML and DL architectures that learn temporal representations directly from data, such as recurrent neural networks and transformer-based models. These architectures have become increasingly prominent due to their ability to capture long- and short-term temporal patterns (Hall and Rasheed 2025).

However, these models are very hard to interpret because their internal mechanisms, representations, and decision boundaries are not easily mapped to human-explainable terms. This tension between accuracy and interpretability has been identified as a key barrier to deployment in safety-critical or high-stakes temporal domains such as healthcare, autonomous systems, and finance (Theissler et al. 2022).

Interpretability in the context of time-series modelling extends beyond understanding the relationship between input variables and outputs; it also requires understanding how temporal structure contributes to a model’s decision. This includes identifying which time steps, temporal patterns, or state transitions influence predictions and how these influences evolve over time. Consequently, interpretability for time-series models is inherently multi-dimensional, involving feature-level, temporal, and sometimes structural explanations (Molnar 2020).

In inherently interpretable time-series models, we can directly see how the model uses information from past time steps and the dependencies between the time steps

in some cases to make predictions (Le Nguyen et al. 2019). For example, linear temporal models quantify the effects of lagged values through coefficients, rule-based models encode temporal conditions explicitly, decision trees split over time-dependent features, and probabilistic models capture dependencies that reflect causal or conditional relationships across time steps (Shalaeva et al. 2018; Mueen et al. 2011; Mrowca et al. 2022). In such models, temporal dynamics are not implicitly embedded in high-dimensional latent spaces with abstract decision boundaries. They are instead expressed through model components that can be directly inspected and reasoned about by humans. In some cases, time-series are first summarised into meaningful features, which are then used as inputs to inherently interpretable models that are not explicitly temporal (Lin et al. 2007; Senin and Malinchik 2013).

As this work focuses exclusively on inherently interpretable models for time-series, referring to models that are transparent by design as mentioned in Section 1. Post-hoc explanation methods are not considered further; instead, the models reviewed in this work are classified according to their structural principles and reasoning mechanisms, forming the basis of the architectural taxonomy introduced in Section 5, which organises inherently interpretable temporal models into linear, rule-based, tree-based, deep neural network-based, and fuzzy systems.

5 Inherently Interpretable Models for Time-Series

This section presents the inherently interpretable models selected from among the reviewed studies, organised according to a taxonomy that groups models by architectural family; namely linear, case- and rule-based, tree-based, neural network-based and fuzzy systems. Each family is first introduced through its foundational concepts, to support readability and understanding. Subsequently, the selected studies within that family are further sub-grouped according to shared characteristics.

5.1 Linear Models

Linear models represent the relationship between a target variable and a set of predictive features as a combination of additive factors. This structure can be modified in various ways, from the addition of regularisation terms to the transformation of individual factors into functions or other structural modifications, thereby giving rise to a wide variety of formulations (Hastie 2009; James 2013). This section introduces the foundational concepts underlying the linear models associated with the reviewed studies, followed by the reviewed studies, which are classified into two groups: studies employing classical linear models, and studies employing more complex architectures.

5.1.1 Foundations

Linear Regression: commonly referred to as **Ordinary Least Squares (OLS)**, it is a foundational instance of linear models described above. It estimates its coefficients by minimising the sum of squared differences between the predicted and observed outcomes (Montgomery et al. 2022), and is formally expressed as shown in Eq. 2. Its interpretability stems directly from this additive structure, as the effect of each feature on the predicted outcome can be seen by directly inspecting the coefficient associated with the feature (Molnar 2020). As the predicted output is unbounded, **OLS** is suited to continuous prediction tasks, but it cannot be directly applied without modification to classification problems where the outcome is categorical (Hosmer Jr et al. 2013).

$$\hat{y} = \beta_0 + \sum_{i=1}^p \beta_i x_i \quad (2)$$

Logistic Regression (LR): It is an extension of **OLS** in which a logistic function maps linear combinations of input features to class probabilities; thereby confining the outputs of a linear equation to values between 0 and 1 (Hosmer Jr et al. 2013). While **OLS** predicts an output according to Eq. 2, **LR** applies the logistic function to this linear predictor, yielding Eq. 3.

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0 + \sum_{i=1}^p \beta_i x_i)}}. \quad (3)$$

The inherent interpretability of an [LR](#) model is tied to the examination of the model weights, which are the outcomes of the model and how much each feature contributes to or affects the outcome. However, due to its use of the logistic function, the weights are not additive, and their individual effects cannot be examined in a linear fashion ([LaValley 2008](#); [Molnar 2020](#)). The odds ratio is used instead, where, by rearranging the logistic equation to isolate the linear term, we find that each weight corresponds to an additive change in the log-odds. Exponentiating the weight for each feature then gives the odds ratio, which tells us how a one-unit increase in that feature multiplies the odds of predicting a given class, while holding other features fixed ([Molnar 2020](#)).

Generalised Linear Models (GLMs): [GLMs](#) extend [OLS](#) by allowing the response variable to follow different distributions from the exponential family (e.g., Gaussian, Bernoulli, Poisson) ([McCullagh 2019](#)). A [GLM](#) consists of three components: a linear predictor, a link function, and a probability distribution. The link function connects the expected value of the response to the linear combination of input features, enabling modelling of non-Gaussian outcomes while retaining a linear structure in the parameters. The relationship between the expected response $\mu = \mathbb{E}[y|\mathbf{x}]$ and the linear predictor is expressed through a link function $g(\cdot)$ as presented in [Eq. 4](#).

$$g(\mu) = \beta_0 + \sum_{i=1}^p \beta_i x_i. \quad (4)$$

Linear and [LR](#) are thus both special cases of the [GLM](#) framework, recovered by choosing the identity and logit link functions, respectively. Interpretability in [GLMs](#) is inherited from this linear structure, as model coefficients quantify the contribution of each feature in the transformed (link) space, making them straightforward to analyse

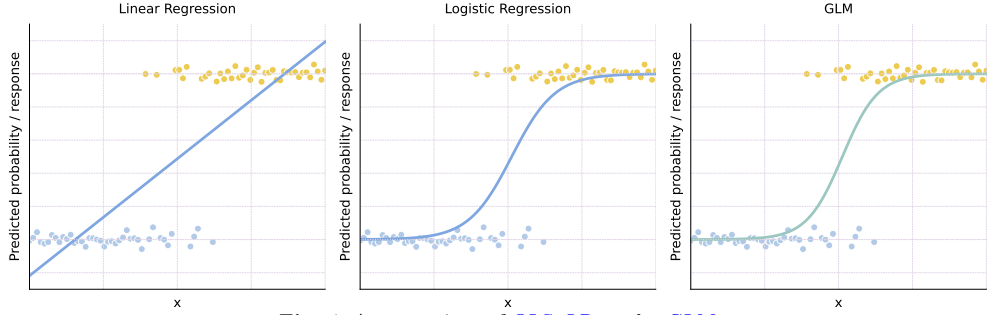


Fig. 4: A comparison of OLS, LR, and a GLM

under the chosen link function (McCullagh 2019; Molnar 2020). An illustration of how OLS, LR, and GLM fit to a sample function can be seen in Fig. 4.

Generalised Additive Models (GAMs): GAMs extend GLMs by relaxing the linearity assumption between features and the response. Instead of a strictly linear combination, GAMs model the prediction as a sum of learned univariate functions, one per feature (Hastie and Tibshirani 1986; Hastie 2017; Larsen 2015). This allows the model to capture nonlinear relationships while preserving an additive structure. Formally, a GAM models the expected response through a link function $g(\cdot)$ applied to a sum of smooth functions $f_i(\cdot)$, one for each feature, as seen in Eq. 5.

$$g(\mu) = \beta_0 + \sum_{i=1}^p f_i(x_i). \quad (5)$$

When g is the identity and each f_i is constrained to be linear, the GAM reduces to a standard GLM. Interpretability is maintained because each feature contributes independently through its learned function, which can be visualised as a shape function showing how different values of the feature influence the prediction.

Explainable Boosting Machines (EBMs): EBMs are a variant of GAMs in which the feature-wise functions are learned using gradient boosting over shallow decision trees instead of predefined smooth functions as in GAMs (Lou et al. 2013; Nori et al. 2019). Each feature is modelled independently through an additive shape function constructed iteratively, with optional pairwise interaction terms included for selected

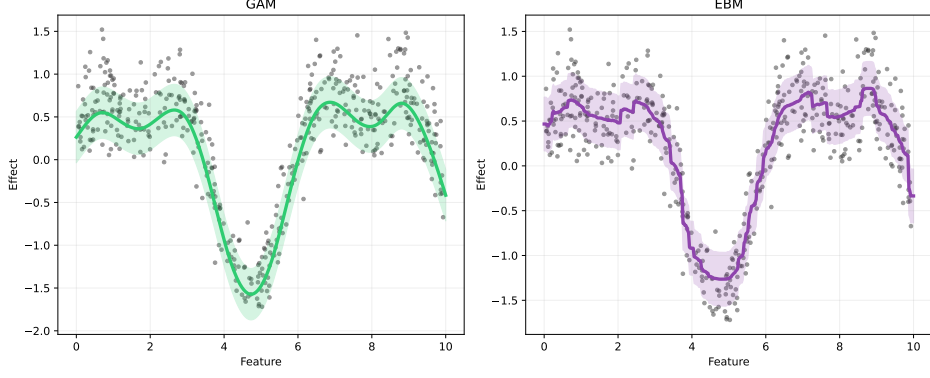


Fig. 5: GAM vs. EBM

feature pairs. Extending the [GAM](#) formulation, [EBMs](#) augment the additive structure with optional pairwise interaction terms, as presented in Eq. 6.

$$g(\mu) = \beta_0 + \sum_{i=1}^p f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) \quad (6)$$

where $f_{ij}(x_i, x_j)$ captures interactions between selected feature pairs, and the model reduces to a standard [GAM](#) when these terms are omitted. Fig. 5 illustrates how the boosted shape functions learned by an [EBM](#) compare to the smooth spline fits of a [GAM](#) across features.

Interpretability follows directly from the model’s additive structure, with each feature’s contribution being fully separable, making the model inherently transparent. Each feature’s effect can be inspected via its learned shape function and importance score. Global importance can be quantified from feature contributions, and local explanations can be obtained by summing per-feature contributions for individual predictions ([Nori et al. 2019](#)).

5.1.2 Selected Studies using Classical and Regularized Linear Models

To model mosquito population size using satellite-derived environmental data, [Mudele et al. \(2021\)](#) presented a weighted Poisson GLM. By doing so, they identified the environmental drivers of dengue risk. The architecture starts by estimating environmental variables from satellite data, followed by an initial Poisson GLM. Feature selection is then performed using Akaike Information Criterion (AIC) to choose the most informative of the environmental variables and reduce model complexity. They then trained a weighted GLM based on the features selected by AIC, with the observation weights obtained from residual errors. The weighted GLM performed comparably to ML models while providing interpretability, as the GLM coefficients can be examined to show which environmental factors affect mosquito population, enabling sensitivity analysis to understand the climatic drivers of dengue risk.

A key contribution towards aero engine exhaust gas temperature (EGT) prediction was presented by [Apostolidis et al. \(2022\)](#). They selected features based on expert knowledge, and correlation analysis with the target variable indicates that the selected variables are highly correlated with it. They implemented and tested a GAM for the prediction of an aero engine's EGT on three synthetic state-of-the-art datasets with increasing complexity. The authors achieved accurate performance and drew conclusions about the relevant and impactful features for successful EGT prediction, what sensor readings to consider, taking a step towards the actualisation and development of models for implementation and deployment on real aircraft.

EBM models are developed and used to classify [electrocardiography \(ECG\)](#) signals as normal or abnormal, and to estimate landslide probability during extreme rainfall events by [Xiaolin et al. \(2022\)](#) and [Caleca et al. \(2024\)](#), respectively. [Xiaolin et al. \(2022\)](#)'s model is a lightweight version, as they use a set of handcrafted ECG features, including peak amplitudes, time intervals, and a few morphology-related amplitudes

to avoid heavy signal processing and keep computation low. The signals classified as abnormal are then sent to the cloud for further computation.

Similarly, in the landslide probability estimation model, [Caleca et al. \(2024\)](#) handled the input time-series data; rainfall, by summarising it into cumulative 24-h rainfall of each landslide event, mean annual rainfall for each station, and rainfall anomaly. They calculated rainfall anomaly by dividing the respective cumulative 24-h precipitation and mean annual rainfall. They employed further variable selection with stepwise forward selection. In both cases, the [EBM](#) builds predictions by summing contributions from individual features, with each feature having its own learned score function, and the final decision is the total score. Since [EBMs](#) are additive models, each variable's global and local (per prediction) contribution can be inspected separately.

[Zheng et al. \(2024\)](#) and [She et al. \(2025\)](#) used [LR](#) as a base model to predict thermal sensation votes (cool/neutral/warm) for elderly people, and to estimate [Intensive Care Unit \(ICU\)](#) mortality in middle-aged and elderly sepsis patients, respectively. They both performed variable selection before training the [LR](#) model. [Zheng et al. \(2024\)](#) used Spearman correlation ranking to select the input features. The problem also involves an imbalanced dataset, which is addressed using SMOTE (Synthetic Minority Oversampling Technique). SMOTE generates synthetic data points for the under-represented class by interpolating between existing minority samples before training ([Chawla et al. 2002](#)). They demonstrated interpretability by examining the odds ratio and probability curves of each feature. [She et al. \(2025\)](#) performed variable selection by first removing variables with high multicollinearity from univariate analysis, and then using random forest importance ranking and LASSO regression. They visualised the [LR](#) model as a nomogram, and the prediction for each observation or case could be seen, interpreted, and understood visually from the nomogram.

5.1.3 Selected Studies using Advanced Linear Models

Jiang et al. (2019) proposed an adapted transfer learning method with a modified regression model as a classifier called γ -Least Squares Regression (LSR) to form the $TLSR_K^L$ framework (Transfer Learning with Knowledge and Label Space) for multi-class electroencephalography (EEG) signal recognition. The model learns from a source dataset, then transfers the learned model weights (knowledge) and a generalised label representation (label space) to a target dataset where labeled data is scarce. The γ -LSR model modifies LSR by modifying the label representation using a label shift vector γ , which expands the distance between class targets in the label space. During training, the model fits the regression model to target-domain data and regularises the weight matrix. Encouraging the target model weights to stay close to the source-domain weights, and the target label shift vector to stay close to the source label shift vector. Effectively combining transfer learning and γ -LSR. Because the model retains the linearity of the original LSR method, each feature’s contribution can be analysed directly through its coefficient in the mapping matrix, making the model inherently interpretable. The authors’ experiments show that $TLSR_K^L$ outperforms other transfer learning methods.

In the same vein, Wijesena and Pradhan (2025) introduced a surrogate model for building a transparent multi-weather insurance index to predict crop yield losses. They first trained a complex model; a neural network, created an expanded dataset by augmenting each observation with 10 random neighboring samples to cover the parameter space more, generated predictions with the complex model on the expanded dataset, and then trained a simpler "surrogate" model, a GLM with fewer number of features to learn and approximate the complex model’s predictions. The final GLM equation has explicit coefficients in which the contribution of each feature can be interpreted easily.

The proposed model by Munkhdalai et al. (2022) is a time-series forecasting model that is initialised as a simple OLS model. Then the regression coefficients are updated

with an [Recurrent Neural Network \(RNN\)](#) at each time step. This effectively rebuilds the regression equation repeatedly, producing a new set of coefficients for every time step, making the linear model locally adaptive, with the influence of each variable changing dynamically over time, instead of having fixed global coefficients while still being able to inspect and interpret the effect of each feature directly.

Table 1: An Overview of Interpretable Linear Models for Time-Series

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Jiang et al. (2019)	$TLSR_K^L$	Transfer Learning, Linear Model	The Bonn EEG Dataset, The TUH EEG Corpus (TUEG)	EEG Signal Multi-class Classification	Epileptic Seizure Recognition	Average Accuracy on all Datasets, Computational Time
Mudele et al. (2021)	GLM for Dengue	Weighted Poisson GLM	Earth observation data from the municipality of Vila Velha and mosquito population data from the same region	Regression, Mosquito Population modelling	Dengue Risk Modelling, Public Health Vector Control Strategies	MAE
Munkhdalai et al. (2022)	RNN-Augmented OLS	OLS , RNN	Electricity, Traffic (PEMS-SF), Appliances Energy Prediction Data Set, Air Quality, Financial Datasets	Regression, Forecasting	Time-series Forecasting	RMSE , MAE , MAPE

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Apostolidis et al. (2022)	GAM for EGT prediction	GAM	N-CMAPSS	Regression for EGT Prediction	Condition-Based Maintenance and Predictive Maintenance for Aircrafts and Safety-Critical Applications	RMSE , R^2
Xiaolin et al. (2022)	EBM for ECG classification	EBM	MIT-BIH Arrhythmia	Binary Classification	ECG classification of normal and abnormal signals	Accuracy, Sensitivity, Specificity, F1 score, Percentage of Data sent to the Cloud.
Zheng et al. (2024)	TSV Prediction	LR	Elderly People Vitals and TSV	Classification	Indoor Thermal Environment Modelling	Accuracy, Precision, F1 Score, AUROC
Caleca et al. (2024)	EBM for Landslide Prediction	Explainable Boosting Machine	Slope unit data on rainfall-triggered landslide	Classification	Landslide Susceptibility modelling	AUROC
Wijesena and Pradhan (2025)	GLM Surrogate Model	GLM , Surrogate Model, DL	Weather Index Datasets	Regression Weather Index Prediction	Crop Yield Loss Prediction	RMSE , MAE , Tail MAE
She et al. (2025)	Nomogram based on LR	Random Forest Importance Ranking, LASSO regression, LR , Nomogram	eICU Collaborative Research Database	Regression for Mortality Risk Prediction	Sepsis Patients Monitoring, Mortality Risk Modelling, Healthcare	AUROC , Sensitivity, Specificity, PPV , NPV , Calibration Curves, DCA

5.2 Tree-Based Models

This section presents tree-based models that reason through a hierarchical sequence of splits, arriving at a prediction by tracing an input through a series of conditions rather than through coefficients or features. As in the previous sections, the selected studies are preceded by a brief theoretical description of tree-based reasoning and its measures of interpretability. It is worth noting that the studies included in this survey are restricted to models where interpretability is a primary concern and has not been sacrificed in favour of model accuracy and complexity.

5.2.1 Foundations

Decision Trees: Decision trees perform prediction by partitioning the input space into regions by recursively splitting on features. Each split is chosen to reduce variance in the target variable within the resulting regions. In regression trees, each final region predicts a constant value, typically the mean of the training samples in that region, producing a piecewise constant approximation of the underlying function. In classification trees, the split criterion is instead Gini impurity or entropy, measuring how mixed the classes are in each resulting region, and the final region predicts the majority class (Quinlan 1986). This structure is inherently interpretable, as the prediction path for each data point can be decomposed and traced from the leaf node to the root by following the paths and the subsets to which the data point belongs. Feature importance is another measure of interpretability in tree-based models, showing how much influence each feature has on the model’s decision-making. It can be determined by analysing all the splits in a tree in which that feature was used and how much it reduced the variance or Gini impurity or entropy (Molnar 2020). The same measures of interpretability apply to all tree-based models; however, that interpretability degrades as model complexity increases.

Random Forests (RFs): RFs are ensemble extensions of decision trees. Multiple trees are trained on bootstrapped samples of data, and at each split, only a random subset of features is considered. The final prediction is obtained by averaging the outputs of all trees. This reduces variance and improves generalization compared to a single tree. The method relies on the idea that averaging many weakly correlated, high-variance models leads to a more stable and accurate predictor. However, increasing the number of trees in such an ensemble method yields a more complex model with reduced interpretability, as the decision process is distributed across many trees (Breiman 2001).

Cubist Tree-Model: A Cubist Tree-model combines decision trees and OLS. Similar to model trees, the input space is first partitioned using decision rules, after which an OLS model is fitted within each resulting region. This produces a piecewise linear model, where the tree structure captures nonlinear relationships and the linear models capture local feature effects. Interpretability is provided through the decision rules derived from the tree, which can be examined and simplified to explain the model's predictions. An optional k-nearest neighbors refinement can further improve prediction

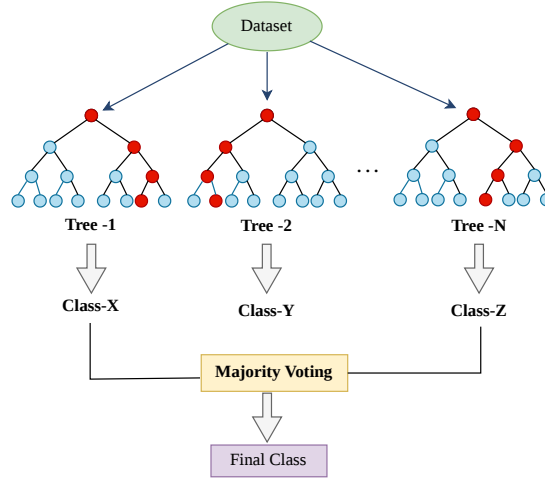


Fig. 6: Random Forest

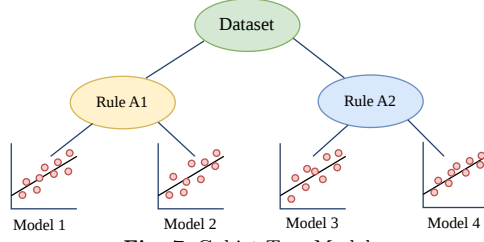


Fig. 7: Cubist Tree Model

accuracy without altering the underlying model structure (Quinlan 1993; Kuhn et al. 2013).

5.2.2 Selected Studies Using Tree-Based Models

For the prediction of European eel landings in Ichkeul Lake, Sahbani et al. (2024a) evaluated the performance of RF and Cubist-Tree models. They worked from 13 environmental variables (water level, salinity, turbidity, chlorophyll a, nutrients, dissolved oxygen, and others) monitored over 2010–2020. Both models are tuned with 5-fold cross-validation. Interpretability is demonstrated through variable importance scores from RF and sensitivity analysis from Cubist Tree, both identifying water level, salinity, and turbidity as the dominant drivers of eel landing; consistent with Pearson correlation results.

Sahbani et al. (2024b) extended this directly, using the same RF and Cubist Tree models and the same lake to forecast eel landings under hypothetical future climate change and anthropogenic pressure scenarios. The environmental variables are projected forward and fed into the trained models to generate a probability distribution of future landings under each scenario. Interpretability is demonstrated in the same way; through variable importance and sensitivity analysis. These insights are now used to understand which pressures are most critical to future eel population trajectories. The authors additionally propose combining both models as a more robust decision-making tool for lake management.

For real-time machine fault diagnosis, [Lu et al. \(2026\)](#) proposed a decision-tree-inspired analogue circuit classifier that works directly from raw vibration signals, bypassing analogue-to-digital conversion entirely. The decision tree logic is physically implemented in hardware, with a filter matrix playing the role of internal nodes; splitting on frequency content, and amplitude quantisation modules select branches. And the final circuit outputs correspond to leaf node class labels. Interpretability is inherited directly from the decision tree structure embedded in the circuit, as the signal’s path through the filter-branch hierarchy is physically traceable.

With a slightly different approach, [Aghahadi et al. \(2026\)](#) developed a four-layer [ML](#) framework for classifying and reconstructing electrical load profiles in smart distribution networks, with separate pipelines for domestic and non-domestic users. The four layers are sequential: clustering-based pattern recognition groups users by consumption shape; [RF](#) classification assigns new users to the appropriate cluster/profile type; [RF](#) regression estimates consumption magnitudes; and a temperature dependency layer applies engineered thermal features to correct for seasonal variations. Interpretability comes from the modular architecture itself, with each layer having a dedicated, inspectable function; cluster representatives are human-readable consumption archetypes, and the bidirectional z-score normalisation provides a transparent mapping between normalised pattern space and actual physical power values. [RF](#) feature importance additionally reveals which engineered features and contextual variables drive the classification and regression steps.

Table 2: An Overview of Interpretable Tree-Based Models for Time-Series

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Sahbani et al. (2024a)	Tree-based models for eel landing	RF and Cubist-Tree	Eel production data and Ichkeul Lake data	Regression, Eel landing prediction	Environmental Modelling	RMSE , MAE , R^2
Sahbani et al. (2024b)	Tree-based models for eel landing	RF and Cubist-Tree	Eel production data and Ichkeul Lake data	Regression, Eel landing prediction	Environmental Modelling	RMSE , MAE , R^2
Lu et al. (2026)	Decision-tree-inspired analog circuit classifier	Decision tree logic implemented in hardware	CWRU Bearing, BLDCM, SRM Dataset	Classification, Predictive Real-time machine fault diagnosis	Maintenance	Accuracy, Sensitivity, Recall, Error Rate
Aghahadi et al. (2026)	ML-based framework for electrical load profiling	Clustering-based pattern recognition + Random Forest Regression and Classification + Temperature Dependency Layers	Utility data from three medium-voltage feeders	Classification and reconstruction of electrical load profiles	Energy Consumption Forecasting	RMAE , MAE , RMSE , R^2

5.3 Interpretable Neural Networks

It is worth noting that many of the models presented in this section are only partially explainable because, despite the authors going to great lengths to develop explainable components in the models, they still include neural network elements whose outputs cannot be fully explained.

5.3.1 Foundations

Attention Mechanism and Attention-Based Explainable Model Construction: Attention is a DL mechanism that allows a model to dynamically focus on the most relevant parts of its input when making a prediction. Instead of treating all features or time steps as equally important, the model learns a set of weights, referred to as attention scores, that quantify the relative importance of the different parts of the input (Vaswani et al. 2017). In time-series and multimodal settings, attention is commonly applied over temporal dimensions, feature dimensions, or structured inputs (e.g., channels, leads, or frequency components) (Barić et al. 2021) as illustrated in Fig. 8. Beyond improving predictive performance, attention also introduces a degree of interpretability. In that, the learned attention weights can be inspected to understand which inputs the model relied on most strongly, offering insight into the underlying decision-making process that led to the model’s final prediction (Tutek and Šnajder 2022).

Sparse Attention Mechanisms: Standard attention typically uses a softmax function, producing dense weight distributions. Sparse alternatives such as *entmax* enforce that only a subset of inputs receive non-zero weights, effectively performing implicit feature or time-step selection (Peters et al. 2019; Correia et al. 2019). This improves interpretability by isolating the most relevant contributors and reducing noise in attention maps (Hastie et al. 2015; Martins and Astudillo 2016).

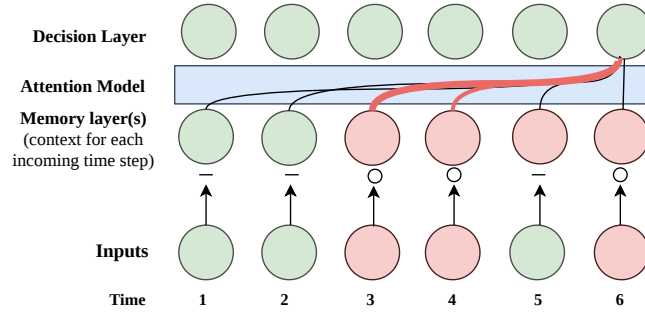


Fig. 8: A simplified schematic illustration of the attention mechanism in time-series data

Interpretable Convolutional Neural Networks: Several models design convolutional architectures to be inherently interpretable by constraining how features are learned. Instead of standard convolutional filters, interpretable [Convolutional Neural Networks \(CNNs\)](#) often parameterize filters in a structured way (e.g., by using band-pass filters with learnable cut-off frequencies, or with each filter representing a certain object part) ([Zhang et al. 2018](#); [Ravanelli and Bengio 2018](#)). This allows direct correspondence between learned features and domain-specific representations such as frequency bands. Depthwise or separable convolutions are also used to disentangle spectral and spatial feature learning, enabling attribution of importance to specific channels, frequency bands, or spatial regions ([Chollet 2017](#); [Ravanelli and Bengio 2018](#)). These models provide explanations in terms of domain-specific meaningful features (e.g., frequency bands, spatial regions, sensor channels), rather than abstract latent representations.

Graph Neural Networks (GNNs): This method extends neural architectures to data with non-Euclidean structure by explicitly modelling relationships between entities as nodes and edges in a graph in an m-dimensional Euclidean space. In spatiotemporal settings, nodes typically represent spatial locations, while edges encode interactions and learned dependencies. [GNNs](#) propagate information through the graph via message passing, where each node updates its representation by aggregating features from its neighbours ([Scarselli et al. 2008](#)). This allows the model to capture spatial dependencies that cannot be represented by grid-based convolutions. Temporal dynamics are commonly incorporated by combining [GNNs](#) with sequence models (e.g., [Gated Recurrent Units \(GRUs\)](#), [Long Short-Term Memorys \(LSTMs\)](#)) or attention mechanisms, resulting in architectures that jointly model spatial structure and temporal evolution ([Yu et al. 2017](#)). An illustration of such a network can be seen in [Fig. 9](#).

Interpretability in [GNN](#)-based models often arises from the explicit graph structure and from mechanisms such as attention over edges or nodes, which indicate the strength of influence between connected entities. This allows the learned relationships

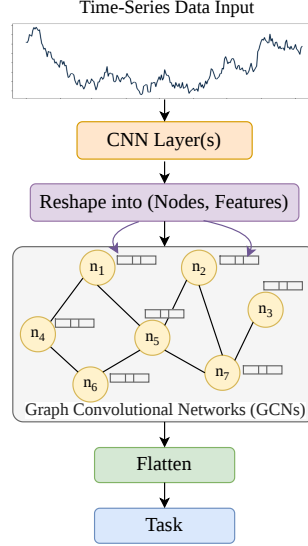


Fig. 9: A sample of GNN implementation for multivariate time-series processing

to be interpreted in terms of domain-specific interactions rather than abstract latent representations (Veličković et al. 2017).

5.3.2 Selected Studies Using Attention-Based Models

A Label-Dependent Attention Model (LDAM) for multimodal disease risk prediction using Electronic Health Records (EHRs) is designed by Niu et al. (2021). Clinical BERT is used to embed medical notes, feature names, and disease labels into a shared semantic space, enabling cross-attention between notes and label embeddings to produce label-specific textual representations. In parallel, time-series features are encoded via channel-wise and cross-channel BiGRUs, and a second cross-attention aligns feature names with labels to weight numerical signals. By weighting the time-series features in this way, alongside encoding the medical notes and disease labels into the same semantic space, we end up with label-conditioned attention weights over both text and structured features. These weights reflect semantic alignment and give the model a high degree of interpretability, as the semantic meaning of the notes and the disease labels directly affect the model’s predictions.

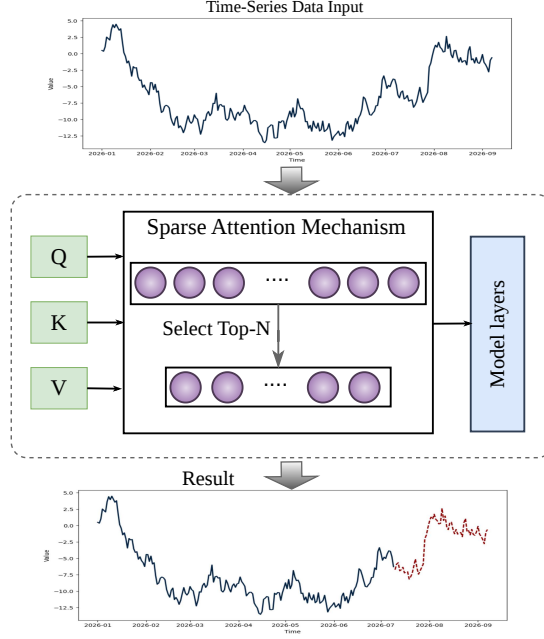


Fig. 10: A schematic of a framework model incorporating a sparse attention mechanism

Xu and Chen (2024) introduced InterFormer, an interpretable transformer for probabilistic short-term load forecasting. A local variable selection network applies causal 1D convolutions and *entmax* activation to produce sparse feature weights at each time step, which are used to reweight the inputs. These representations are processed by a transformer with sparse self-attention (also using *entmax*), aggregated across heads into a single attention map, as illustrated in Fig. 10. The model outputs multi-step probabilistic forecasts via quantile regression. Interpretability is demonstrated through global feature importance derived from selection weights, which select only the relevant input variables, and this is further honed with sparse attention. Another dimension interpretability that can be derived from is that temporal importance maps can be extracted from the sparse attention matrices. The importance maps show how much each time step j contributes to predictions at time t for each layer (Peters et al. 2019).

For the prediction of Bispectral Index (BIS) scores, [Hwang et al. \(2023\)](#) proposed a partially interpretable EEG-based model. EEG signals are transformed via Fast Fourier Transform (FFT) into amplitude and phase components, these components are processed through parallel BiLSTM streams before recombination. The frequency components are modeled with fully connected layers with a time-decayed effect in the form of a Weibull distribution. These are followed by attention layers. The attention weights are constrained to align with FFT dimensions, enabling analysis of frequency importance. Interpretability is demonstrated by stratifying the attention distributions across clinically relevant BIS ranges and examining how amplitude and phase contributions vary.

A sleep staging model was developed by [Ito and Tanaka \(2025\)](#) using single-channel EEG with both time- and frequency-domain inputs. Separate branches are used to process the waveform and spectral features with self-attention applied independently, and their outputs are fused for classification. To incorporate long-term context, multiple sequential epochs are processed in the base model, and the outputs of each epoch are combined in a fully connected layer, and a final classification of the sleep stage is obtained after that. The attention mechanisms are explicitly aligned with clinically relevant waveform and frequency patterns, and interpretability is further demonstrated through visualization of the attention maps, highlighting important time points and frequency bands in accordance with expert sleep scoring criteria.

[Jin et al. \(2021\)](#) presented a Dual-Level Attentional Convolutional LSTM Neural Network (DLA-CLSTM) for arrhythmia detection on two levels: beat and rhythm level. ECG signals are segmented into beats and processed through convolutional and BiLSTM layers. Beat-level attention highlights important waveform segments, while rhythm-level attention captures longer temporal dependencies across beats. The final classification uses a weighted multilabel sigmoid-based classifier to address class imbalance. It is worth mentioning that the architectural specifics (length of beat segments,

input points, and layer sizes) of the model need to be adjusted depending on the dataset. Interpretability is demonstrated by overlaying the attention weights on [ECG](#) signals and verifying alignment with clinically relevant patterns.

DLA-CLSTM was extended into Dual Attentional Convolutional [LSTM](#) Neural Network (DAC-LSTM) in [Jin et al. \(2023\)](#). As opposed to applying attention on a beat and rhythm level, they apply channel-wise attention directly to the raw inputs, followed by convolutional and BiLSTM layers. And then, they follow it up with a second temporal attention mechanism. The final classification is performed using softmax, and they demonstrated interpretability in the same way as in DLA-CLSTM.

With a view to classify the severity of myocardial infarction from 12-lead [ECG](#), [Prabhakararao and Dandapat \(2020\)](#) proposed Multi-Lead Diagnostic Attention-based Recurrent Neural Network (MLDA-RNN). Each lead is encoded separately using [GRUs](#), followed by attention applied to the time steps of each lead (intra-lead attention) to identify important time segments and inter-lead attention to determine the relative importance of the different leads. The aggregated representation is used for classification. Interpretability is achieved by analysing both temporal and lead-level attention patterns, which align with cardiological diagnostic reasoning.

[Jin et al. \(2025\)](#) introduced CrossTabNet for tabular risk prediction. A feature interaction block explicitly constructs nonlinear feature combinations, which are then processed by a TabNet encoder (an encoder for tabular data proposed by [Arik and Pfister \(2021\)](#)) with sequential sparse attention masks. At each decision step, a subset of features (including interaction terms) is selected, and representations are aggregated across steps, behaving like an internal boosting ensemble in order to obtain the predicted risk. Interpretability is obtained from the learned sparse masks, which provide feature importance scores, and from sensitivity analysis that quantifies the directional influence of the features.

An explainable uncertainty-aware [Convolutional Recurrent Neural Network \(CRNN\)](#) (eUA-CRNN) was designed by [Tan et al. \(2020\)](#) for data generation to supplement irregular medical time-series, in order to perform mortality risk prediction. A Gaussian Process first regularizes the data and provides uncertainty estimates, with high uncertainty where data are sparse, and low uncertainty where measurements are dense. The model then decomposes the signal into frequency-based subseries using convolutional filters, processes them with [CNN](#) and [RNN](#) layers, and applies uncertainty-aware attention to weight variables and time steps. Therefore, the variables are weighted according to clinical relevance and uncertainty values. Interpretability is demonstrated by analysing how attention weights balance clinical relevance against uncertainty and how different frequency components contribute to predictions, as opposed to earlier methods that only weigh variables based on uncertainty weights.

[Li et al. \(2022\)](#) presented ILSTM.Soil, a model that predicts soil moisture and temperature. An [LSTM](#) is used to process the variables as a multivariate time-series with multi-feature attention applied to the output, combining variable and time interactions. Attention is applied to this multi-feature attention output in combination with the input data across variables, aggregated over time and across time steps, and aggregated over variables separately. This produces separate variable-level and temporal importance representations. These are combined for the final prediction. Interpretability is derived from the learned attention weights, which indicate both variable importance and influential historical time steps.

A new framework for real-time [Heart Rate \(HR\)](#) monitoring from facial videos using remote photoplethysmography (rPPG) was presented by [Liu et al. \(2024a\)](#). Empirical mode decomposition decomposes the rPPG signal into intrinsic mode functions, and a channel-wise attention mechanism learns to weight these frequency components. The model emphasizes physiologically relevant IMFs while suppressing

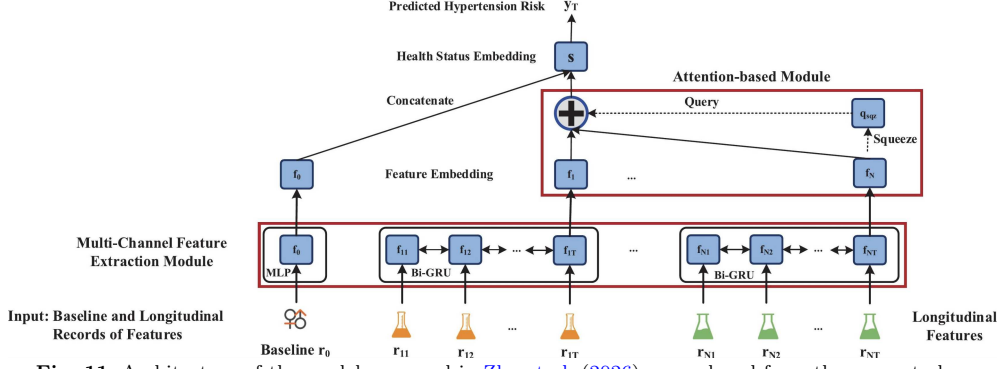


Fig. 11: Architecture of the model proposed in [Zhu et al. \(2026\)](#), reproduced from the same study

noise and irrelevant components, providing interpretable insight into frequency contributions while maintaining real-time performance. Experimental results show the model achieves strong agreement with ground truth while maintaining inference under 100 ms on mobile CPUs.

For hypertension risk prediction, [Zhu et al. \(2026\)](#) proposed an attention-based model using longitudinal EHR data. Dynamic features like vitals, labs, etc., are encoded with BiGRUs and static features like age, sex, etc., are encoded with an Multi-layer Perceptron (MLP). The final representation for each entry is a concatenation of the dynamic and static features. Attention is applied using a query derived from the combined patient representation and keys from feature embeddings. A sparse activation function is used to provide feature importance, improving interpretability. Additionally, the decision boundaries from the learned attention weights are used to derive dynamic risk thresholds. The features, attention weights, and identified thresholds are used to construct causal graphs that identify medically meaningful causal relationships via the PC (Peter-Clark) causal inference algorithm. A visual of the model architecture can be seen in [Fig. 11](#).

Fault diagnosis under variable speed conditions, where traditional methods struggle due to temporal signal distortion is a critical challenge due to the need for precise localization of fault related patterns in time-series signals. [Kim et al. \(2026\)](#) proposed

a Multi-Scale Signal Transformer (MSSIT) with a novel multi-scale self-attention mechanism that adjusts key and value scales while preserving query resolution. They also introduced a Signal Processing-Based Attention Interpretation (SPAI) method to analyze attention weights in time and frequency domains. Validation on two bearing datasets showed superior performance over existing methods.

AM-PTCN (Attention Mechanism-Parallel Temporal Convolutional Network) was presented for the joint estimation of three battery states: State of Charge (SOC), State of Health (SOH), and State of Temperature (SOT) from multivariate time-series inputs by Wang et al. (2026). The architecture uses two parallel Temporal Convolutional Network (TCN) branches, each processing the input independently via causal dilated convolutions to capture temporal dependencies at multiple scales, as the different states rely on different scales. An attention mechanism is then applied to each branch’s extracted features, re-weighting the input variables according to their contribution to that branch’s target state before the three outputs are softly fused. Finally, uncertainty is quantified using Gaussian Negative Log-Likelihood (NLL) loss, so the model outputs both a prediction and a confidence estimate. Interpretability is demonstrated through the per-branch attention weights, which reveal the distinct physical dependencies of SOC, SOH, and SOT on the measured signals.

Similarly, Jeyarani (2026) proposed Arrhythmia Classification with Interpretative Temporal Convolutional Transformer (AC-ITCT) for 12-lead ECG arrhythmia classification, combining dilated window CNNs for morphological feature extraction with a TCT and time attention mechanism for capturing temporal dependencies, with interpretability demonstrated by visualising attention weights over ECG segments to highlight which temporal regions drive each classification decision.

Zhang et al. (2026) proposed TANet (ThermoAttention Network) for dynamic Mean Skin Temperature (MST) estimation during physical exercise. Rather than

directly regressing [MST](#), the model is trained on a proxy task: classifying the exercise condition (rest, hiking, lifting) and then using the produced attention weights as adaptive body segment weights for [MST](#) estimation, replacing the fixed anatomical weights of traditional formulas. N parallel [LSTM](#) encoders process the local skin temperature time-series from each body segment independently. Their outputs are fused by an additive attention layer, with a shared linear scoring function producing a scalar per channel, and a *softmax* normalises these into weights β_i , and the weighted sum is passed to a linear classifier. After training, β_i directly serves as the segment contribution weights for [MST](#). Interpretability is demonstrated by averaging attention weights per segment per condition across subjects, revealing physiologically consistent patterns. The attention interpretation is further validated with perturbation masking and gradient attribution as independent verification, with the authors explicitly acknowledging that attention weights alone are not absolute importance measures.

5.3.3 Selected Studies Using Interpretable [CNNs](#)

[Borra et al. \(2020\)](#) introduced Sinc-ShallowNet, an interpretable [CNN](#) for decoding motor execution and imagery from [EEG](#). The model replaces standard convolutions with a sinc-convolutional layer that learns band-pass filters using only two parameters per filter, making spectral features directly interpretable. This is followed by a spatial depthwise convolution that learns electrode-specific spatial filters for each frequency band, enabling identification of relevant brain regions. Interpretability is further supported through gradient-based temporal sensitivity analysis, which highlights the most relevant frequency bands and filters for each class. Based on the result, Sinc-ShallowNet outperformed traditional methods and other [CNNs](#) while using significantly fewer parameters, and the learned features matched known neurophysiology to show contralateral sensorimotor activity for hand movements and central activity for feet movements. This framework was extended with MS-Sinc-ShallowNet in [Borra et al. \(2023\)](#). They maintained the separation of spectral and spatial feature learning

while introducing a multi-scale design and evaluating the model under within-subject, cross-subject, and transfer learning settings. They further enhance interpretability by incorporating saliency maps to localize EEG features relevant for decoding continuous hand movement (position and velocity), rather than discrete classes. The model remained lightweight and outperformed comparable deep networks while being 30 times smaller. The delta frequency band (0.18–4 Hz) consistently emerged as the most important feature across participants, particularly over sensorimotor and parieto-occipital brain regions.

SincNet was also applied as an interpretable DL architecture to study emotion recognition in individuals with Autism Spectrum Disorder (ASD) by Mayor-Torres et al. (2021). They implemented it using EEG signals during a facial emotion recognition task. The first layer of SincNet, which learns band-pass filters parametrised by only two cutoff frequencies per filter and makes the learned features directly interpretable in the frequency domain. After training on EEG data from 88 adolescents, the authors analysed the frequency responses of the learned filters and discovered that the model automatically attenuated high-alpha and beta bands for ASD participants. This finding aligns with neuroscience research and shows that suppression in these bands is associated with behavioural deficits in ASD. The model outperformed both human performance and standard CNN baselines and demonstrated that interpretability need not come at the cost of performance.

Li et al. (2025a) introduced a framework for detecting Alzheimer’s disease risk in cognitively healthy individuals using task-related EEG and addressed the critical challenge of identifying preclinical AD before symptoms emerge. The authors employed an interpretable CNN that combines pointwise and depthwise convolutions with class activation mapping to reveal which spatial and spectral features drive predictions. The model was trained on EEG data from 51 middle-aged participants during two cognitive tasks and achieved good performance. Crucially, the interpretability analysis

revealed that the model consistently focused on theta and alpha band activity in parietal and temporal regions for high-risk individuals, while non-risk samples showed frontal activation patterns and aligned with established neuropathology.

An interpretable temporal convolutional framework (ITCF) based on [TCN](#) with a novel convolutional hierarchical group Lasso (cHGL) regularization was proposed by [Dong et al. \(2025\)](#). They integrated the iterative soft-thresholding algorithm (ISTA) into backpropagation to induce sparsity both across channels and within each channel’s temporal kernels. This design enables the model to identify which variables are causally related and automatically determine the maximum time delay by interpreting kernel weights along the dilated convolution paths. The framework was validated on both linear (VAR) and nonlinear (Lorenz-96) simulations, as well as real-world [Functional Magnetic Resonance Imaging \(fMRI\)](#) and gene expression data.

5.3.4 Selected Studies Using Graph NNs

Towards the development of new traffic flow prediction models, [Ji et al. \(2020\)](#) proposed a new interpretable [DL](#) framework that models urban traffic dynamics through latent spatiotemporal potential energy fields (ST-PEFs) and draws an analogy to water flow driven by gravity. The key insight is that traffic flow, while containing cycles that violate potential field concepts, can be decomposed into acyclic poly-tree components using a Wind field Decomposition algorithm. The model first predicts how potential energy fields evolve using [GRU](#) for temporal dynamics and graph attention networks for spatial dependencies, then derives predicted traffic flows from the fields. Experiments on three real-world taxi datasets show the model outperforms state-of-the-art methods like ST-ResNet and STDN in prediction accuracy.

[Xin et al. \(2021\)](#) introduced an explainable [DL](#) framework called WTRPNet for epileptic [EEG](#) classification that transforms raw time-series signals into interpretable graph-based images. The method first applied wavelet transform to decompose [EEG](#) signals into different frequency components, then used recurrence plot analysis on the

low-frequency coefficients to create 2D images that visually capture the dynamic recurrence patterns of brain activity. These images are then fed into a convolutional neural network for classification. The WTRP block makes the feature extraction process inherently explainable. Clinicians can visually inspect the recurrence plots to understand why the network classifies a signal as epileptic or not, as focal seizures create characteristic geometric structures different from non-focal activity. Experiments show the method outperform traditional methods while providing visual interpretability.

The proposed interpretable DL framework for predicting high-speed train station delays by Zhang et al. (2022a) combines graph neural networks with fuzzy decisions. The model first applies community detection using the Louvain algorithm, to partition the complex railway network into subgraphs where trains have strong mutual influence and reduce noise from negligible inter-community interactions. It then used a multi-attention graph convolutional network (C-MATGCN) with spatial, temporal, and multi-factor attention mechanisms to capture spatiotemporal dependencies and external factors such as weather and holidays. To make the model interpretable, the authors fuzzified numerical attributes into linguistic terms and trained a time-series fuzzy decision tree that approximated the C-MATGCN’s predictions. Experiments on a real-world dataset shown C-MATGCN significantly outperforms baseline models across 1-6 hour predictions. Zurlo et al. (2026) proposed a hybrid framework for anomaly detection in multivariate IT time-series using Generative Adversarial Networks (GANs), autoencoders, and GNNs. The authors represented time-series as correlation-based static graphs and trained two interconnected GANs to regularize reconstruction. Anomalies were detected via reconstruction error and discriminator scores, while graph visualization enabled interpretable root cause analysis by highlighting anomalous nodes and edges. Validation on a real data center case study and two benchmarks demonstrated robust performance.

Nelson et al. (2026) proposed a brain-inspired graph neural network for multimodal neonatal sepsis detection using both EEG waveforms and electronic clinical records. The authors modeled EEG channels as anatomically grounded graph nodes and employed a forward-forward graph convolutional network with spike-timing-dependent plasticity instead of traditional backpropagation. This biologically plausible architecture effectively captured temporal and spatial dependencies across modalities. The experimental evaluation demonstrated reliable and robust classification performance for sepsis assessment in NICU settings.

5.3.5 Selected Studies Using Other Interpretable Neural Models

To detect Atrial Fibrillation (AF) from ECGs, Jo et al. (2021) designed a model with two separate interpretable modules that explicitly learn the two key clinical features of AF, including irregular heart rhythm and absence of P-waves. By concatenating these modules, the final model not only detects AF but also provides the specific reasoning behind each prediction. The model was developed using ECGs from two Korean hospitals and externally validated on three public datasets with different lead configurations. Results showed acceptable performance for AF detection on all validation sets, significantly outperforming standard ECG machine interpretations.

Kobler et al. (2022) introduced a novel geometric DL framework for unsupervised domain adaptation in EEG-based brain-computer interfaces to address the critical challenge of poor generalization across sessions and subjects. They proposed SPD domain-specific momentum batch normalization (SPDDSMBN), which operated on symmetric positive definite (SPD) covariance matrices. The method enabled end-to-end learning of tangent space mapping (TSM) models, which are inherently interpretable because they can be decomposed to reveal the underlying brain sources driving predictions. By tracking domain-specific Fréchet means on the SPD manifold with theoretical convergence guarantees, SPDDSMBN transforms domain-specific

inputs into domain-invariant outputs, enabling effective multi-source/target adaptation. The proposed architecture, TSMNet, combines a feature extractor, SPDDSMBN layers, and a linear classifier, achieving state-of-the-art performance across 6 diverse EEG datasets in mental imagery and workload estimation tasks.

A transparent and explainable DL model for ECG beat classification that clinicians can trust and understand is introduced by Prakash et al. (2023). The authors first train a modified 1D U-Net to detect the exact locations of P waves, QRS complexes, and T waves, the same features cardiologists look at when reading an ECG. Once these important points are found, the model extracts over 12 meaningful features like wave durations and amplitudes, and feeds them into a random forest classifier. This two-step process makes the whole system interpretable because doctors can see that the decisions are based on real clinical features, not hidden patterns. The model outperformed other methods while staying transparent and explainable.

To address point-level detection and classification of EEG artifacts, Yu et al. (2024) constructed WaveNet model. It is a critical task in clinical diagnosis where artifacts like eye movements, muscle activity, or electrode issues can obscure brain signals. The model combines an invertible wavelet decomposition block that extracts signal features without information loss with a novel tree generator that automatically learns the optimal wavelet decomposition structure for each channel based on the signal’s frequency spectrum. This tree generator provides explainability during inference. The authors also introduced Event-Aligned Compensation Scoring (EACS), a fairer event-level metric that evaluates detection of artifact events regardless of their duration. Based on the results, WaveNet achieves superior F1-scores, outperforming larger CNN and Transformer baselines while providing interpretable frequency-based explanations.

Deng et al. (2024) proposed a new counterfactual explanation framework named DiffusionCF for block maxima forecasting models that predict extreme values in time-series. Fig. 12 shows a schematic illustration of DiffusionCF. The method generates

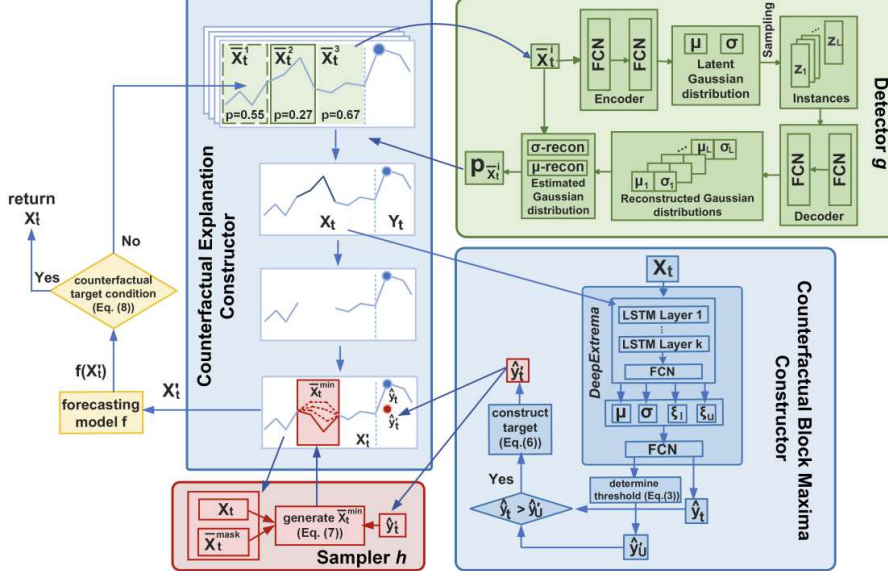


Fig. 12: A schematic illustration of the proposed DiffusionCF framework, adapted from Deng et al. (2024)

counterfactual instances to address three key challenges including informative, close to the original data, and realistic. DiffusionCF first uses the Generalized Extreme Value (GEV) distribution to define an appropriate counterfactual target value, then applies anomaly detection via a variational autoencoder to identify anomalous segments in the input time-series and finally employs a conditional diffusion model (CSDI) to generate realistic replacements for those segments that would lead to a non-extreme forecast. Experiments on weather, financial, and traffic datasets show DiffusionCF significantly outperforms baselines in sparsity and consecutiveness metrics. A case study on traffic data near a stadium demonstrates how the method correctly identifies that an anomalous traffic pattern on a game day explains an extreme traffic forecast for the following day.

XDL-ESI is an explainable framework for EEG/MEG source imaging that reconstructs brain activity from scalp recordings, designed by Jiao et al. (2024). The method connects iterative optimisation algorithms with neural networks by unfolding the update steps of ISTA into network modules. They introduced the topological loss,

which uses geodesic distances on the cortical surface to apply stronger penalties to localisation errors farther from the true source. The framework also allows hyperparameter tuning after training to give clinicians flexibility to adjust sparsity and see multiple solution options. Validated on synthetic data and real clinical datasets, XDL-ESI consistently outperformed benchmark methods, producing more concentrated and accurate source reconstructions for epileptic spike localisation and intracerebral stimulation mapping.

An interpretable disease risk prediction model that first clusters patients into subtypes and then performs disease prediction based on the cluster was built by [Niu et al. \(2024\)](#). The model leverages a Dirichlet Process Mixture Model to infer the number of clusters dynamically, allowing the subtype structure to adapt to the data without pre-specification. Patient visits are encoded into latent representations using Clinical-BERT and expressed as mixtures over learned cluster centers, which serve as clinically meaningful subgroups. The model follows a variational Bayesian framework to capture temporal dependencies across visits. The resulting latent state is used for downstream risk prediction. Training jointly optimizes predictive performance and consistency between prior and posterior distributions. They provide interpretability on two levels. At the local level, Clinical-BERT highlights clinically relevant features contributing to the representation. At the global level, the learned clusters correspond to patient subtypes, enabling subgroup-level reasoning and more structured insight into disease risk patterns.

To model dam deformation, [Zhu et al. \(2024\)](#) developed a hybrid deformation prediction model for high-arch dams that combines interpretable statistical modelling with DL-based residual correction. The approach begins with preprocessing environmental signals with FFT to extract dominant periodic components, reducing input complexity while preserving meaningful physical patterns. A Bayesian Ridge Regression model is first used to capture the primary relationship between the environmental

factors and structural deformation. To account for the remaining nonlinear and temporal effects, the residual errors from this model are then modeled using a [LSTM](#) network. The final prediction is obtained by combining the outputs of both stages. Interpretability comes from the structure itself. The Bayesian Ridge component remains transparent and physically meaningful, allowing analysis of how temperature-related factors influence deformation. In contrast, the [LSTM](#) is restricted to modelling only the residuals, which limits its role and prevents it from becoming a full black-box model. This creates a balance where most of the prediction is explainable, and only the complex leftover patterns are handled by the deep model.

[Lei et al. \(2025\)](#) introduced [Temp²-Kolmogorov–Arnold Network \(KAN\)](#) architecture for climate temperature forecasting based on the Kolmogorov-Arnold representation theorem, which states that any multivariate function can be decomposed into sums and compositions of univariate functions. Unlike conventional multilayer perceptrons that use fixed activation functions on neurons, [Temp²-KAN](#) places learnable univariate functions on the network edges. They enabled dynamic adaptation to complex temporal patterns while maintaining inherent interpretability. The learned activation functions could be directly visualised to determine which features drive predictions. The model was evaluated on U.K. Met Office climate data from multiple stations and predicted multi-step temperature forecasts from historical observations. [Temp²-KAN](#) achieved state-of-the-art performance and outperformed [RNN](#), [LSTM](#), [GRU](#), and Transformer baselines across most stations.

A framework for monitoring punch wear in fine blanking processes was presented by [Becker et al. \(2025\)](#) by combining variational autoencoders with causal inference principles. They used an Ada-GVAE with the independent causal mechanisms (ICM) constraint, which forces the model to learn latent factors that correspond to distinct real-world causes of variation in the force signals. The framework successfully disentangled two factors, including one capturing that is a bias from machine stops during

wear measurements, and another isolating genuine wear-related changes. A predictor network then estimates wear from these interpretable latent factors, and gradient-based analysis reveals feature importance. Experiments show that while the model achieves excellent predictive performance, the explainability analysis reveals it was primarily relying on the measurement bias factor rather than the true wear feature. The framework also successfully identified synthetically injected features, validating its causal disentanglement capability.

To address battery capacity estimation, [Li and Zhang \(2025b\)](#) introduced a transparent approach using a Zero-Pole-Gain (ZPG) model fitted to electrochemical impedance spectroscopy (EIS) data. The fourth pole (Pole4) is extracted as an interpretable feature that reflects the time constant of charge transfer processes around 40 Hz, and its shift during aging corresponds to increased resistance and slower battery response. The method requires no prior battery knowledge, is robust to temperature variation and measurement noise, and achieves estimation errors as low as 1.53% on 12 commercial cells tested at different temperatures. With integrating into time-series models like [AutoRegressive with eXogenous inputs \(ARX\)](#) and [LSTM](#), the Pole4 feature significantly improves multi-step ahead capacity forecasting.

[Zhao et al. \(2026\)](#) built a traffic forecasting model that considers the spatial and temporal aspects of multiple temporal scales in the traffic sequence, a Multi-Scale Spatio-Temporal Neural Network (MSTNN). To enhance interpretability, they used [KANs](#) instead of standard [MLPs](#). [KANs](#) replace standard linear weight matrices with learnable nonlinear functions, making feature transformations more explicit and interpretable ([Liu et al. 2024b](#)). The model processes the input sequence by decomposing it into patches representing different temporal resolutions, allowing parallel learning of short- and long-term patterns. Spatial and temporal dependencies are modeled through dedicated modules that operate on each temporal scale independently. Spatial relationships between traffic nodes are learned using attention-aware mechanisms,

while temporal dynamics are captured through channel-wise and cross-channel interactions. The outputs from different temporal scales are then aggregated to produce the final forecast. The interpretability of their method comes from their usage of [KANs](#) as opposed to regular [MLPs](#), and the spatial attention weights in the SA-KAN module also allow the examination of which nodes influence prediction the most.

Also leveraging the interpretable power of [KANs](#), [Liu et al. \(2026\)](#) proposed a non-invasive method for detecting atrial fibrillation using heart sound signals instead of [ECG](#). The authors introduced a framework that combines a dual-branch feature extraction, attention-based fusion, and interpretable nonlinear modelling. It includes mel-frequency cepstral coefficients (MFCCs) and bispectrum to fuse complementary features, followed by an attention mechanism for adaptive feature integration. These features are then classified by a [KAN](#) to capture complex nonlinear patterns in cardiac acoustics.

Neural network-based models are generally limited by non-negligible drawbacks such as interpretability issues of feature learning. To address that, [Lou et al. \(2026\)](#) introduced a framework to monitor molten iron quality (MIQ) in blast furnaces. It combines [Shallow Correlation Network Representation Regression \(ShCNRR\)](#) and [Deep Correlation Network Representation Regression \(DeCNRR\)](#) models based on [Canonical Correlation Analysis \(CCA\)](#). By integrating a correlation-based loss function and an inverse twin network, they established an explicit and interpretable mapping between process variables and molten iron quality. Their method was validated on real-world blast furnace data and showed superior regression and fault detection performance.

5.3.6 Selected Studies Using Case- and Rule-Based Models

Another group of models reviewed in this survey that employ neural architectures are case- and rule-based models. However, case- and rule-based reasoning constitutes a

distinct modelling paradigm in its own right, rather than a sub-type of neural network architecture; the neural components in these models merely serve as a means of achieving that paradigm. For this reason, these models are discussed separately in Section 5.4, rather than as a subsection of interpretable neural networks.

Table 3: An Overview of Interpretable Deep Neural Network Models for Time-Series

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Tan et al. (2020)	eUA-CRNN	Interpolation, Frequency Decomposition, CNN, RNN, Attention	PUB data set, MIMIC-III Clinical, eICU Collaborative Research data set	Classification	Mortality Risk Prediction	ROC, AUROC Score
Prabhakararao and Dandapat (2020)	MLDA-RNN	Attention, RNN (GRU)	STAFF III and PTB	Classification	Myocardial Infarction Severity Prediction	Overall Sensitivity, Specificity, and Accuracy, Class-wise Sensitivity, Specificity, and Accuracy
Borra et al. (2020)	Sinc-ShallowNet	A CNN with sinc-convolutional layer + spatial depthwise convolution + temporal sensitivity analysis for feature relevance	High-Gamma, BCI Competition IV-2a	Motor execution and motor imagery classification	BCI	Accuracy

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Ji et al. (2020)	ST-PEF	Physics-inspired interpretable framework: community partition, Wind field Decomposition (WD) algorithm, Temporal modelling (GRU) + spatial modelling, Flow prediction from predicted fields	Xi'an, Beijing, Porto taxi datasets	Traffic flow prediction	Urban traffic management, congestion prediction, crowd safety, urban planning	RMSE , R^2
Mayor-Torres et al. (2021)	SincNet-based emotion recognition	Interpretable CNN with sinc-convolutional layer + 3 standard conv-pool blocks + fully connected layers + softmax classifier	EEG from 88 adolescents	Emotion recognition	Clinical assessment of emotion recognition deficits in autism, potential biomarker discovery for ASD	Accuracy, PSD , F-tests
Jo et al. (2021)	DLM	Supervised DL with two parallel modules using residual blocks and 1D convolutional layers	Sejong ECG , PTB-XL, Chapman-Shaoxing ECG Database, PhysioNet/CINC challenge 2017	Atrial fibrillation detection	Clinical ECG analysis for arrhythmia screening and diagnosis	AUROC , Sensitivity, Specificity, PPV , NPV , Accuracy

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Xin et al. (2021)	WTRPNet	Wavelet transform + recurrence plot (RP) + CNN with multiple convolutional, pooling, and dense layers	Bern-Barcelona, Bonn EEG Datasets	Epileptic EEG classification (binary and multi-class)	Epilepsy diagnosis, Seizure detection, clinical decision aid	Accuracy, Specificity, Sensitivity
Niu et al. (2021)	LDAM	Clinical-BERT, Multimodal label-dependent attention	MIMIC-III Clinical	Classification	Disease Prediction	Micro- and Macro-Precision, Recall, AUROC
Jin et al. (2021)	DLA-CLSTM	CNN , LSTM , Attention	MIT-BIH Arrhythmia, CPSC2018	Classification	Arrhythmia Detection	Accuracy, macro F1 score
Kobler et al. (2022)	TSMNet	SPD domain-specific momentum batch normalization (SPDDSMBN)	BCI Competition IV-2a, BNCI Horizon 2015-001 Motor Imagery, OpenBMI EEG , Stieger2021, EEG Motor Imagery (ETH Zurich), Mental Workload, Tetraplegic Pilot BCI Datasets	Unsupervised domain adaptation for EEG classification	BCI , clinical EEG applications, mental state monitoring	Balanced accuracy, Fit time

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Zhang et al. (2022a)	C-MATGCN + Fuzzy Decision Tree	Community detection, multi-attention GCN with spatial, temporal, and multi-factor attention for each community, multi-objective optimized time-series fuzzy decision tree for rule extraction	High-speed railway network dataset	High-speed train station delays prediction (1-6 hour forecasts)	Railway dispatching, preventive maintenance, passenger information systems, operational efficiency	MAE , RMSE , MAPE , prediction approximation
Li et al. (2022)	ILSTM_Soil	LSTM, Attention	FLUXNET2015	Regression	Soil Moisture and Temperature Prediction	R^2 , RMSE , MAE , Bias
Prakash et al. (2023)	1D U-Net + Random Forest	Modified 1D U-Net + Random Forest	LUDB, MIT-BIH Arrhythmia	ECG beat fiducial point detection, AAMI standard beat classification	Clinical arrhythmia diagnosis, patient-specific ECG monitoring	Accuracy, Sensitivity, Positive Predictivity, F-score
Borra et al. (2023)	MS-Sinc-ShallowNet	A CNN with sinc-convolutional layers for spectral filtering + depthwise spatial convolution + dual-scale temporal feature extraction	EEG signals recorded by Graz BCI group	Trajectory decoding	Healthcare BCI	RMSE

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Hwang et al. (2023)	BIS Inter-pretable DL	FFT, LSTM, Attention	VitalDB, AMC dataset	Regression, Classification	Bispectral Index Score Prediction, Hypnosis Maintainance	RMSE, Accuracy, Specificity, Sensitivity, AUROC
Jin et al. (2023)	DAC-LSTM	CNN, LSTM, Attention	PhysioNet / CINC challenge 2011	Classification	ECG Quality Evaluation	Accuracy, Sensitivity, Specificity
Liu et al. (2024a)	EMD + Channel-wise Attention CNN	EMD decomposition into 5 IMFs + SE-block channel attention	VIPL-HR dataset, self-collected videos	HR estimation	Remote patient monitoring, stress assessment, fitness tracking, driver monitoring	SD, MAE, RMSE, Pearson's r
Yu et al. (2024)	WaveNet	U-Net + wavelet decomposition blocks + tree generator + Transformer-based decoder	Temple University EEG Artifact Corpus (TUAR)	Point-level EEG artifact detection and multi-class classification	Clinical EEG diagnosis support, automated artifact screening, reducing clinician workload	Precision, Recall, F1

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Deng et al. (2024)	DiffusionCF	Combining GEV distribution for counterfactual target selection, VAE for anomaly detection in time-series segments, (3) CSDI for generating realistic segment replacements	GSOD, S&P 500, Dodgers loop sensor	Counter-factual explanation for block maxima forecasting	Disease outbreak early warning, extreme weather explanation, traffic anomaly attribution, financial risk assessment	Sparsity, Consecutive-ness, Proximity, NLL , Precision
Jiao et al. (2024)	XDL-ESI	Unrolled optimization network (UONN)	Synthetic EEG data, Brainstorm tutorial epilepsy data, simultaneous scalp EEG/SEEG	Epileptogenic zone localization, intracerebral stimulation mapping, presurgical evaluation	Electro-physiological Sources Imaging	(LE), AUROC , computational time
Xu and Chen (2024)	InterFormer	Local Variable Selection Network, Sparse Self-Attention	Real-world household PV Data	Regression	Residential Load Point and Probabilistic Forecasting	RMSE , MAE , MAAPE , Mean Pinball Score, Mean Winkler Score
Zhu et al. (2024)	Hybrid Deformation Prediction Model	FFT + Bayesian Ridge Regression + Residual error modelling with LSTM	Self-collected meteorological and Dam Deformation Data	Deformation Prediction	Environmental and Climate Monitoring	R^2 , RMSE , MAE

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Niu et al. (2024)	DirPred	Predictive Clustering, Dirchlet Process, Clinical-BERT	MIMIC-III Clinical, N2C2-2014	Classification	Disease Prediction	Micro- and Macro-Precision, Recall, F1 score AUROC , and Accuracy
Li et al. (2025a)	ICNN	Interpretable CNN with pointwise convolution + depthwise convolution + CAM-based visualization	PEARL-Neuro Database	Binary classification of Alzheimer's genetic risk	Preclinical Alzheimer's screening	AUROC , Cohen's Kappa, Sensitivity
Lei et al. (2025)	Temp ² -KAN	KAN with B-spline parameterized univariate functions on edges	U.K. Met Office Central England Temperature (CET) dataset	Multi-step temperature forecasting	Climate informatics, operational weather forecasting, environmental monitoring, carbon emission tracking	MSE , RMSE , MAE , MAPE
Becker et al. (2025)	Ada-GVAE + Predictor	Causal representation learning + MLP predictor with gradient-based feature importance + generative decoding	Fine Blanking Punch Force Signals	Explainable condition monitoring, Punch wear estimation	Industrial sheet metal shearing, predictive maintenance, tool wear monitoring	R^2 , Mean feature importance

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Li and Zhang (2025b)	ZPG model with Pole4 feature	Data-driven system identification combined with Polynomial, SVM , GPR , ARX , LSTM	12 <i>LiCoO₂</i> /graphite cells	Battery capacity estimation, multi-cycle forecasting	Battery health monitoring, predictive maintenance	RMSPE , Spearman correlation
Ito and Tanaka (2025)	Sleep Satelight FTC	Attention	Sleep-EDF Database Expanded	Classification	Sleep Scoring	Accuracy, Macro-F1 score, Kappa Coefficient (Cohen's kappa κ), No. of Model Parameters
Jin et al. (2025)	CrossTab Net	Polynomial Feature Interaction, Sparse Attention, TabNet	Self-collected traffic data	Regression	Risk-collision Prediction	RMSE , MAE
Dong et al. (2025)	ITCF	TCN with convolutional hierarchical group Lasso (cHGL) and ISTA-optimized sparse backpropagation	Mathematically generated datasets: VAR simulation and Lorenz-96, BOLD fMRI , DREAM-3 gene expression	Granger causality detection and time delay estimation	General Applications	AUROC , Lag selection accuracy
Zhao et al. (2026)	MSTNN	KAN , Spatial Attention	PEMS03, PEMS04, PEMS07, PEMS08	Regression	Traffic Forecasting	RMSE , MAE , MAPE

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Zhu et al. (2026)	IDL for Hypertension Risk Prediction	Attention	HCBDMF	Hypertension Risk Prediction, Classification	Healthcare	AUROC , AUPRC , Recall, Specificity, Precision, Accuracy, F1 Score
Liu et al. (2026)	DenseNet- KAN	Dual-branch DenseNet for feature extraction, channel-attention fusion, and KAN classifier	Private clinical dataset, Classification of Heart Sound Signal Using Multiple Features	Classification, AF detection	Healthcare	Accuracy, AUC , Precision, Recall, F1-score, Specificity
Lou et al. (2026)	ShCNRR / DeCNRR	Combining shallow and deep correlation network regression models based on CCA	Industrial data from a blast furnace	Regression and Fault detection	Industrial process monitoring	RMSE , R^2 , detection time, Spearman & Kendall correlation tests
Wang et al. (2026)	AM-PTCN	Attention, Parallel Temporal Convolutional Networks, Causal Dilated Convolutions	BatteryAgingARC-FY08Q4 from NASA, cylindrical LiFePO4 battery aging dataset (LFP, 3.7 V, 20 Ah)	Regression; Battery State Estimation	Battery Health Monitoring	RMSE , MAE
Zhang et al. (2026)	TANet	Proxy classification task before regression, LSTM , Attention	Self-collected Mean Skin Temperature Data	Classification + Regression	Healthcare	Accuracy, Macro Precision, Macro Recall

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Jeyarani (2026)	AC-ITCT	Attention, TCT, Dilated Window Convolutions	MIT-BIH Arrhythmia	Classification between types of Arrhythmia, Arrhythmia Detection	Healthcare	Accuracy, Precision, Recall, F1 Score
Nelson et al. (2026)	Bi-GNN	Forward-Forward Graph Convolutional Network (F-FGCN) with spike-timing-dependent plasticity (STDP), anatomically informed graph structures, and multimodal (EEG + ECR) fusion.	MIMIC-III Clinical, MIMIC-III Waveform Databases	Neonatal sepsis detection	Healthcare	Accuracy, Precision, Recall, F1-score, Loss, AUC
Zurlo et al. (2026)	Hybrid GAN-GNN-Auto-encoder	Reconstruction-based framework with two coupled GANs (WGAN-GP), VGAE, correlation-based static graph construction	INFN CNAF Tier-1 data center, The Pooled Server Metrics (PSM), The Server Machine Dataset (SMD)	Anomaly detection	IT infrastructure monitoring	F_β scores
Kim et al. (2026)	MSSIT	Transformer with multi-scale self-attention, SPAI via attention rollout and STFT	SNU Bearing, Ottawa University Bearing Datasets	Rotating machinery fault diagnosis	Industry	Accuracy, Time consumption

5.4 Case and Rule-Based Models

This section presents models that reason by comparing new instances of data to a set of example cases, or through explicit if-then rules, rather than modelling decision boundaries through explicit parameters or coefficients. As with the previous section, the selected studies are preceded by a brief theoretical description of prototypical networks and rule-based reasoning. It should be noted that, while case- and rule-based reasoning is conceptually independent of any particular model architecture, almost all of the studies reviewed in this section implement this paradigm using neural network components.

5.4.1 Foundations

Case-Based Reasoning: Analogous to the human problem-solving process, in which a person draws on similar past experiences to solve a problem, this method of modelling uses previously encountered cases, or created representative examples; learned prototypes of the different classes within the training data to make predictions ([Aamodt and Plaza 1994](#); [Kolodner 1992](#)). This form of reasoning is inherently interpretable, as predictions can be traced back to specific, human-understandable reference points, i.e., the particular case or prototype on which it was classified.

Prototypes: These are representative instances or summaries learned from the training data for each class or region of interest, rather than retaining raw historical cases directly. When a new point is to be classified or predicted, it is compared against these learned prototypes, using a measure of similarity ([Kim et al. 2014](#); [Li et al. 2018](#)). An additional advantage of using prototypes is that domain experts can inspect them to verify their correctness.

Rule-Based Reasoning: In contrast, this method encodes knowledge through rules (typically explicit if-then rules) derived by domain experts, or in the case of some

architectures, automatically generated rules to perform prediction. The interpretability of this method can be attributed to the fact that each prediction can be examined to see which rules were applied to reach the given conclusion (Buchanan and Shortliffe 1984).

Some approaches combine both paradigms, for instance, by first applying rules to obtain a preliminary solution, then drawing on cases to handle exceptions the rules fail to cover, or use cases to support or refine rule-based inference (Golding and Rosenbloom 1996; Hatzilygeroudis and Prentzas 2004).

5.4.2 Selected Studies using Case and Rule-Based Models

For sequence classification that explains predictions by comparing inputs to learned prototype sequences, Ming et al. (2019) proposed ProSeNet. The model used a LSTM encoder to map sequences into a latent space, then computed similarity scores with prototype vectors stored in a prototype layer. It was followed by a non-negative linear classifier that shows how much each prototype contributes to each class. After training, each prototype is projected onto its closest real training sequence and simplified via beam search to keep only the most relevant events. Experiments illustrated that ProSeNet achieved accuracy comparable to state-of-the-art LSTMs while providing intuitive explanations.

In a similar vein, Ni et al. (2021) introduced SCNPro for sequence classification that explains predictions by learning prototype segments from training data. The model compared every part of an input sequence with learned prototypes, calculated similarity scores, and used them for classification with non-negative weights, making it clear which prototype contributed to each decision. After training, each prototype was replaced with its closest real training segment, ensuring every prototype corresponds to an actual, human-understandable subsequence.

On the other hand, the presented model by De La Cruz et al. (2024) predicts hypoglycemia in diabetic patients using structured grammatical evolution. Their model

generated white-box rules in the form of if-then-else statements that combined glucose values, HR, steps, and calories from the previous two hours. The method produced personalized, cluster-based, and general models, with personalized models achieving the best accuracy while maintaining full interpretability through human-readable logical expressions.

Another model developed for a medical application is one proposed to localize the origin of Premature Ventricular Contractions (PVCs) from ECG signals by Kujime et al. (2024). Thereby helping doctors to plan catheter ablation procedures. This model performs semantic segmentation to classify every single point in the ECG signal into four categories, including background, normal rhythm, PVC from the left ventricle, or PVC from the right ventricle, with an underlying ECG-SegNet deep architecture. The model also included a "Neutral" category for unclear cases, flagging them for doctor review rather than forcing a wrong answer. The model achieved high accuracy while providing visual explanations of its decisions.

Haule et al. (2025) introduced a novel interpretable approach called CLaI for detecting and predicting clinical events from low-resolution physiological time-series data commonly available in ICUs. The method first used collaborative learning to map patient data into a latent space where similar physiological states cluster together. Then, for a new patient, collaborative inference compared their latent trajectory with reference patients to predict future events and provide interpretability through case-based reasoning. Based on the results, CLaI achieved competitive F1 scores while being notably robust to class imbalance.

ProtoEEG-kNN is presented for Interictal Epileptiform Discharges (IED) detection in EEG by Tang et al. (2025). IEDs are brief abnormal electrical spikes that are a key biomarker of epilepsy. The model is prototype-based: it learns an EEG embedding space using ProtoPNet; a prototypical network that learns representative training

examples as prototypes, then switches to kNN classification in that space. So predictions take the form of "this EEG looks like these training EEGs". The similarity metric is designed to capture both spike morphology and spatial distribution across channels, aligning the model's notion of similarity with how neurologists actually compare EEGs. Channel masking produces channel-wise importance weights, revealing which electrode locations the model focuses on. Interpretability is justified by showing the most similar training examples and the channels that drove the comparison for each prediction.

An intrinsically interpretable framework for biomedical time-series classification, called ProtoKDM, was introduced by [Montenegro et al. \(2026\)](#). The model combines a bifurcated convolutional autoencoder with a Kernel Density Matrix module that selects real training instances from high-density latent regions as prototypes. Predictions are explained by showing similar prototypes along with temporal similarity maps highlighting waveform deviations. Validation on two datasets showed strong performance with narrow confidence intervals. The modular design decouples representation learning from prototype selection and enables consistent interpretability across different biomedical modalities.

Table 4: An Overview of Interpretable Case and Rule-Based Models for Time-Series

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Ming et al. (2019)	ProSeNet	Combining prototype learning and RNNs, Clustering and evidence regularization	Vehicle fault logs, Yelp Open Dataset, UniProtKB, MIT-BIH Arrhythmia	Sequence classification	Sentiment Analysis, Protein Family Classification, ECG arrhythmia diagnosis (healthcare)	Accuracy, Recall, MAP, Precision
Ni et al. (2021)	SCNPro	Convolutional sequence model by learning local prototypes with adaptation regularization	Yelp reviews, Amazon Product Reviews, MIT-BIH Arrhythmia, PTB Diagnostic ECG	Sequence classification with prototype-based explanations	Sentiment analysis, ECG diagnosis, protein classification	Accuracy, Prototype Visualization
De La Cruz et al. (2024)	SGE / DSGE with rule-based grammar	Structured Gram-matical Evolution using if-then-else grammar with logical/relational/numeric operations interpretable white-box rules	24 T1DM patients with CGM and smartwatch data	Hypoglycemia prediction	Diabetes Management, Early Warning System	TPR, TNR, Bayesian probability of winning
Kujime et al. (2024)	ECG-SegNet + rule-based localization	1D-CNN + bidirectional LSTM for point-wise semantic segmentation + rule-based algorithm for final classification	Private NCVC dataset, Ningbo dataset	Point-wise ECG segmentation + PVC origin localization	Preoperative planning for catheter ablation	Accuracy, Sensitivity, Specificity, F1-score, AUROC, Neutral Rate

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Haule et al. (2025)	CLaI	Collaborative learning with LSTM to embed patients into latent space	KidsBrainIT, MIMIC-IV, CHB-MIT Scalp EEG	Clinical event detection and prediction	ICU monitoring, early warning systems, clinical decision support (healthcare)	AUROC , AUPRC , Sensitivity, Specificity, F1-score
Tang et al. (2025)	ProtoEEG-kNN	Prototypical network ProtoPNet learning prototypes + kNN classification	IEDs in EEG Dataset	IED Detection	Epilepsy Detection, Healthcare	Binary Classification Accuracy, AUROC , R^2
Montenegro et al. (2026)	ProtoKDM	CNN autoencoder + supervised contrastive loss + KDM + UMAP & temporal similarity visualization	GaitRec, PTB-XL	Classification	Healthcare	AUC , Accuracy

5.5 Fuzzy and Neuro-Fuzzy Models

This section reviews fuzzy and neuro-fuzzy models used for interpretable time-series prediction and classification. We first introduce the core theoretical and architectural concepts on which these models are built, before summarizing each method with emphasis on its specific contributions, task, and design choices. The models are grouped according to shared architectural basis.

5.5.1 Foundations

Fuzzy Set Theory: Fuzzy set theory extends classical set theory by allowing elements to belong to a set with varying degrees of membership rather than crisp binary inclusion. Each element is assigned a membership value in $[0, 1]$, enabling gradual transitions between linguistic concepts such as "low", "medium", and "high" as opposed to just belonging or not belonging to a set (Zadeh 1965). This property makes fuzzy systems particularly suitable for modelling uncertainty, vagueness, and gradual change, as is common in real-world time-series data. **Membership Functions (MFs)** define how crisp input values are mapped to fuzzy membership degrees, i.e. to what extent an input belongs to a fuzzy set. Common choices include triangular, trapezoidal, and Gaussian functions (Pedrycz and Gomide 1998). The shape, center, and width of MFs control how uncertainty and variability are represented. In data-driven fuzzy systems, MF parameters are often learned or adapted automatically, allowing the fuzzy partitions to reflect empirical data distributions (Jang 1993; Zadeh 1965).

Fuzzy Rules, Antecedents, and Consequents: **Fuzzy Inference Systems (FISs)** are built from IF-THEN rules. The antecedent (IF-part) specifies conditions on input variables using fuzzy sets, whose membership was determined using MFs, while the consequent (THEN-part) defines the output. An illustration of this can be seen in Fig. 15. In Mamdani-type systems, consequents are fuzzy sets (Mamdani and Assilian 1975), whereas in **Takagi-Sugeno-Kang (TSK)** systems, consequents are crisp functions (typically linear or polynomial, the order of which lends itself to the system name) of the inputs (Takagi and Sugeno 1985). Rule generation can be exhaustive, heuristic, or learned directly from data. The firing strength of a fuzzy rule quantifies the degree to which the rule applies to a given input. It is computed by aggregating the membership degrees of the antecedent conditions using fuzzy operators such as product, minimum, or sum. Firing strengths determine each rule's contribution to the final output and are central to both inference and interpretability (Zadeh 1978). A different

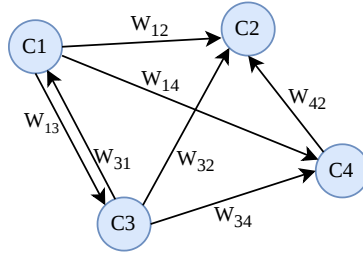


Fig. 13: An illustration of an FCM

type of fuzzy set used to provide an additional layer of uncertainty modelling is called *Interval type-2 fuzzy sets*. These generalise regular fuzzy sets by representing membership degrees themselves as intervals rather than crisp values. To obtain crisp outputs, type-reduction and defuzzification are required (Zadeh 1975; Mendel and John 2002).

Fuzzy Cognitive Maps (FCMs): FCM represent systems as directed graphs where nodes correspond to concepts (variables), and edges represent causal influences with signed fuzzy weights. Node states evolve dynamically over time through recurrent updates, enabling both prediction and causal reasoning (Kosko 1986; Glykas 2010). A sample FCM can be seen in Fig. 13. Interpretability arises from the explicit concept-to-concept relationships encoded in the graph (Kosko 1986).

Fuzzy Neural Networks (FNNs) and Deep FNNs (DFNNs): FNNs integrate fuzzy inference mechanisms within neural architectures. Typically, an FNN consists of an input layer, a fuzzification layer, a fuzzy rule layer computing firing strengths, and a consequent and defuzzification stage as illustrated in Fig. 14. Deep fuzzy neural networks (DFNNs) stack multiple FNNs or embed fuzzy layers within deeper architectures to harness the advantages of DL while maintaining rule-based interpretability (Hu et al. 2017) (Gu et al. 2023).

Rough Set Theory: Rough set theory is a different way of modelling vagueness and uncertainty by positing whether an element of a set either belongs to a set certainly or possibly. This is implemented using lower and upper approximations, with the elements

definitely belonging to a set in the lower approximation and the elements possibly belonging to a set in the upper approximation (Pawlak 1982). In fuzzy-rough models, rough layers aggregate multiple fuzzy rules into compact representations, reducing redundancy while preserving predictive performance and interpretability (Cao et al. 2020).

Neuroevolution and Multiobjective Optimization: Neuroevolution uses evolutionary or genetic algorithms such as genetic mutation, crossover, and selection to optimize neural network structures, architecture, or parameters. In the case of FNNs, it is used to optimize fuzzy network structures (Cord et al. 2001). Multiobjective Evolutionary Algorithms (MOEAs) extends this idea by simultaneously optimizing several objectives; like prediction accuracy, generalization, and interpretability (Ishibuchi and Nojima 2007).

5.5.2 Selected Studies Using Evolutionary Fuzzy and Rough Neural Networks

For rainfall prediction, a Mamdani-type FIS combined with an Artificial Neural Network (ANN) was presented by Kajornrit et al. (2014). The parameters of the FNN are optimized using a genetic algorithm. Temporal dependencies are handled explicitly through lag selection guided by autocorrelation and partial autocorrelation analysis, while seasonality is encoded using a discrete time coefficient to limit rule explosion

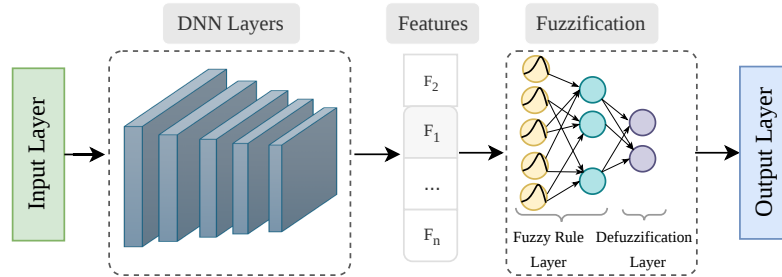


Fig. 14: An illustration of a Fuzzy Neural Network

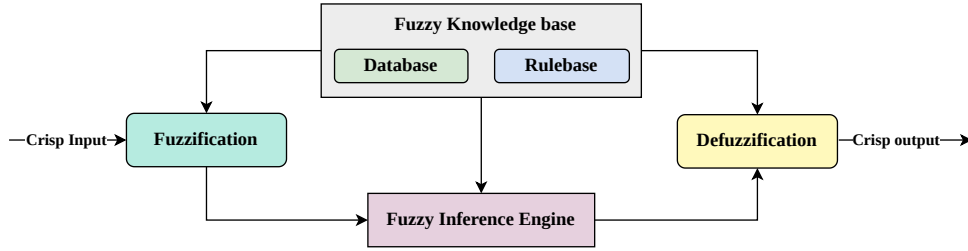


Fig. 15: A summary of a Fuzzy Inference System

and preserve interpretability. MFs are defined in a data-driven manner based on percentile changes in rainfall, resulting in transparent fuzzy rules that capture temporal uncertainty.

Cao et al. (2020) proposed a Fuzzy Rough Neural Network (FRNN) whose structure and parameters are optimized using a MOEA. The model follows a layered fuzzy architecture in which inputs are fuzzified, candidate fuzzy rules are generated and evaluated via firing strengths, and a rough layer aggregates redundant rules into fewer rough nodes. This reduces rule redundancy before mapping to consequents and producing outputs via defuzzification. The evolutionary optimization balances prediction error, generalization, and interpretability, favouring compact rule bases and simple network structures.

The aforementioned framework was extended by stacking multiple FRNN blocks to form a Deep Fuzzy Rough Neural Network (DFRNN) by Zhao et al. (2025). A supernet with predefined maximum depth and width is constructed, and architectural elements, including the number of FRNN blocks, fuzzy rules, MF types, inter-layer connections, and consequent parameters are encoded and optimized via a MOEA. Compared to its 2020 predecessor, this work represents a clear continuation, introducing depth while preserving the same evolutionary and fuzzy-rough principles. On air pollution time-series datasets, the model achieves superior forecasting performance while producing simpler and more interpretable structures than competing deep fuzzy approaches.

To perform energy forecasting, [Lerner et al. \(2021\)](#) used fuzzy rules and evolutionary strategies to generate synthetic data, which are then used to train a simple ANN as the final model. While simpler than later deep fuzzy models, it relies on the same foundational fuzzy rule mechanisms and illustrates how, once fuzzy systems are used to boost the data quality, a simple model can achieve competitive results.

5.5.3 Selected Studies Using Interpretable Fuzzy Transfer Learning

The proposed model by [Jiang et al. \(2025\)](#) is an interpretable fuzzy transfer learning framework for [Steady-State Visual Evoked Potential \(SSVEP\)](#)-based (EEG) BCI. The model is built on a [TSK](#)-type fuzzy neural network, using fuzzy antecedents and functional consequents to maintain transparency. [EEG](#) features are clustered to define fuzzy rule centres, with Gaussian [MFs](#) determining firing strengths. Each centroid represents a prototypical [EEG](#) pattern, with its centre and width learned and optimised during training. A fuzzy attention mechanism, implemented through spatial and temporal fuzzy filters, captures channel relevance and temporal dynamics in an interpretable manner. Transfer learning is achieved in a zero-shot setting by aligning fuzzy rules across subjects rather than retraining, enabling strong cross-subject performance with minimal calibration. Interpretability is demonstrated through explicit rule analysis and physiological consistency with known SSVEP harmonics.

[Jiang et al. \(2017\)](#) extended [TSK](#) fuzzy classification to [EEG](#) seizure detection by integrating transductive transfer learning and semi-supervised learning. Transductive transfer learning is applied to address inter-patient variability by minimizing maximum mean discrepancy between source and target outputs, improving generalization to unseen patients without requiring target labels. Semi-supervised learning is incorporated through fuzzy label memberships for unlabeled target samples, with each sample being assigned membership. These memberships act as additional constraints during training, encouraging samples that are close in feature space to receive similar predictions and helping to refine the decision boundaries. Two variants are proposed:

a simpler fixed-membership model and an adaptive alternating-optimization model. Interpretability is demonstrated by explicitly visualizing fuzzy membership functions, extracting linguistic IF–THEN rules, and tracing individual predictions to dominant firing rules, making the model suitable for clinical EEG analysis.

5.5.4 Selected Studies Using Deep FCMs

For haze prediction and causality mining in haze pollution modelling, Peng et al. (2020) developed a Primary Sub-Fuzzy Cognitive Map (PS-FCM). The complex systems are modeled as interacting subsystems. Separate sub-FCMs represent pollutant factors and meteorological variables, each evolving dynamically over time, while a primary FCM models haze itself. The primary haze state depends on its previous state and weighted inputs from each sub-FCM, with additional coupling weights learned to quantify subsystem influence. All variables are fuzzified before inference; causal weights are learned via gradient descent to minimise prediction error across all variables and time steps, in each subsystem, and across the primary system. All concepts retain explicit semantic meaning, and interpretability is obtained directly from the learned causal weights and relationships between the named concepts. This enables both quantitative forecasting and explicit analysis of how different pollutants and meteorological factors contribute to haze formation.

A deep FCM (DFCM) was introduced by Wang et al. (2021) for multivariate time-series prediction. This is achieved by replacing the linear update equations of classical FCMs with learned nonlinear functions. Concept states are first fuzzified into activation degrees in $[0, 1]$, and each concept update is decomposed into two components: an f -function modelling internal system dynamics (influences from other concepts) and a u -function modelling unknown exogenous effects. The f -function is implemented as an MLP with architectural constraints (no bias, linear output) to ensure consistency with classical FCM behaviour, while the u -function is modeled with an LSTM to capture temporal and seasonal effects. Training uses an alternating

optimization strategy that separately updates f and u to preserve interpretability of internal versus external influences. Instead of fixed causal weights, concept-to-concept influence strengths are recovered via partial derivatives, yielding signed, quantitative causal measures analogous to FCM weights. The model improves predictive accuracy while maintaining interpretable causal relationship analysis.

Another extension of FCMs is deep attention fuzzy cognitive maps (DAFCM) as proposed by Qin et al. (2023) for long-term prediction of nonstationary multivariate time-series. The model extends FCMs by combining a spatiotemporal FCM (STFCM), LSTM, and a temporal FCM (TFCM) with attention mechanisms and residual connections. STFCM captures correlations between concepts at the same time step and short-term temporal dependencies, while the LSTM models longer-term dynamics and global context. TFCM further refines the temporal modelling by using attention to adaptively weight nonstationarity over time. Residual connections are introduced to stabilize training by preventing vanishing or exploding gradients and preserve input information when stacking the graph-based FCM and recurrent components. While interpretability is theoretically grounded in the FCM structure and attention-weighted causal links, the paper primarily demonstrates interpretability through feature ablation and performance degradation analysis rather than explicit extraction of learned causal rules.

5.5.5 Selected Studies Using Other Neuro-Fuzzy Architectures

Tian et al. (2019) proposed an interpretable framework for detecting epileptic seizure from EEG signals by combining multi-view feature learning with a fuzzy rule-based classifier. Their method first constructed initial features from three complementary views including time domain, frequency domain (FFT), and time-frequency domain. Three separate CNNs then learnt deep features from each view and reduced dimensionality while preserving discriminative information. Finally, a Multi-View Takagi-Sugeno-Kang Fuzzy System (MV-TSK-FS) classified seizures using these deep

features. This fuzzy system provided interpretability through if-then rules that can be understood by clinicians. Experiments on the CHB-MIT dataset show the method achieves good performance while outperforming single-view approaches and traditional feature methods such as PCA. The fuzzy rule-based classifier added transparency to the DL pipeline, enabling explanation of decisions through human-readable rules.

To predict wind speed, especially with highly noisy and volatile signals, Shi et al. (2025) presented an Infinite-Dimensional Convolutional Neural Network (ICNN)-based fuzzy time-series framework. The pipeline consists of intelligent data repair, fuzzy time-series prediction, and rolling evaluation. Data repair combines fuzzy information granulation for missing-value handling with variational mode decomposition (VMD) for denoising, where VMD parameters are optimized via an improved artificial hummingbird algorithm to produce a repaired input signal. Prediction follows the fuzzy time-series paradigm with explicit IF-THEN rules, but rule matching is performed using an ICNN rather than heuristic matching. The ICNN uses continuous convolution kernels to better capture randomness and heteroscedasticity while remaining lightweight. Interpretability is preserved because the fuzzy rules remain explicit and inspectable; with the ICNN functioning only as a rule matcher rather than a black-box decision model. Although it was worth noting that the data repair stage entails black-box elements, the authors provide explanations for that part of the pipeline using SHAP.

Building on the standard interval type-2 fuzzy neural network (IT2FNN), Yao et al. (2025) proposed a self-organizing multi-output extension called SOIT2FNN-MO, for multi-step time-series forecasting. While conventional IT2FNNs effectively handle uncertainty, they are primarily designed for single-output prediction and do not explicitly model dependencies between prediction horizons. SOIT2FNN-MO addresses these limitations by introducing horizon-specific antecedent effects, improved firing-strength aggregation, separate consequents for each forecasting horizon, and explicit

links between successive prediction steps. The model automatically adapts its structure through rule growth and pruning, while maintaining interpretability through the analysis of fuzzy rule activations and their contributions across different prediction horizons.

Pseudo Outer-Product (POP) learning constructs fuzzy IF–THEN rules in a single pass by associating the most active input and output membership functions per sample, while Rough Set-Based POP (RSPOP) incorporates rough set reduction to eliminate redundant rule components and attributes. Traditional RSPOP is conducted offline, and it assumes the availability of a full and stationary dataset. [Das et al. \(2016\)](#) proposed an incremental extension of RSPOP for online time-series prediction under streaming data, enabling sample-by-sample learning without retraining. [MFs](#) are generated incrementally using discrete incremental clustering, allowing fuzzy partitions to evolve as new data arrives. To prevent fragmentation and loss of interpretability, a two-phase clustering strategy and membership-function merging are introduced. Rough set reduction is applied selectively to control rule-base growth, and ensemble learning is used to mitigate catastrophic forgetting by retaining pruned models as ensemble members. The result is an online, interpretable fuzzy predictor with explicit rules and robustness to nonstationarity.

HCFN (Hierarchically Collaborative Fuzzy Network) was introduced by [Sun et al. \(2025\)](#) for motor imagery [EEG](#) decoding. The pipeline has three stages. First, a multiscale spatiotemporal convolutional module extracts spatial (cross-channel) and temporal features from the raw [EEG](#). Second, an Adaptive Fuzzy Temporal Convolutional Network (AFTCN) replaces the static weights of standard causal convolutions with dynamically generated fuzzy kernels. The kernel weights are functions of the input state, computed via fuzzy membership functions. This means the convolutional filter adapts to the current [EEG](#) segment, making it more robust to the nonstationarity of [EEG](#) signals. The final part is a Fuzzy Attention-Guided [TSK](#) classifier, where each

rule has a linear consequent (IF antecedent THEN linear output), so the final prediction is a weighted sum of linear outputs, with the weights being the firing strengths of the fuzzy rules. The firing strengths are fed back into a temporal self-attention mechanism to guide which time steps the attention focuses on, so the fuzzy rules actively steer the attention as well as acting as a classifier. Interpretability is demonstrated in the form of the fuzzy rules being inspectable, with each rule’s antecedent defining a region of feature space, and its firing strength shows how much it activates for a given input. On another level, the fuzzy-guided attention weights show which time segments the model focused on for each prediction, and since the attention is steered by the fuzzy firing strengths, there is a direct link between the rules and the temporal focus.

[Azma et al. \(2026\)](#) proposed WPD-SSA-FAT, a hybrid DL model for multivariate water quality forecasting (dissolved oxygen, pH, water temperature) in the Taihu Lake Basin. The pipeline has three sequential stages. First, Wavelet Packet Decomposition (WPD) which is a signal processing technique that decomposes the raw time-series into multiple sub-signals at different frequency scales, separating short-term disturbances like rainfall pulses from longer seasonal cycles. The second part of the pipeline; Singular Spectrum Analysis (SSA) is a data-driven method that further denoises and reconstructs the WPD sub-signals by decomposing them into trend, oscillatory, and noise components and discarding the noise, producing cleaner inputs for the model. The third part is a Fuzzy Attention Transformer (FAT), a transformer where the standard attention mechanism is replaced with fuzzy membership functions, which assign soft degrees of relevance to each input rather than hard weights, making the model more robust to noisy environmental data and better at capturing nonlinear interactions between water quality variables. Inherent interpretability is demonstrated through transparent signal processing and fuzzy attention weights, which show how much each environmental variable contributes to predicting the target variable at each time step.

They take it a step further with post hoc sensitivity and uncertainty analyses that show which inputs the model is most responsive to.

For household energy load forecasting, [Mehrotra et al. \(2026\)](#) introduced CFNO (Causal Fuzzy Neural Operator), a four-stage pipeline. A causal reweighting layer first scales each input feature through a sigmoid-activated, L1-regularized weight vector, promoting sparsity so only the most relevant features retain high weights. Each feature is then mapped to seven Gaussian fuzzy sets with learnable centres and widths, expanding the feature space to capture uncertainty and partial membership in consumption patterns. A masked self-attention layer next enforces that predictions at time t only use information available up to t , before spectral neural operator blocks apply an [FFT](#), learn weights in the frequency domain, and inverse-transform the result to capture global long-range dependencies across all time steps simultaneously. Interpretability is demonstrated through the learned causal weights, which rank input features by their relevance to energy consumption, and through the fuzzy membership degrees, which show how each feature value maps to soft linguistic categories (e.g. "high temperature") at each time step.

For univariate time-series forecasting, an end-to-end interpretable model called LFIG-MHAM-FAR was developed by [Wu et al. \(2026\)](#). The model transforms raw time-series data into Linear Fuzzy Information Granules (LFIGs), which summarise local temporal trends and uncertainty, and then predicts future granules using fuzzy association rules. A multi-head attention mechanism is first used to identify meaningful temporal segments, after which the resulting granules are clustered and represented through IF-THEN fuzzy association rules. Interpretability is maintained throughout the pipeline, as the segmentation, granule representations, clustering structure, and fuzzy rules can all be directly inspected and related to the forecasting process.

Table 5: An Overview of Fuzzy and Neuro-Fuzzy Models for Time-Series

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Kajornrit et al. (2014)	Mamdani FNN	Genetic Algorithm, Mamdani FISs	Rainfall Time-series from northeast region of Thailand	Regression	Rainfall Prediction (or Environmental monitoring)	RMSE, MAE
Das et al. (2016)	ieRSPOP	Pseudo Outer-Product Algorithm, Ensemble Learning, Incremental Learning, FNN, Rough Set Theory	Singapore Highway Traffic Flow Dataset, Stock Data from Straits Time Index (STI) for 5 Stocks (CapitaMall, SMRT, SPH, SembCorp, Singtel), Citigroup, J.P. Morgan	Regression, Forecasting	Time-series prediction	RMSE, R^2 (Coefficient of Determination), MAE, MSE
Jiang et al. (2017)	TL-SSL-TSK Model	Transductive Transfer Learning, Semi-supervised Learning, TSK Fuzzy Systems	The Bonn EEG Dataset	Classification	Seizure Classification	Classification Accuracy
Tian et al. (2019)	MV-TSK-FS with CNN features	Multi-view features from time domain, FFT, time-frequency, 3 CNNs for deep feature extraction from each view, MV-TSK-FS	CHB-MIT scalp EEG dataset	Epileptic seizure detection	Clinical epilepsy monitoring (healthcare)	Accuracy, Sensitivity, Specificity, detection delay
Cao et al. (2020)	Multi-objective FRNN	EA, Fuzzy & Rough Set Theory	Stock Data from Yahoo Finance (0941.HK, 1288.HK, and 0005.HK)	Regression	Stock Prediction	HV

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Peng et al. (2020)	PS-FCM	Back Propagation, FCMs	Self-curated from Beijing environmental protection detection center & https://rp5.ru/	Classification, Haze Regression & Causality Min-ing/Modelling	Haze Formation & Prediction (or Environmental Monitoring)	Classification Probability Score & RMSE
Wang et al. (2021)	DFCM	DL, FCMs, LSTM	Air Quality Indexes (KDD Cup Dataset), Traffic Dataset of Beijing Road, Individual Household Electric Power Consumption, SML2010 (Temperature)	Regression	Time-series Prediction	MAE, RMSE
Lerner et al. (2021)	Hybrid Fuzzy, EA & NNs	EA, ANN & Fuzzy Set Theory	Appliances Energy Prediction (AEP) dataset	Data Synthesis, Regression	Energy Forecasting	Gini Index, Fuzzy Entropy Function, RMSE, MSE, MAE
Qin et al. (2023)	DAFCM	DL, Attention Mechanisms, Spatiotemporal FCMs, LSTM	Traffic Dataset of Beijing Road, Beijing Multi-Site Air Quality, Individual Household Electric Power Consumption, SML2010 (Temperature), StockNet, CWRU Bearing Data	Regression	Nonstationary Time-series Forecasting	MAE, RMSE
Jiang et al. (2025)	iFuzzyTL	Transfer Learning, FISs & Attention Mechanisms	12JFPM, Benchmark, eldBETA	Classification	SSVEP-based BCI	Accuracy, ITR

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Zhao et al. (2025)	Multi-objective DFRNN	DL, EA, Fuzzy & Rough Set Theory	Air quality dataset, Beijing PM2.5 dataset	Regression	Environmental monitoring	HV, Time
Shi et al. (2025)	ICNN-FTS	DL, Infinite-Dimensional CNN for Fuzzy Rule Learning, Fuzzy Time-series	Sotavento Wind Farm, Penglai Power Plant, NREL WIND Toolkit, Chinese State Grid Renewable Energy Generation Forecasting Competition	Regression, Forecasting	Wind Speed Prediction	RMSE, MAPE, ESD
Yao et al. (2025)	SOIT2FNN-MO	Self-Organizing Learning Mechanism, Fuzzy Neural Network	Simulated Mackey–Glass chaotic time-series, District Microgrid Dataset	Regression, Forecasting	Time-series prediction	RMSE, MPE
Sun et al. (2025)	HCFN	Multiscale spatiotemporal convolution + Adaptive Fuzzy Temporal Convolution + Fuzzy Attention-Guided TSK classification	BCI Competition IV-2a, BCI Competition IV-2b, OpenBMI EEG Dataset	Motor Imagery Decoding	BCI, clinical EEG applications, mental state monitoring, Healthcare	Accuracy, Cohen's kappa coefficient
Azma et al. (2026)	WPD-SSA-FAT	Wavelet Packet Decomposition, Singular Spectrum Analysis, Fuzzy Attention Transformer	Taihu Lake Physical and Chemical Surveillance Data, High Resolution Water Quality Dataset of Chinese Lakes and Reservoirs from 2000 to 2023	Regression, Time-series Forecasting	Water Quality Forecasting	RMSE, MAE, MAPE, R^2

Continued on next page

References	Model	Paradigm & Architecture	Dataset	Task	Application	Metrics
Mehrotra et al. (2026)	CFNO	Fuzzy Membership + Masked Self-attention + Spectral Neural Operator (FFT) + iFFT	UK Power Networks-led Low Carbon London project Energy Dataset	Regression, Time-series forecasting	Energy Forecasting	RMSE, R^2
Wu et al. (2026)	LFIG-MHAM-FAR	Multi-Head Attention + Conversion to Linear Fuzzy Information Granule + k-means + Fuzzy Association Rules	Chaotic Time-series, Monthly Sunspots, Daily Low Temperature (1), Daily Low Temperature (2), Commodity Exchange Values and International Diseases	Univariate Time-series Forecasting	All domains	MAE, MSE, RMSE

6 Datasets

This section provides an overview of the time-series datasets encountered in the studies included in this survey. Table 6 presents publicly available datasets, while Table 7 summarises proprietary and self-collected datasets used by the reviewed works. The datasets cover a wide range of application domains and learning tasks and are intended to provide a consolidated reference to the data sources commonly used in time-series ML research.

For the publicly available datasets shown in Table 6, the *Time-Series Type* column indicates the modality or source of the time-series data, *Dataset Name* identifies the dataset, *Task Type* specifies the ML task(s) performed, *Application* indicates the application domain, and *Data Source* column provides clickable links that give access to the dataset source. In Table 7 for the proprietary and self-collected datasets, the *Description* column summarises the dataset, including relevant collection details where

available, while the remaining columns indicate the associated task type, application domain, and the study in which the dataset was used.

To improve readability and facilitate comparison, datasets with similar characteristics, data types, or application domains are grouped together. The different groups are separated by horizontal lines within Table 6.

Table 6: Publicly Available Time-Series Datasets Used in Inherently Interpretable Studies

Time-Series Type	Dataset Name	Task Type	Application	Data Source
Hourly averaged chemical sensor responses	Air Quality (De Vito et al. 2008 ; Vito 2008)	Forecasting, Regression	Environmental monitoring	Link
Hourly PM2.5 sensor responses and hourly meteorological measurements	Beijing PM2.5 (Liang et al. 2015 ; Chen 2015)	Forecasting, Regression	Air quality prediction, public health assessment, environmental decision support	Link
Hourly air quality sensor responses and hourly meteorological measurements	Beijing Multi-Site Air Quality (Zhang et al. 2017 ; Chen 2017)	Regression	Climate and Environmental Monitoring	Link
Monitoring system weather sensor measurements	SML2010 (Romeu-Gualart and Zamora-Martinez 2014)	Regression	Indoor environmental monitoring and energy prediction	Link
Hourly air quality levels	Air Quality Indexes (KDD Cup Dataset)	Regression	Climate and Environmental Monitoring	Link
Univariate timeseries of weather measurements	Monthly Sunspots (Wu et al. 2026)	Regression, Time-series forecasting	Weather prediction, Climate and Environmental Monitoring	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
Univariate timeseries of weather measurements	Daily Low Temperature (1) (Wu et al. 2026)	Regression, Time-series forecasting	Weather prediction, Climate and Environmental Monitoring	Link
Univariate timeseries of weather measurements	Daily Low Temperature (2) (Wu et al. 2026)	Regression, Time-series forecasting	Weather prediction, Climate and Environmental Monitoring	Link
Appliance energy usage and weather sensor measurements	Appliances Energy Prediction Data Set (Candanedo et al. 2017)	Regression	Energy Consumption Forecasting	Link
Minute-level household energy consumption	Individual Household Electric Power Consumption (Hebrail and Berard 2006)	Regression, Clustering	Energy consumption prediction	Link
Wind farm, power generated, and meteorological data	Sotavento Wind Farm Data	Regression, Forecasting	Renewable Energy Modelling	Link
Wind farm, power generated, meteorological, and forecast data	NREL Wind Integration National Dataset (WIND) Toolkit (Draxl et al. 2015)	Regression	Renewable Energy Modelling	Link
Wind farm, solar station, power generated, and meteorological data	Solar and wind power data from the Chinese State Grid Renewable Energy Generation Forecasting Competition (Chen and Xu 2022)	Regression	Renewable Energy Modelling	Link 1 or Link 2
Hourly carbon intensity and microgrid load-price-generation data	District Microgrid Dataset (Yao et al. 2024)	Regression, Forecasting	Microgrid Monitoring	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
Hourly electricity consumption of 370 clients	Electricity Dataset (Trindade 2015)	Regression, Forecasting	Time-series and Electricity Load Forecasting	Link
Ecosystem fluxes	FLUXNET 2015 (Pastorello et al. 2020)	Regression	Ecological Modelling	Link
Daily meteorological observations (temperature, precipitation, wind, etc) from ground stations in the U.S.A	Global Surface Summary of the Day	Extreme value forecasting	Climate and Environmental Monitoring	Link 1 or Link 2
Daily and monthly temperature recordings in the U.K	U.K. Met Office Central England Temperature (CET) dataset	Temperature Forecasting	Climate and Environmental Monitoring	Link
Monthly observations of physicochemical parameters for water quality in Taihu Lake China	Taihu Lake Physical and Chemical Surveillance Data	Water Quality Forecasting	Climate and Environmental Monitoring	Link
Monthly estimates of physicochemical parameters for water quality for approximately 180,000 lakes and reservoirs in China	High Resolution Water Quality Dataset of Chinese Lakes and Reservoirs from 2000 to 2023 (Luan et al. 2025)	Water Quality Forecasting	Climate and Environmental Monitoring	Link
SSVEP-based EEG	12JFPM (Nakanishi et al. 2015)	Classification, Frequency Recognition	BCI	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
SSVEP-based EEG	Benchmark Database (Liu et al. 2020)	Classification, Frequency Recognition	BCI	Link
Elderly subjects SSVEP-based EEG	eldBETA (Liu et al. 2020)	Classification, Frequency Recognition	BCI	Link
Healthy and epileptic subjects' EEG	The Bonn EEG Dataset (Andrzejak et al. 2001)	Classification	Seizure detection and prediction, BCI	Link
Healthy and epileptic subjects' EEG	Bern-Barcelona EEG Database (Andrzejak et al. 2012)	Classification	Seizure detection and prediction, BCI	Link
Patients' EEG data, histories, and physicians' interpretations	The TUH EEG Corpus (TUEG) (Harati et al. 2014; Obeid and Picone 2016)	Classification	Epileptic Seizure detection and prediction, BCI	Link
Patients' EEG artifact data	The TUH EEG Artifact Corpus (TUAR) (Harati et al. 2014; Obeid and Picone 2016)	Classification	BCI, EEG Artifact Detection, Clinical EEG diagnosis support	Link
EEG Recordings	CHB-MIT scalp EEG dataset (Guttag 2010; Shoeb 2009; Goldberger et al. 2000)	Classification	Epileptic Seizure detection and prediction, BCI	Link
EEG and MRI Recordings of an epileptic patient	Brainstorm tutorial epilepsy data (Gramfort et al. 2014)	Source Image Reconstruction	Electrophysiological Sources Imaging, Epileptic Seizure detection and prediction, BCI	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
SEEG and hd- EEG recordings of drug-resistant epileptic patients	Simultaneous SEEG and High Density Scalp EEG (Parmigiani et al. 2022)	Source Image Reconstruction	Electrophysiological Sources Imaging, Epileptic Seizure detection and prediction, BCI	Link
EEG , fMRI , health and lifestyle data of middle-aged people at risk of dementia	PEARL-Neuro Database (Dzianok and Kublik 2024)	Classification of Alzheimer's genetic risk	Preclinical Alzheimer's screening	Link
EEG Recordings of executed motor movements	High-Gamma Dataset (Schirrmester et al. 2017)	Classification	BCI , Decoding executed movement type: left hand, right hand, feet, rest	Link
EEG Recordings of 4-class imagined motor movements	BCI Competition IV-2a (Brunner et al. 2008 ; Tangermann et al. 2012)	Classification	BCI , Decoding imagined movement type: left hand, right hand, feet, rest	Dataset #1 on page
EEG Recordings of 2-class motor imagery dataset	BCI Competition IV-2b (Leeb et al. 2008 ; Tangermann et al. 2012)	Classification	BCI , Decoding motor imagery	Dataset #4 on page
EEG Recordings of imagined motor movements	BNCI Horizon 2015-001 Motor Imagery Dataset (Faller et al. 2012)	Classification	BCI , Decoding imagined right hand movement vs. rest	Dataset #10 on page
EEG Recordings of imagined motor movements	OpenBMI EEG Dataset (Lee et al. 2019)	Classification	BCI , ERP, SSVEP decoding, Motor imagery: left hand vs. right hand	Link
EEG Recordings of imagined motor movements	Stieger2021 (Stieger et al. 2020, 2021)	Classification	BCI Motor Control	Link
EEG Recordings of imagined motor movements	EEG Motor Imagery Dataset (ETH Zurich) (Robinson et al. 2021)	Classification	BCI Motor Imagery	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
EEG recordings during cognitive workload tasks	Mental Workload Dataset (Hinss et al. 2021)	Classification	BCI Workload Estimation	Link
PolySomnoGraphic Sleep recordings, containing EEG , EOG , chin EMG , and event markers	Sleep-EDF Database Expanded (Kemp et al. 2000) (Goldberger et al. 2000)	Classification	Sleep Scoring	Link
EHRs	MIMIC-III Clinical dataset (Johnson et al. 2016 ; Goldberger et al. 2000)	Regression, Classification	Healthcare, Disease Risk Prediction	Link
EHRs	MIMIC-III Waveform dataset (Johnson et al. 2016 ; Goldberger et al. 2000 ; Moody et al. 2020)	Regression, Classification	Healthcare, Disease Risk Prediction	Link
EHRs	MIMIC-IV dataset (Johnson et al. 2024, 2023 ; Goldberger et al. 2000)	Regression, Classification	Healthcare, Disease Risk Prediction	Link
EHRs	N2C2-2014 dataset (Stubbs et al. 2015)	Regression, Classification	Healthcare, Disease Risk Prediction	Link
EHRs	eICU Collaborative Research Database (Pollard et al. 2018)	Regression, Classification	Healthcare, Disease Risk Prediction	Link
Monitoring signals from surgical patients	VitalDB (Lee et al. 2022)	Regression, Classification	Healthcare	Link
Face video recordings for remote HR estimation	VIPL-HR dataset (Niu et al. 2018, 2019)	Regression, Classification	Healthcare	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
Ground reaction forces signals during walking for gait classification	GaitRec (Horsak et al. 2020)	Classification, Gait Recognition	Rehabilitation and Healthcare	Link
ECG Recordings	CPSC2018 (Liu et al. 2018 ; Goldberger et al. 2000)	Regression, Classification	Healthcare - Arrhythmia detection	Link
ECG Recordings	MIT-BIH Arrhythmia (Moody and Mark 2001 ; Goldberger et al. 2000)	Regression, Classification	Healthcare - Arrhythmia detection	Link
ECG Recordings	PhysioNet / CINC challenge 2011 (Silva et al. 2011 ; Goldberger et al. 2000)	Regression, Classification	Healthcare - ECG Quality Assessment	Link
ECG Recordings	PhysioNet/CINC challenge 2017 (Clifford et al. 2017 ; Goldberger et al. 2000)	Regression, Classification	Healthcare - Atrial fibrillation detection	Link
ECG Recordings	Staff III (Martínez et al. 2017 ; Goldberger et al. 2000)	Regression, Classification	Healthcare - Myocardial Infarction Severity Prediction	Link
ECG Recordings	PTB Diagnostic ECG (Goldberger et al. 2000 ; Bousseljot et al. 1995)	Regression, Classification	Healthcare - Myocardial Infarction Severity Prediction	Link
ECG Recordings	PTB-XL (Wagner et al. 2020 ; Goldberger et al. 2000)	Regression, Classification	Healthcare - Myocardial Infarction Severity Prediction, Atrial fibrillation detection	Link
ECG Recordings	Chapman-Shaoxing ECG Database (Zheng et al. 2020a)	Classification	Healthcare - Arrhythmia study, Atrial fibrillation detection	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
ECG Recordings	Ningbo dataset (Zheng et al. 2020b)	Classification	Healthcare - Arrhythmia and Cardiac catheter ablation	Link
ECG Recordings	LUDB (Kalyakulina et al. 2020, 2021 ; Goldberger et al. 2000)	Regression, Classification	Healthcare - Arrhythmia detection	Link
Heart Sound Signals	Classification of Heart Sound Signal Using Multiple Features (Jamil et al. 2025 ; Yaseen et al. 2018)	Classification	Healthcare - Atrial Fibrillation	Link
Univariate timeseries on disease data	International Diseases (Wu et al. 2026)	Regression, Time-series forecasting	Public Health Data Modelling	Link
Protein amino acid sequences with functional annotations	UniProtKB	Regression, Classification	Sequence classification (e.g. protein function prediction, subcellular localization)	Link
Highway Traffic Datasets	PEMS03, PEMS04, PEMS07, PEMS08 (Chen et al. 2001)	Regression	Traffic Forecasting	Link
Traffic speeds from 6 road segments	Traffic Dataset of Beijing Road	Regression	Traffic and Transport Modelling	Link
Freeway hourly occupancy rate traffic dataset	Traffic dataset (PEMS-SF) (Cuturi 2011)	Regression	Time-series and Traffic Forecasting	Link
Traffic volume dataset of an on ramp near the Dodger's stadium	Dodgers Loop Sensor (Hutchins 2006 ; Ihler et al. 2006)	Regression, Classification	Time-series analysis and Traffic Forecasting	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
GPS trajectories of taxicab	Xi'an taxi dataset	Traffic flow prediction, Spatiotemporal forecasting, Regression	Time-series and Traffic Forecasting	Link
GPS trajectories of taxicab	Porto taxi dataset	Traffic flow prediction, Spatiotemporal forecasting, Regression	Time-series and Traffic Forecasting	Link
High-speed train operation records and weather data	High-speed railway network dataset (Zhang et al. 2022b)	Regression	Transport Systems Analysis	Link
Text sequence data - natural language user reviews with star ratings	Yelp Open Dataset	NLP, Image Analysis, Classification	Sentiment Analysis, and Prediction	Link
Text sequence data - natural language product reviews with star ratings	Amazon Product Reviews (Ni et al. 2019)	NLP, Classification	Sentiment Analysis and Prediction	Link
Historical stock prices	StockNet (Xu and Cohen 2018)	Regression	Stock movement and price prediction	Link
Standard financial market data, consisting of the daily closing prices of the S&P-500 index	S&P 500	Extreme value forecasting	Financial risk modelling	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
Univariate timeseries on trade data	Commodity Exchange Values (Wu et al. 2026)	Regression, Time-series forecasting	Financial Forecasting	Link
Motor bearing accelerometer readings	Case Western Reserve University (CWRU) Bearing Data	Regression	Fault detection and condition monitoring of machinery	Link
Vibration signals from bearings of different health conditions	Ottawa University Bearing Dataset (Huang and Baddour 2018)	Regression	Fault detection and condition monitoring of machinery	Link
Punch force measurements from a fine-blanking machine	Fine Blanking Punch Force Signals (Niemietz 2022)	Regression	Wear Assessment, Predictive Maintenance	Link
EIS measurements from lithium-ion batteries at different charge levels	12 $LiCoO_2$ /graphite cells (Zhang et al. 2020)	Regression	Battery Health Monitoring, Predictive Maintenance	Link
Simulated readings from 14 aircraft engine sensors	N-CMAPSS (Frederick et al. 2007)	Regression	Predicting the remaining useful life of plane engines	Link
Server records for anomaly detection from eBay	The Pooled Server Metrics (PSM) (Abdulaal et al. 2021)	Anomaly Detection	IT infrastructure monitoring	Link
Server records for anomaly detection from a large-scale internet service provider	The Server Machine Dataset (SMD) (Su et al. 2019)	Anomaly Detection	IT infrastructure monitoring	Link

Continued on next page

Time-Series Type	Dataset Name	Task Type	Application	Data Source
Charge/discharge cycle time-series (voltage, current, temperature, capacity, impedance) from Li-ion 18650 batteries from NASA	BatteryAgingARC-FY08Q4	Regression	Battery Health Monitoring, Remaining Useful Life Prediction	Link
Univariate timeseries	Chaotic time-series (Wu et al. 2026)	Regression, Time-series forecasting	Time-series Forecasting	Link

Table 7: Proprietary and Self-Collected Time-Series Datasets Used in Inherently Interpretable ML Studies

Description	Task Type	Application	Referenced In
Self-curated three-hourly meteorological time-series data for Beijing (2015/01/02–2017/03/10) obtained from the RP5 open weather archive https://rp5.ru/ . Variables include temperature (T), atmospheric pressure (P), relative humidity (Rh), and wind speed (Wp).	Forecasting, Regression	Climate and Environmental Monitoring	Peng et al. (2020)
Three-hourly air pollutant concentration time-series data for Beijing (2015/01/02–2017/03/10) collected from the Beijing Environmental Protection Detection Center. Includes PM2.5, SO ₂ , NO ₂ , O ₃ , and CO, along with haze indicators	Forecasting, Regression, Classification	Climate and Environmental Monitoring	Peng et al. (2020)
Eight monthly rainfall time-series collected from the study area, the northeast region of Thailand.	Forecasting, Regression	Climate and Environmental Monitoring	Kajornrit et al. (2014)
Penglai Power Plant Wind Data	Regression, Forecasting	Renewable Energy Generation (Wind)	Shi et al. (2025)

Continued on next page

Description	Task Type	Application	Referenced In
Net load, solar power generation, and meteorological data collected from 28 houses in the Western Victoria region of Australia. Data is stated to be confidential.	Regression, Forecasting	Microgrid Monitoring and Renewable Energy Generation (Solar)	Xu and Chen (2024)
Utility data from three medium-voltage feeders (E07L01, E07L05, and E07L12) from the XXV Aprile primary substation in Brescia. The authors state that data is available upon request Aghahadi et al. (2026) .	Classification and reconstruction of electrical load profiles	Energy Consumption/Load Forecasting	Aghahadi et al. (2026)
Self-collected environmental (meteorological measurements like indoor temperature), physiological (subjects' vital measurements like blood pressure), and subjective (the subjects' assessment of the current TSV) datasets on elderly people and their perception of indoor thermal environment sensations. It contains data from 14 elderly people (4 males and 10 females) in a pensioners' building. The authors state that data is available upon request.	Classification	Indoor Thermal Environment Modelling, HVAC Systems Control	Zheng et al. (2024)
Earth observation data from the municipality of Vila Velha and field-collected mosquito population measured weekly using 791 mosquito traps in Vila Velha city, Brazil, for 36 weeks in 2017, and 40 weeks in 2018 from the same region. The authors state that they do not have permission to share the data	Regression	Mosquito Population Modelling	Mudele et al. (2021)
Self-collected meteorological and dam deformation data from a hydropower station located on the southwest border of the Sichuan Province of China	Regression	Dam Deformation Monitoring, Climate and Environmental modelling	Zhu et al. (2024)
Self-collected vehicle trajectory data from a freeway tunnel located in Chongqing, China. The tunnel is a unidirectional, two-lane structure with a total length of 266 m. The authors state that data is available on request.	Regression, Forecasting	Traffic and Transport Modelling	Jin et al. (2025)

Continued on next page

Description	Task Type	Application	Referenced In
EHR data from 6367 patients with PUB disease, from the Endoscopy Center, Prince of Wales Hospital, Hong Kong, between 2007 and 2016. Data is available on request from the hospital.	Regression, Classification	Healthcare, Disease Risk Prediction	Tan et al. (2020)
Healthcare and Nursing Community Big Data Management Platform (HCBDMF) is an EHR database with a specific focus on elderly care. It contains data from 17 communities across six provinces in China, encompassing a population of approximately 2 million people, including 1.2 million seniors. Each entry includes a wide array of information, with over 600 variables. The authors state that they do not have permission to share data.	Regression, Classification	Healthcare, Disease Risk Prediction	Zhu et al. (2026)
Glucose (collected via Continuous Glucose Management (CGM) system) and exercise-related (collected via FitBit) measurements of 24 patients with Type 1 Diabetes from the Hospital Universitario Príncipe de Asturias, Madrid (Spain). The authors state that the data is available upon reasonable request.	Classification	Hypoglycemia prediction, Diabetes Management	De La Cruz et al. (2024)
Self-recorded videos for remote HR estimation. The authors state that data will be made available upon request.	Regression, Facial Detection	HR estimation, Remote patient monitoring, stress assessment, fitness tracking, driver monitoring	Liu et al. (2024a)
Preoperative ECGs of patients with PVCs were recorded and acquired at the National Cerebral and Cardiovascular Center (NCVC dataset). The authors state that the data is available upon request.	Multi-class Classification	Point-wise ECG segmentation, PVC origin localization, Preoperative planning for catheter ablation	Kujime et al. (2024)
ECG recordings from Mediplex Sejong Hospital (MSH) and Sejong General Hospital (SGH). The dataset is an internal dataset with no publicly available link.	Classification	Atrial fibrillation detection	Jo et al. (2021)

Continued on next page

Description	Task Type	Application	Referenced In
Raw heart sound recordings from 159 subjects, including healthy individuals and clinically diagnosed AF patients from First Medical Center of the Chinese People's Liberation Army General Hospital. Data is not available publicly because it contains sensitive patient information. An anonymised version is available upon request	Classification	Atrial fibrillation detection	Liu et al. (2026)
KidsBrainIT Database Lo et al. (2018) (Paediatric Brain Monitoring with Information Technology) contains minute-by-minute ICU monitoring signals (intracranial pressure, mean arterial blood pressure, cerebral perfusion pressure) from pediatric traumatic brain injury patients. There isn't a publicly downloadable link to the dataset. Access requires collaboration with the KidsBrainIT consortium directly.	Classification	Clinical event detection and outcome prediction	Haule et al. (2025)
Tetraplegic Pilot BCI Data: EEG recordings while a tetraplegic person mentally imagined 4 different movements to control a racing game avatar, recorded across 26 sessions over 14 months. Data not publicly available as it is a single case study on one known participant; data available on request from gernot.mueller@tugraz.at and results published in Hehenberger et al. (2021)	Classification	BCI Control	Kobler et al. (2022)
EEG from 88 adolescents with ASD and without ASD recorded during an emotion recognition task. The data was collected and preprocessed by the developers of SincNet. No mention is made of data availability.	Classification	Emotion Recognition and Prediction	Mayor-Torres et al. (2021)
EEG signals recorded by the Graz BCI group in the following works: Mondini et al. (2020) ; Martínez-Cagigal et al. (2020) . There isn't a statement of data availability	Regression	BCI Trajectory decoding	Borra et al. (2023)
IEDs in EEG Dataset. Data is available on request from authors Li and Zhang (2025b)	Classification	Epilepsy Detection and Prediction, Healthcare	Tang et al. (2025)

Continued on next page

Description	Task Type	Application	Referenced In
Self-collected Mean Skin Temperature Data from different parts of the body, collected under different exercise conditions. The authors state that the data is available upon request.	Regression, Classification	Body Temperature Prediction	Zhang et al. (2026)
Self-collected dataset with information about terrain patches bounded by natural drainage and ridgeline boundaries referred to as slope units, tracking a rainfall-triggered landslide event. Each slope unit is described by 18 features spanning terrain morphology, geology, land-use, soil type, rainfall characteristics, and landslide labels. Authors provide access to the data at: https://github.com/fcaleca1/landslide_susceptibility_EBM	Classification	Landslide Susceptibility modelling	Caleca et al. (2024)
A collection of time-series correlated to the financial market, including Nasdaq stock market price, trade-weighted U.S. dollar index (DTWEXB), and six other time-series obtained from https://fred.stlouisfed.org .	Regression	Time-series Forecasting	Munkhdalai et al. (2022)
Vehicle Fault Logs are described as sequences of fault codes logged by vehicle onboard diagnostics (OBD). Each fault is described with a five-digit Diagnostic Trouble Code (DTC), which is standard across different makes and models.	Multi-class risk classification	Vehicle Fault Risk Level Prediction	Ming et al. (2019)
SNU Bearing Dataset is an experimental dataset collected from a specialized rolling element bearing (REB) testbed for vibration-based machinery fault diagnosis. Data is available upon request from authors Park et al. (2023)	Regression	Fault detection and condition monitoring of machinery	Kim et al. (2026)
A brushless direct current motor (BLDCM) bearing dataset collected using an NI data acquisition card. Data is available upon request from the authors Lu et al. (2026)	Classification	Fault detection and condition monitoring of machinery	Lu et al. (2026)
Dataset from a Switched Reluctance Motor (SRM) test rig Song et al. (2024)	Classification	Fault detection and condition monitoring of machinery	Lu et al. (2026)

Continued on next page

Description	Task Type	Application	Referenced In
Multivariate time-series from cooling, electrical, and IT systems of the INFN CNAF Tier-1 data centre, sampled every 10-15 minutes from January 2022 to July 2023. Data is not publicly available.	Anomaly Detection	IT infrastructure monitoring	Ronchieri et al. (2024)

Tables 6 and 7 provide a comprehensive resource and reference of the standard datasets used in time-series modelling. They also reveal notable patterns in certain aspects. Firstly, in terms of application domain, healthcare datasets dominate, including EEG, ECG, EHRs, and other medical data. EEG datasets in particular are used far more than any other single data type, appearing in models across almost every model family, with the most concentration in the interpretable neural network models, where BCI and EEG-based studies make up a large share of the reviewed studies.

This is also reflected in the learning tasks, with a clear split between healthcare and non-healthcare domains. EEG datasets are used almost exclusively for classification tasks such as seizure detection, motor imagery decoding, or artifact detection, with a few exceptions. ECG and EHR datasets, while also commonly used for classification in the selected studies, are frequently suited to regression as well. In contrast, datasets such as air quality, weather, energy consumption, and electricity generation data are overwhelmingly used for regression and forecasting tasks, reflecting their prevalence in applications such as renewable energy modelling, energy consumption forecasting, and weather quality monitoring. Datasets in traffic and transport modelling, financial data, and machinery and industrial data, though fewer in number, are similarly predominantly used for regression and forecasting with a few instances of anomaly detection. This pattern is broadly consistent across both Tables 6 and 7, indicating that task type is closely tied to application domain rather than any other factor.

This distinction is worth noting, as publicly available datasets substantially outweigh proprietary and self-collected ones, with 82 and 30 datasets identified, respectively. This suggests that the majority of reviewed studies rely on established, accessible data sources rather than collecting their own data, which in turn contributes to the reuse of certain datasets across studies. Among these, the following emerge as the most frequently reused benchmarks in their respective applications:

- MIT-BIH Arrhythmia Database and PhysioNet / CINC Challenge Datasets for arrhythmia classification
- The Bonn EEG Dataset and BCI Competition IV-Series Dataset for seizure classification and BCI applications
- MIMIC-III and MIMIC-IV Clinical Databases both EHRs primarily for mortality and clinical risk prediction tasks
- Air Quality and Beijing PM2.5 for climate and environmental modelling
- The PEMS03, 04, 07 and 08 datasets for traffic forecasting and
- The CWRU Bearing Dataset as the most used industrial dataset

7 Discussion

This review presents inherently interpretable time-series models in different application domains. To obtain the reviewed studies, we followed the methodology outlined in Section 3, starting out with 568 records.

To provide further context on the scope and nature of these results, bibliometric metadata was extracted from the WoS export. The direct export of metadata was cleaned, merged and standardised on publisher names. Similarly, publishers with fewer than 5 records were grouped into an "Others" category for clarity in the visualisations. The distribution of records across publishers is shown in Fig. 16, with notable publishers like Elsevier dominating the corpus. The spread of publication years is shown in Fig. 18, showing a clear growth trend in the field over recent years. The research

areas the initial query records belong to, as categorised on WoS, are shown in Fig. 17, with areas having fewer than 20 records grouped as "Others". It should be noted that WoS research area categories are not mutually exclusive, and individual publications are sometimes assigned to multiple categories; therefore, the category counts should not be interpreted as disjoint or summing to the total number of unique publications. This is reflected directly in Fig. 17. Each bar is segmented by the number of WoS categories the contributing papers belong to, so the coloured segments reveal what proportion of papers in a given area are exclusively classified there versus shared across two, three, or four or more areas simultaneously.

A similar bibliometric analysis was conducted on the final 77 papers included in the final review after initial title/abstract screening, full-text assessment, and backward reference search. The distributions are shown in Figures 19, 20, and 21.

As seen in the figures of the distributions presented, the breadth across publishers and research domains suggests the query was not inadvertently biased toward a particular outlet or discipline while ensuring that only relevant and qualitative works were included.

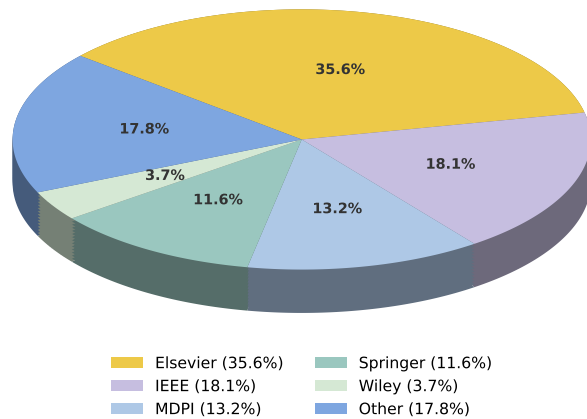


Fig. 16: Distribution of Publishers in Initial Results

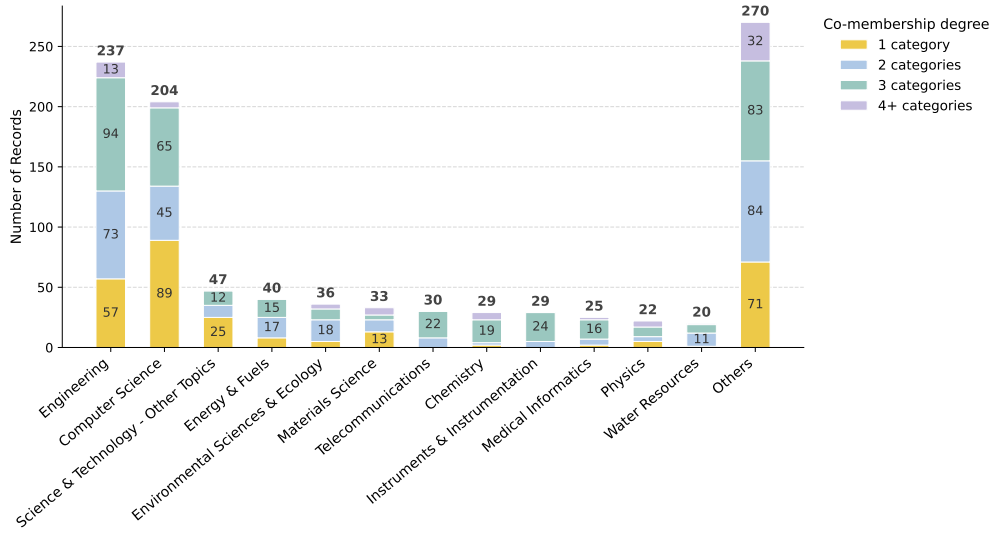


Fig. 17: Distribution of Research Areas in Initial Results

In addition to publisher and research area distributions, a distribution of the class of inherently interpretable models presented in this survey is shown in Fig. 22, as this was the central characterisation that was employed in this survey.

As shown in Fig. 22, Fuzzy and Neuro-Fuzzy models take the biggest share at 22.1%, closely followed by Attention-Based models at 20.8%. This distribution makes a lot of sense because time-series data is naturally noisy and filled with uncertainty.

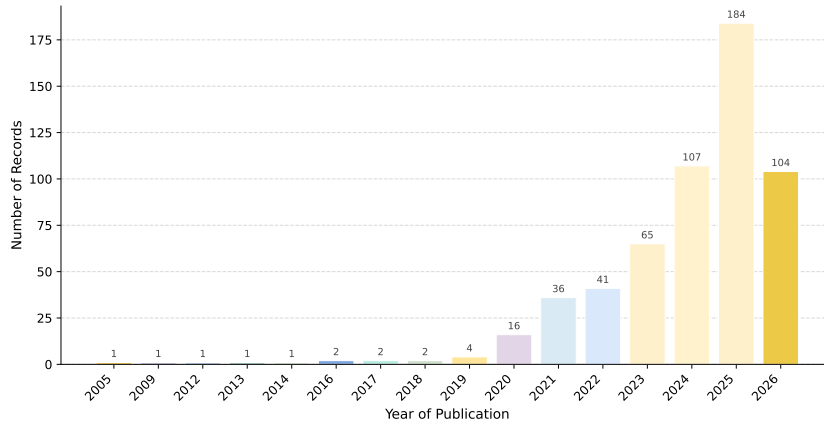


Fig. 18: Distribution of Publications by Years in Initial Results

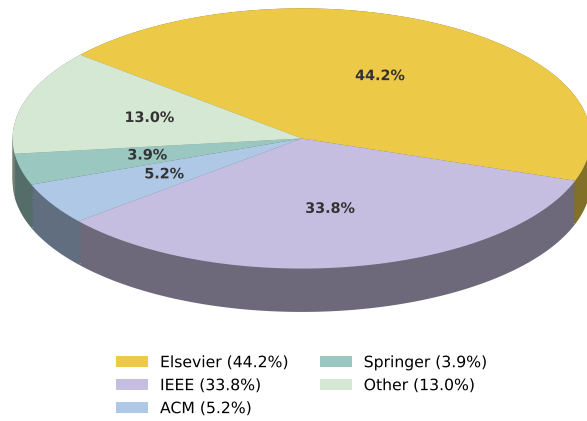


Fig. 19: Distribution of Publishers in Final Review

Fuzzy systems, with their intuitive "If-Then" rules, offer high inherent interpretability, while attention mechanisms are excellent at highlighting exactly which time steps or features matter most. Interestingly, other DL categories like Interpretable CNNs (6.5%), GNNs (6.5%), and Other Interpretable Neural Architectures (18.2%) also hold a massive, combined share. Looking at recent papers from 2025 and 2026, this spread

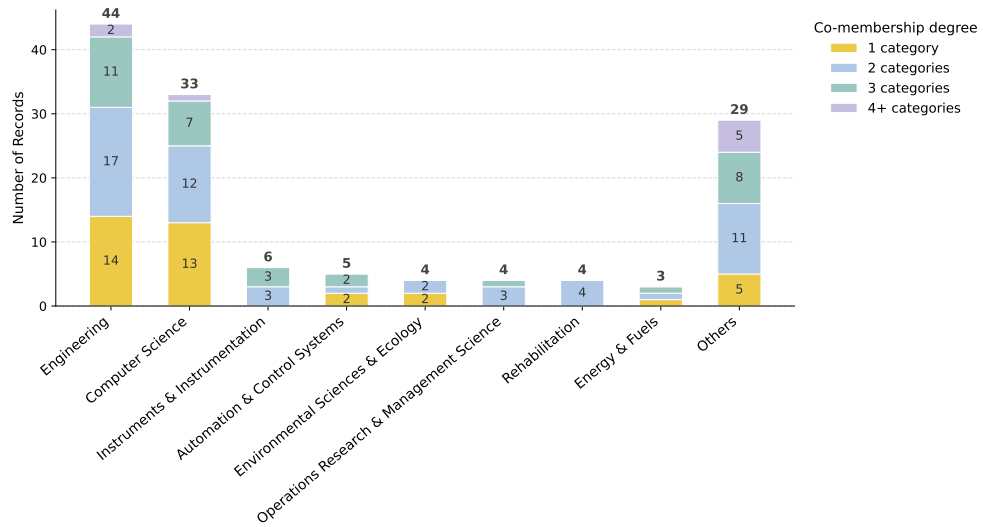


Fig. 20: Distribution of Research Areas in Final Review

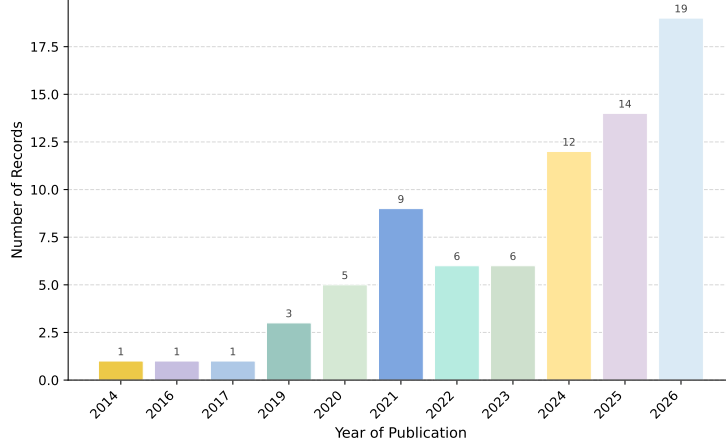


Fig. 21: Distribution of Publications by Years in Final Review

highlights a major trend: instead of just using pure fuzzy rules or classic linear models (11.7%), the research community is moving fast toward hybrid architectures. By combining the solid reasoning of fuzzy logic with the feature-learning power of attention and deep networks, researchers are successfully building models that do not sacrifice accuracy for the sake of transparency.

The systematic review of 77 selected eligible papers shows that inherently interpretable ML models for time-series are mainly used in four important areas:

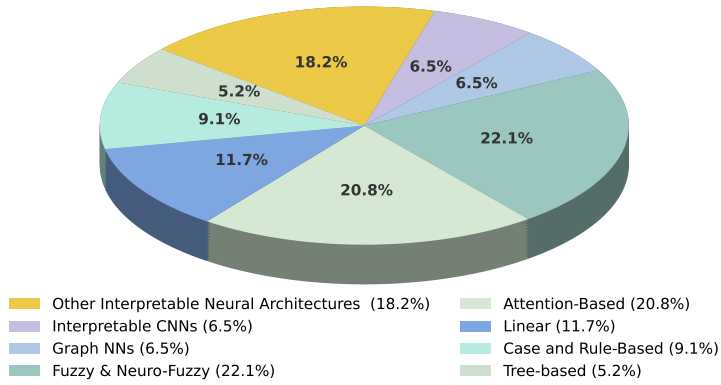


Fig. 22: Distribution of Final Model Categories in Review

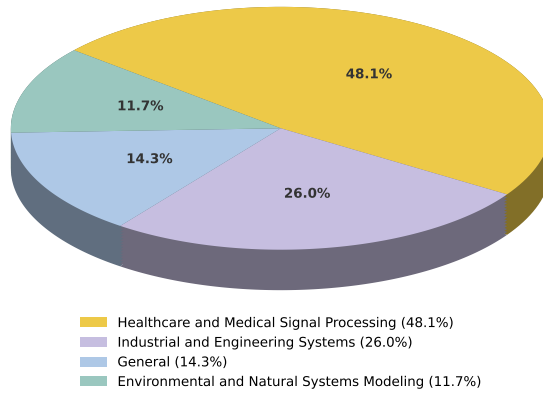


Fig. 23: Distribution of Application Areas of Models in Review

- **Healthcare and Medical Signal Processing:** A large number of studies focus on analysing medical signals like [EEG](#) and [ECG](#). In this field, interpretability is very helpful. Doctors cannot trust a "black-box" model because a wrong decision can risk a patient's life. Inherently interpretable models help by showing the exact medical rules behind each diagnosis.
- **Industrial and Engineering Systems:** Another key application is modelling complex and non-linear physical systems, such as water flow dynamics, hydraulic power, and energy consumption. These systems generate highly noisy time-series data. Using hybrid models helps engineers predict system behaviours while understanding how different physical variables affect the final output.
- **Environmental and Natural Systems modelling:** Some of the models are designed to model natural phenomena, environmental processes, and accomplish tasks such as rainfall prediction, air quality monitoring, and landslide susceptibility assessment. The decisions undertaken based on the results of these models have a huge impact on our environment and climate. Therefore, interpretability is vital in understanding the factors influencing the predictions produced.
- **General-Purpose Time-Series Modelling:** 14.3% of the studies develop inherently interpretable models that are not tied to a specific application domain but

are designed for general time-series modelling, prediction and forecasting tasks. These models are quite advantageous due to their global applicability. These models could also offer non-domain experts the opportunity to apply these models to their particular use case.

Fig. 23 shows the distribution of these applications.

When looking at the ML tasks, a large number of the studied papers focus on time-series forecasting and regression problems rather than classification. This strong preference makes perfect sense because most real-world time-series applications such as predicting future water flow dynamics, forecasting energy consumption, or tracking shifting financial trends are continuous by nature and require predicting exact future values over time. While classification tasks are still popular in medical fields for decoding specific signal patterns like EEG or ECG segments, where an event needs to be detected and classified as either a negative or a positive occurrence, the main drive in XAI research is to understand how a model calculates its future predictions. By making these regression and forecasting models inherently interpretable, researchers can see which past time steps or features had the biggest impact on the final numbers, making the entire forecasting process trustworthy for engineers and analysts alike.

8 Conclusion, Limitations and Future Work

This systematic literature review provided comprehensive inherently interpretable ML models designed specifically for time-series. By shifting the focus from post-hoc explanations to inherent interpretability, this work establishes a clear foundational landscape for developing inherently transparent predictive systems in safety-critical domains. To achieve this, we systematically identified and synthesised 77 eligible studies following the PRISMA framework. We carefully analysed the utilised datasets across various time-series modalities, learning tasks, and application domains, uncovering key trends and data sources in current research. Furthermore, we clarified

the conceptual distinctions between interpretability and explainability in temporal contexts and proposed a comprehensive architectural taxonomy that spans from traditional rule-based frameworks to advanced DL systems. The reviewed inherently interpretable models are essential as they are increasingly deployed in areas that impact human lives and society at large. Regardless, several limitations were observed across the reviewed studies. The noted challenges and the potential future directions are listed below:

- **Varying degrees of interpretability provided:** The level of interpretability provided by the reviewed studies varies significantly. It ranges from studies that present entirely transparent models that are easy to understand even to non-experts in the field, to several that require a substantial knowledge of AI, physics, or mathematics. This hinders practitioners in the field from more easily adapting these methods to their needs and specific use cases.
- **Transparent components combining to form complex models:** Although a lot of the models are built on components that are fundamentally inherently interpretable; when these components are used to build up the model, in combination with a lot of other components; or in some cases just a lot more of the same component; a lot of the interpretability is lost as the model becomes increasingly complex. Resulting in a model whose individual parts are interpretable in isolation but not as a whole.
- **Trade-off between model complexity and interpretability:** As architectures become more complex, they can capture more complicated temporal patterns and, by so doing, achieve higher accuracy and better predictive performance. However, this increase in components and parameters introduces more factors to consider when evaluating and interpreting a model's structure and predictions. Developing methods that are not overly complex without sacrificing accuracy would be a promising direction for future research to focus on.

- **Lack of consistent evaluation criteria for interpretability:** Due to the burgeoning nature of the field, many studies use traditional metrics to evaluate model performance and accuracy. While the quality of the explanations is usually analysed and assessed with very subjective criteria. An attempt to provide a unified or universally acceptable framework for assessing the interpretability of a model would make it easier to determine how much impact these models are having on user trust.
- **Focus of models on specific application domains:** A majority of the reviewed studies, 85.7%, focus on specific application domains. This hinders transfer learning, domain adaptation, and makes it difficult to ascertain how effectively the proposed models would generalize to other domains. Future work could explore more generalizable inherently interpretable time-series forecasting and predicting architectures for ease of application to diverse domains.

Overall, the field is moving towards models that attempt to balance accuracy, transparency, and usability. Inherently interpretable time-series models represent an important step towards developing ML systems that are not only accurate but also understandable and trustworthy in real-world applications. And, achieving models that are both highly performant and genuinely understandable remains an open research challenge, particularly for complex time-dependent systems.

Acknowledgements. This study is supported by the Crafoord Foundation (Grant No. 2025-0618).

Declarations

- **Funding:** This study is supported by the Crafoord Foundation (Grant No. 2025-0618).
- **Competing interests:** The authors have no competing interests to declare that are relevant to the content of this article.
- **Ethics approval and consent to participate:** Not applicable

- Consent for publication: Not applicable
- Data availability: No datasets were generated in this study, and all datasets mentioned with publicly available links are provided
- Materials availability: Not applicable
- Code availability: Not applicable
- Author contribution: R.N. and A.A. jointly identified the research gap and conceptualized the study. R.N. contributed to the methodology, investigation, data curation, formal analysis, visualization of the bibliometric analysis, and preparation of the original draft. Z.A.V. contributed to the methodology, investigation, data curation, formal analysis, preparation of the original draft, visualization of the architectural components, and review and editing of the manuscript. R.M. contributed to validation and review and editing of the manuscript. A.A. contributed to the methodology, funding acquisition, project administration, supervision, validation, and review and editing of the manuscript. All authors reviewed and approved the final manuscript.

References

- Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access* **6**, 52138–52160 (2018)
- Aghahadi, M., Bosisio, A., Dacco, E., Falabretti, D., Ruffini, A., Cirocco, A., Berizzi, A.: Multi-layer machine learning approach for bottom-up classification and reconstruction of electrical load profiles in smart distribution networks. *International Journal of Electrical Power & Energy Systems* **176**, 111691 (2026)
- Apostolidis, A., Bouriquet, N., Stamoulis, K.P.: Ai-based exhaust gas temperature prediction for trustworthy safety-critical applications. *Aerospace* **9**(11), 722 (2022)
- Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado,

- A., García, S., Gil-López, S., Molina, D., Benjamins, R., *et al.*: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion* **58**, 82–115 (2020)
- Abdulaal, A., Liu, Z., Lancewicki, T.: Practical approach to asynchronous multivariate time series anomaly detection and localization. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 2485–2494 (2021)
- Andrzejak, R.G., Lehnertz, K., Mormann, F., Rieke, C., David, P., Elger, C.E.: Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E* **64**(6), 061907 (2001)
- Azma, A., Liu, Y., Rezania, S.: A hybrid wpd–ssa machine learning transformer for modeling nonlinear dynamics of flow systems. *Journal of Water Process Engineering* **82**, 109448 (2026)
- Aamodt, A., Plaza, E.: Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications* **7**(1), 39–59 (1994)
- Arik, S.Ö., Pfister, T.: Tabnet: Attentive interpretable tabular learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 6679–6687 (2021)
- Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., Atkinson, P.M.: Explainable artificial intelligence: an analytical review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **11**(5), 1424 (2021)
- Andrzejak, R.G., Schindler, K., Rummel, C.: Nonrandomness, nonlinear dependence, and nonstationarity of electroencephalographic recordings from epilepsy patients.

- Physical Review E—Statistical, Nonlinear, and Soft Matter Physics **86**(4), 046206 (2012)
- Arsenault, P.-D., Wang, S., Patenaude, J.-M.: A survey of explainable artificial intelligence (xai) in financial time series forecasting. *ACM Computing Surveys* **57**(10), 1–37 (2025)
- Barić, D., Fumić, P., Horvatić, D., Lipic, T.: Benchmarking attention-based interpretability of deep learning in multivariate time series predictions. *Entropy* **23**(2), 143 (2021)
- Borra, D., Fantozzi, S., Magosso, E.: Interpretable and lightweight convolutional neural network for eeg decoding: Application to movement execution and imagination. *Neural Networks* **129**, 55–74 (2020)
- Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M.: *Time Series Analysis: Forecasting and Control*. John Wiley & Sons, Hoboken, NJ (2015)
- Bousseljot, R., Kreiseler, D., Schnabel, A.: *Nutzung der ekg-signaldatenbank cardiostat der ptb über das internet* (1995)
- Brunner, C., Leeb, R., Müller-Putz, G., Schlögl, A., Pfurtscheller, G.: Bci competition 2008–graz data set a. Institute for knowledge discovery (laboratory of brain-computer interfaces), Graz University of Technology **16**(1-6), 1 (2008)
- Borra, D., Mondini, V., Magosso, E., Müller-Putz, G.R.: Decoding movement kinematics from eeg using an interpretable convolutional neural network. *Computers in Biology and Medicine* **165**, 107323 (2023)
- Becker, M., Niemietz, P., Bergs, T.: Explainable neural network for time series-based condition monitoring in sheet metal shearing. *Journal of Intelligent Manufacturing*,

- Breiman, L.: Random forests. *Machine learning* **45**(1), 5–32 (2001)
- Buchanan, B.G., Shortliffe, E.H.: Rule Based Expert Systems: the Mycin Experiments of the Stanford Heuristic Programming Project (the Addison-Wesley Series in Artificial Intelligence). Addison-Wesley Longman Publishing Co., Inc., Boston, MA (1984)
- Cord, O., *et al.*: Genetic Fuzzy Systems: Evolutionary Tuning and Learning of Fuzzy Knowledge Bases vol. 19. World Scientific, Singapore (2001)
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* **16**, 321–357 (2002)
- Caleca, F., Confuorto, P., Raspini, F., Segoni, S., Tofani, V., Casagli, N., Moretti, S.: Shifting from traditional landslide occurrence modeling to scenario estimation with a “glass-box” machine learning. *Science of the total environment* **950**, 175277 (2024)
- Candanedo, L.M., Feldheim, V., Deramaix, D.: Data driven prediction models of energy use of appliances in a low-energy house. *Energy and buildings* **140**, 81–97 (2017)
- Chen, S.: Beijing PM2.5. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5JS49> (2015)
- Chen, S.: Beijing Multi-Site Air Quality. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5RK5G> (2017)

- Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1251–1258 (2017)
- Clifford, G.D., Liu, C., Moody, B., Lehman, L.-w.H., Silva, I., Li, Q., Johnson, A.E., Mark, R.G.: Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In: 2017 Computing in Cardiology (CinC), pp. 1–4 (2017). IEEE
- Correia, G.M., Niculae, V., Martins, A.F.: Adaptively sparse transformers. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2174–2184 (2019)
- Chen, C., Petty, K., Skabardonis, A., Varaiya, P., Jia, Z.: Freeway performance measurement system: mining loop detector data. Transportation research record **1748**(1), 96–102 (2001)
- Cuturi, M.: PEMS-SF. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C52G70> (2011)
- Chen, Y., Xu, J.: Solar and wind power data from the chinese state grid renewable energy generation forecasting competition. Scientific data **9**(1), 577 (2022)
- Cao, B., Zhao, J., Lv, Z., Gu, Y., Yang, P., Halgamuge, S.K.: Multiobjective evolution of fuzzy rough neural network via distributed parallelism for stock prediction. IEEE Transactions on Fuzzy Systems **28**(5), 939–952 (2020) <https://doi.org/10.1109/TFUZZ.2020.2972207>
- Das, R.T., Ang, K.K., Quek, C.: ierspop: A novel incremental rough set-based pseudo outer-product with ensemble learning. Applied Soft Computing **46**, 170–186 (2016)

- Draxl, C., Clifton, A., Hodge, B.-M., McCaa, J.: The wind integration national dataset (wind) toolkit. *Applied Energy* **151**, 355–366 (2015)
- Deng, Y., Galib, A.H., Tan, P.-N., Luo, L.: Unraveling block maxima forecasting models with counterfactual explanation. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 562–573 (2024)
- Dzianok, P., Kublik, E.: Pearl-neuro database: Eeg, fmri, health and lifestyle data of middle-aged people at risk of dementia. *Scientific Data* **11**(1), 276 (2024)
- De La Cruz, M., Garnica, O., Cervigon, C., Velasco, J.M., Hidalgo, J.I.: Explainable hypoglycemia prediction models through dynamic structured grammatical evolution. *Scientific Reports* **14**(1), 12591 (2024)
- Dong, A., Starr, A., Zhao, Y.: An interpretable temporal convolutional framework for granger causality analysis. *IEEE/CAA Journal of Automatica Sinica* (2025)
- De Vito, S., Massera, E., Piga, M., Martinotto, L., Di Francia, G.: On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical* **129**(2), 750–757 (2008)
- Frederick, D.K., DeCastro, J.A., Litt, J.S.: User’s guide for the commercial modular aero-propulsion system simulation (c-mapss). Technical report, NASA (2007)
- Felzmann, H., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A.: Towards transparency by design for artificial intelligence. *Science and engineering ethics* **26**(6), 3333 (2020)
- Faller, J., Vidaurre, C., Solis-Escalante, T., Neuper, C., Scherer, R.: Autocalibration and recurrent adaptation: Towards a plug and play online ERD-BCI. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **20**(3), 313–319 (2012)

<https://doi.org/10.1109/TNSRE.2012.2189584>

- Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation* **101**(23), 215–220 (2000)
- Gu, X., Angelov, P., Han, J., Shen, Q.: Multilayer evolving fuzzy neural networks. *IEEE transactions on fuzzy systems* **31**(12), 4158–4169 (2023)
- Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), pp. 80–89 (2018). IEEE
- Graziani, M., Dutkiewicz, L., Calvaresi, D., Amorim, J.P., Yordanova, K., Vered, M., Nair, R., Abreu, P.H., Blanke, T., Pulignano, V., *et al.*: A global taxonomy of interpretable ai: unifying the terminology for the technical and social sciences. *Artificial intelligence review* **56**(4), 3473–3504 (2023)
- Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., Hämäläinen, M.S.: Mne software for processing meg and eeg data. *neuroimage* **86**, 446–460 (2014)
- Glykas, M.: *Fuzzy Cognitive Maps: Advances in Theory, Methodologies, Tools and Applications* vol. 247. Springer, Berlin, Heidelberg (2010)
- Golding, A.R., Rosenbloom, P.S.: Improving accuracy by combining rule-based and case-based reasoning. *Artificial Intelligence* **87**(1-2), 215–254 (1996)
- Guttag, J.: CHB-MIT Scalp EEG Database. PhysioNet (2010) <https://doi.org/10.1>

- Hastie, T.: The elements of statistical learning: data mining, inference, and prediction. springer (2009)
- Hastie, T.J.: Generalized additive models. Statistical models in S, 249–307 (2017)
- Hebrail, G., Berard, A.: Individual Household Electric Power Consumption. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C58K54> (2006)
- Huang, H., Baddour, N.: Bearing vibration data collected under time-varying rotational speed conditions. Data in brief **21**, 1745–1749 (2018)
- Hu, Z., Bodyanskiy, Y.V., Tyshchenko, O.K.: A deep cascade neural network based on extended neo-fuzzy neurons and its adaptive learning algorithm. In: 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON), pp. 801–805 (2017). IEEE
- Hinss, M.F., Darmet, L., Somon, B., Jahanpour, E., Lotte, F., Ladouce, S., Roy, R.N.: An EEG Dataset for Cross-session Mental Workload Estimation: Passive BCI Competition of the Neuroergonomics Conference 2021. <https://doi.org/10.5281/zenodo.4917217> . <https://doi.org/10.5281/zenodo.4917217>
- Hosmer Jr, D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression. John Wiley & Sons, Hoboken, NJ, USA (2013)
- Hehenberger, L., Kobler, R.J., Lopes-Dias, C., Srisrisawang, N., Tumfart, P., Uroko, J.B., Torke, P.R., Müller-Putz, G.R.: Long-term mutual training for the cybathlon bci race with a tetraplegic pilot: a case study on inter-session transfer and intra-session adaptation. Frontiers in Human Neuroscience **15**, 635777 (2021)
- Harati, A., Lopez, S., Obeid, I., Picone, J., Jacobson, M., Tobochnik, S.: The tuh eeg

- corpus: A big data resource for automated eeg interpretation. In: 2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–5 (2014). IEEE
- Hatzilygeroudis, I., Prentzas, J.: Integrating (rules, neural networks) and cases for knowledge representation and reasoning in expert systems. *Expert Systems with Applications* **27**(1), 63–75 (2004)
- Haule, H., Piper, I., Jones, P., MillyLo, T.-Y., Escudero, J.: Clai: Collaborative learning and inference for low-resolution physiological signals–validation in clinical event detection and prediction. *IEEE Transactions on Biomedical Engineering* (2025)
- Hwang, E., Park, H.-S., Kim, H.-S., Kim, J.-Y., Jeong, H., Kim, J., Kim, S.-H.: Development of a bispectral index score prediction model based on an interpretable deep learning algorithm. *Artificial intelligence in medicine* **143**, 102569 (2023)
- Hall, T., Rasheed, K.: A survey of machine learning methods for time series prediction. *Applied Sciences* **15**(11), 5957 (2025)
- Horsak, B., Slijepcevic, D., Raberger, A.-M., Schwab, C., Worisch, M., Zeppelzauer, M.: Gaitrec, a large-scale ground reaction force dataset of healthy and impaired gait. *Scientific data* **7**(1), 143 (2020)
- Hastie, T., Tibshirani, R.: Generalized additive models. *Statistical science* **1**(3), 297–310 (1986)
- Hastie, T., Tibshirani, R., Wainwright, M.: Statistical learning with sparsity. *Mono-graphs on statistics and applied probability* **143**(143), 8 (2015)
- Hutchins, J.: Dodgers Loop Sensor. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C51P50> (2006)
- Islam, M.R., Ahmed, M.U., Barua, S., Begum, S.: A systematic review of explainable

- artificial intelligence in terms of different application domains and tasks. *Applied Sciences* **12**(3), 1353 (2022)
- Ihler, A., Hutchins, J., Smyth, P.: Adaptive event detection with time-varying poisson processes. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 207–216 (2006)
- Ishibuchi, H., Nojima, Y.: Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning. *International Journal of Approximate Reasoning* **44**(1), 4–31 (2007)
- Ito, A., Tanaka, T.: Sleepsatelightftc: A lightweight and interpretable deep learning model for single-channel eeg-based sleep stage classification. *IEEE Access* (2025)
- Jamil, S., *et al.*: Facegest: A comprehensive facial gesture dataset for human-computer interaction. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 337–347 (2025)
- James, G.: *An introduction to statistical learning with applications in R*. Springer (2013)
- Jang, J.-S.: Anfis: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics* **23**(3), 665–685 (1993)
- Johnson, A., Bulgarelli, L., Pollard, T., Gow, B., Moody, B., Horng, S., Celi, L.A., Mark, R.: MIMIC-IV. *PhysioNet* (2024) <https://doi.org/10.13026/kpb9-mt58> . Version 3.1
- Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., *et al.*: Mimic-iv, a freely accessible electronic health record dataset. *Scientific data* **10**(1), 1 (2023)

- Jo, Y.-Y., Cho, Y., Lee, S.Y., Kwon, J.-m., Kim, K.-H., Jeon, K.-H., Cho, S., Park, J., Oh, B.-H.: Explainable artificial intelligence to detect atrial fibrillation using electrocardiogram. *International journal of cardiology* **328**, 104–110 (2021)
- Jiang, X., Cao, B., Ou, L., Chang, Y.-C., Do, T., Lin, C.-T.: ifuzzytl: Interpretable fuzzy transfer learning for steady-state visual evoked potentials brain–computer interfaces system. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **55**(12), 9495–9508 (2025) <https://doi.org/10.1109/TSMC.2025.3614244>
- Jiang, Z., Chung, F.-L., Wang, S.: Recognition of multiclass epileptic eeg signals based on knowledge and label space inductive transfer. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **27**(4), 630–642 (2019)
- Jeyarani, A.: An interpretable temporal transformer for 12-lead ecg arrhythmia classification. *International Journal of Machine Learning and Cybernetics* **17**(1), 14 (2026)
- Jovic, A., Frid, N., Brkic, K., Cifrek, M.: Interpretability and accuracy of machine learning algorithms for biomedical time series analysis—a scoping review. *Biomedical Signal Processing and Control* **110**, 108153 (2025)
- Jin, Y., Liu, J., Liu, Y., Qin, C., Li, Z., Xiao, D., Zhao, L., Liu, C.: A novel interpretable method based on dual-level attentional deep neural network for actual multilabel arrhythmia detection. *IEEE Transactions on Instrumentation and Measurement* **71**, 1–11 (2021)
- Jin, Y., Li, Z., Qin, C., Liu, J., Liu, Y., Zhao, L., Liu, C.: A novel attentional deep neural network-based assessment method for ecg quality. *Biomedical Signal Processing and Control* **79**, 104064 (2023)
- Jin, J., Li, J., Tian, S., Ye, Q.: Zone-specific real-time traffic conflict risk modeling

- for freeway tunnels: a crosstabnet approach. *Accident Analysis & Prevention* **223**, 108274 (2025)
- Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. *Scientific data* **3**(1), 1–9 (2016)
- Jiang, Y., Wu, D., Deng, Z., Qian, P., Wang, J., Wang, G., Chung, F.-L., Choi, K.-S., Wang, S.: Seizure classification from eeg signals using transfer learning, semi-supervised learning and tsf fuzzy system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **25**(12), 2270–2284 (2017) <https://doi.org/10.1109/TNSRE.2017.2748388>
- Ji, J., Wang, J., Jiang, Z., Ma, J., Zhang, H.: Interpretable spatiotemporal deep learning model for traffic flow prediction based on potential energy fields. In: 2020 IEEE International Conference on Data Mining (ICDM), pp. 1076–1081 (2020). IEEE
- Jiao, M., Xian, X., Wang, B., Zhang, Y., Yang, S., Chen, S., Sun, H., Liu, F.: Xdl-esi: Electrophysiological sources imaging via explainable deep learning framework with validation on simultaneous eeg and ieeg. *Neuroimage* **299**, 120802 (2024)
- Kobler, R., Hirayama, J.-i., Zhao, Q., Kawanabe, M.: Spd domain-specific batch normalization to crack interpretable unsupervised domain adaptation in eeg. *Advances in Neural Information Processing Systems* **35**, 6219–6235 (2022)
- Kuhn, M., Johnson, K., *et al.*: *Applied Predictive Modeling* vol. 26. Springer, New York, NY (2013)
- Kim, S., Lee, S., Lee, J., Kim, M., Yoon, H., Youn, B.D.: Multi-scale signal transformer with signal processing-based attention interpretation for fault diagnosis of rotating machinery under variable speed conditions. *Reliability Engineering & System Safety*,

112243 (2026)

Kolodner, J.L.: An introduction to case-based reasoning. *Artificial intelligence review* **6**(1), 3–34 (1992)

Kosko, B.: Fuzzy cognitive maps. *International journal of man-machine studies* **24**(1), 65–75 (1986)

Krishnan, M.: Against interpretability: a critical examination of the interpretability problem in machine learning. *Philosophy & Technology* **33**(3), 487–502 (2020)

Kim, B., Rudin, C., Shah, J.: The bayesian case model: A generative approach for case-based reasoning and prototype classification. *Advances in neural information processing systems* **27** (2014)

Kujime, K., Seno, H., Nakajima, K., Yamazaki, M., Sakuma, I., Yamagata, K., Kusano, K., Tomii, N.: Explainable localization of premature ventricular contraction using deep learning-based semantic segmentation of 12-lead electrocardiogram. *Journal of Arrhythmia* **40**(4), 948–957 (2024)

Kajornrit, J., Wong, K.W., Fung, C.C., Ong, Y.S.: An integrated intelligent technique for monthly rainfall time series prediction. In: 2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1632–1639 (2014). <https://doi.org/10.1109/FUZZ-IEEE.2014.6891619>

Kalyakulina, A.I., Yusipov, I.I., Moskalenko, V.A., Nikolskiy, A.V., Kosonogov, K.A., Osipov, G.V., Zolotikh, N.Y., Ivanchenko, M.V.: Ludb: a new open-access validation tool for electrocardiogram delineation algorithms. *IEEE access* **8**, 186181–186190 (2020)

- Kalyakulina, A., Yusipov, I., Moskalenko, V., Nikolskiy, A., Kosonogov, K., Zolotykh, N., Ivanchenko, M.: Lobachevsky University Electrocardiography Database. PhysioNet (2021) <https://doi.org/10.13026/eegm-h675> . Version 1.0.1
- Kemp, B., Zwinderman, A.H., Tuk, B., Kamphuisen, H.A., Obery, J.J.: Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering* **47**(9), 1185–1194 (2000)
- Larsen, K.: Gam: the predictive modeling silver bullet. *Multithreaded. Stitch Fix* **30**, 1–27 (2015)
- LaValley, M.P.: Logistic regression. *Circulation* **117**(18), 2395–2399 (2008)
- Leeb, R., Brunner, C., Müller-Putz, G., Schlögl, A., Pfurtscheller, G.: Bci competition 2008–graz data set b. *Graz University of Technology, Austria* **16**, 1–6 (2008)
- Lou, Y., Caruana, R., Gehrke, J., Hooker, G.: Accurate intelligible models with pairwise interactions. In: *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 623–631 (2013)
- Lei, Y., Deng, B., Wang, Z.: Climate temporal temperature prediction via an interpretable kolmogorov-arnold neural network. *IEEE Transactions on Instrumentation and Measurement* (2025)
- Liu, Y., Guo, X., Zhang, Y.: Lightweight and interpretable convolutional neural network for real-time heart rate monitoring using low-cost video camera under realistic conditions. *Biomedical Signal Processing and Control* **87**, 105461 (2024)
- Larsson, S., Heintz, F.: Transparency in artificial intelligence. *Internet policy review* **9**(2), 1–16 (2020)
- Liu, B., Huang, X., Wang, Y., Chen, X., Gao, X.: Beta: A large benchmark database

- toward ssvep-bci application. *Frontiers in neuroscience* **14**, 627 (2020)
- Lee, M.-H., Kwon, O.-Y., Kim, Y.-J., Kim, H.-K., Lee, Y.-E., Williamson, J., Fazli, S., Lee, S.-W.: Eeg dataset and openbmi toolbox for three bci paradigms: An investigation into bci illiteracy. *GigaScience* **8**(5), 002 (2019) <https://doi.org/10.5524/100542>
- Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing sax: a novel symbolic representation of time series. *Data Mining and knowledge discovery* **15**(2), 107–144 (2007)
- Li, O., Liu, H., Chen, C., Rudin, C.: Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018)
- Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., *et al.*: An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics* **8**(7), 1368–1373 (2018)
- Le Nguyen, T., Gsponer, S., Ilie, I., O’reilly, M., Ifrim, G.: Interpretable time series classification using linear models and multi-resolution multi-domain symbolic representations: T. le nguyen *et al.* *Data mining and knowledge discovery* **33**(4), 1183–1222 (2019)
- Lo, T., Piper, I., Depreitere, B., Meyfroidt, G., Poca, M., Sahuquillo, J., Durduran, T., Enblad, P., Nilsson, P., Ragauskas, A., *et al.*: Kidsbrainit: a new multi-centre, multi-disciplinary, multi-national paediatric brain monitoring collaboration. In: *Intracranial Pressure & Neuromonitoring XVI*, pp. 39–45. Springer, Cham, Switzerland (2018)

- Luan, S., Pan, H., Shen, R., Xia, X., Duan, H., Yuan, W., Wei, J.: High resolution water quality dataset of chinese lakes and reservoirs from 2000 to 2023. *Scientific Data* **12**(1), 572 (2025)
- Lee, H.-C., Park, Y., Yoon, S.B., Yang, S.M., Park, D., Jung, C.-W.: Vitaldb, a high-fidelity multi-parameter vital signs database in surgical patients. *Scientific Data* **9**(1), 279 (2022)
- Lerner, M., Reich, C., Abdeslam, D.O.: Hybrid ai improves energy forecasts by combining fuzzy rules, evolutionary strategies and neural networks. In: *IECON 2021 – 47th Annual Conference of the IEEE Industrial Electronics Society*, pp. 1–6 (2021). <https://doi.org/10.1109/IECON48115.2021.9589186>
- Li, Z., Wang, H., Li, L.: Task-related eeg as a biomarker for preclinical alzheimer’s disease: An explainable deep learning approach. *Biomimetics* **10**(7), 468 (2025)
- Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y., Tegmark, M.: Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756* (2024)
- Lou, S., Yang, C., Liu, Z., Zhang, H., Liu, C., Wu, P.: From shallow to deep: A novel correlation network representation regression framework for modeling and monitoring miq-driven blast furnace ironmaking processes. *IEEE/CAA Journal of Automatica Sinica* **13**(2), 281–299 (2026)
- Li, C., Zhang, L.: Data-driven system identification, model transfer, and forecasting of battery capacity with impedance measurements. *Mechanical Systems and Signal Processing* **237**, 113041 (2025)
- Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., Chen, S.X.:

- Assessing beijing’s pm2. 5 pollution: severity, weather impact, apec and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **471**(2182), 20150257 (2015)
- Li, Q., Zhu, Y., Shangguan, W., Wang, X., Li, L., Yu, F.: An attention-aware lstm model for soil moisture and soil temperature prediction. *Geoderma* **409**, 115651 (2022)
- Liu, J., Zhao, F., Yang, H., Chen, Y., Wang, W.: Detecting atrial fibrillation from heart sounds using dual-branch feature fusion and kolmogorov–arnold networks. *Applied Soft Computing*, 114895 (2026)
- Lu, S., Zhu, J., Zheng, L., Song, J., Hu, Z., Wang, X., Wang, D.: Real-time classification and fault diagnosis of machine signals via a novel decision-tree-inspired analog circuit classifier. *Mechanical Systems and Signal Processing* **249**, 114074 (2026)
- Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies* **7**(1), 1–13 (1975)
- Martins, A., Astudillo, R.: From softmax to sparsemax: A sparse model of attention and multi-label classification. In: *International Conference on Machine Learning*, pp. 1614–1623 (2016). PMLR
- McCullagh, P.: *Generalized Linear Models*. Routledge, New York, NY (2019)
- Martínez-Cagigal, V., Kobler, R.J., Mondini, V., Hornero, R., Muller-Putz, G.R.: Non-linear online low-frequency eeg decoding of arm movements during a pursuit tracking task. In: *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 2981–2985 (2020). IEEE
- Mudele, O., Frery, A.C., Zanandrez, L.F., Eiras, A.E., Gamba, P.: Modeling dengue

- vector population with earth observation data and a generalized linear model. *Acta Tropica* **215**, 105809 (2021)
- Montenegro, A., González, F.A., Franco, H.: Protokdm: An interpretable prototype-based framework for biomedical time-series analysis. *Biomedical Signal Processing and Control* **116**, 109473 (2026)
- Mrowca, A., Gyrock, F., Günnemann, S.: Temporal state change bayesian networks for modeling of evolving multivariate state sequences: model, structure discovery and parameter estimation. *Data Mining and Knowledge Discovery* **36**(1), 240–294 (2022)
- Mendel, J.M., John, R.B.: Type-2 fuzzy sets made simple. *IEEE Transactions on fuzzy systems* **10**(2), 117–127 (2002)
- Mondini, V., Kobler, R.J., Sburlea, A.I., Müller-Putz, G.R.: Continuous low-frequency eeg decoding of arm movement for closed-loop, natural control of a robotic arm. *Journal of neural engineering* **17**(4), 046031 (2020)
- Mueen, A., Keogh, E., Young, N.: Logical-shapelets: an expressive primitive for time series classification. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1154–1162 (2011)
- Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G.: Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *Bmj* **339** (2009)
- Moody, G.B., Mark, R.G.: The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine* **20**(3), 45–50 (2001)
- Munkhdalai, L., Munkhdalai, T., Pham, V.-H., Li, M., Ryu, K.H., Theera-Umpon, N.: Recurrent neural network-augmented locally adaptive interpretable regression

- for multivariate time-series forecasting. *IEEE Access* **10**, 11871–11885 (2022)
- Moody, B., Moody, G., Villarroel, M., Clifford, G., Silva, I.: Mimic-iii waveform database (version 1.0). *PhysioNet* **3** (2020)
- Molnar, C.: *Interpretable Machine Learning*. Lulu. com, Durham, NC (2020)
- Martínez, J.P., Pahlm, O., Ringborn, M., Warren, S., Laguna, P., Sörnmo, L.: The staff iii database: Ecg recorded during acutely induced myocardial ischemia. In: *2017 Computing in Cardiology (CinC)*, pp. 1–4 (2017). IEEE
- Mehrotra, A., Pavithra, S., Singh, I.: Causal-fuzzy neural operators for smart energy forecasting: A multi context learning framework for grid aware load prediction. *Energy Reports* **15**, 108946 (2026)
- Montgomery, D.C., Peck, E.A., Vining, G.G.: *Introduction to Linear Regression Analysis*, 6e Solutions Manual. John Wiley & Sons, ??? (2022)
- Mayor-Torres, J.M., Ravanelli, M., Medina-DeVilliers, S.E., Lerner, M.D., Riccardi, G.: Interpretable sincnet-based deep learning for emotion recognition from eeg brain activity. In: *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 412–415 (2021). IEEE
- Minh, D., Wang, H.X., Li, Y.F., Nguyen, T.N.: Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review* **55**(5), 3503–3568 (2022)
- Ming, Y., Xu, P., Qu, H., Ren, L.: Interpretable and steerable sequence learning via prototypes. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 903–913 (2019)
- Nelson, M., Alotaibi, Y., Rajendran, S., Jibinsingh, B.R.: Multimodal learning for neonatal sepsis detection in the nicu using brain-inspired graph neural networks.

- Ni, J., Chen, Z., Cheng, W., Zong, B., Song, D., Liu, Y., Zhang, X., Chen, H.: Interpreting convolutional sequence model by learning local prototypes with adaptation regularization. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management, pp. 1366–1375 (2021)
- Niu, X., Han, H., Shan, S., Chen, X.: Vipl-hr: A multi-modal database for pulse estimation from less-constrained face video. In: Asian Conference on Computer Vision, pp. 562–576 (2018). Springer
- Niemietz, P.: Series of Time Series Representing Fine-blanking Punch Force Strokes with Wear Assessment. <https://doi.org/10.7910/DVN/OYNDZO> . <https://doi.org/10.7910/DVN/OYNDZO>
- Nori, H., Jenkins, S., Koch, P., Caruana, R.: Interpretml: A unified framework for machine learning interpretability. arXiv preprint arXiv:1909.09223 (2019)
- Ni, J., Li, J., McAuley, J.: Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 188–197 (2019)
- Niu, X., Shan, S., Han, H., Chen, X.: Rhythmnet: End-to-end heart rate estimation from face via spatial-temporal representation. IEEE Transactions on Image Processing **29**, 2409–2423 (2019)
- Nakanishi, M., Wang, Y., Wang, Y.-T., Jung, T.-P.: A comparison study of canonical correlation analysis based methods for detecting steady-state visual evoked potentials. PloS one **10**(10), 0140703 (2015)

- Niu, S., Yin, Q., Ma, J., Song, Y., Xu, Y., Bai, L., Pan, W., Yang, X.: Enhancing healthcare decision support through explainable ai models for risk prediction. *Decision Support Systems* **181**, 114228 (2024)
- Niu, S., Yin, Q., Song, Y., Guo, Y., Yang, X.: Label dependent attention model for disease risk prediction using multimodal electronic health records. In: 2021 IEEE International Conference on Data Mining (ICDM), pp. 449–458 (2021). <https://doi.org/10.1109/ICDM51629.2021.00056>
- Obeid, I., Picone, J.: The temple university hospital eeg data corpus. *Frontiers in neuroscience* **10**, 196 (2016)
- Pawlak, Z.: Rough sets. *International journal of computer & information sciences* **11**(5), 341–356 (1982)
- Prabhakararao, E., Dandapat, S.: Myocardial infarction severity stages classification from ecg signals using attentional recurrent neural network. *IEEE Sensors Journal* **20**(15), 8711–8720 (2020)
- Pedrycz, W., Gomide, F.: *An Introduction to Fuzzy Sets: Analysis and Design*. MIT press, Cambridge, MA (1998)
- Pollard, T.J., Johnson, A.E., Raffa, J.D., Celi, L.A., Mark, R.G., Badawi, O.: The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data* **5**(1), 180178 (2018)
- Peng, Z., Liu, W., An, S.: Haze pollution causality mining and prediction based on multi-dimensional time series with ps-fcm. *Information sciences* **523**, 307–317 (2020)
- Page, M.J., McKenzie, J.E., Bossuyt, P.M., Boutron, I., Hoffmann, T.C., Mulrow,

- C.D., Shamseer, L., Tetzlaff, J.M., Akl, E.A., Brennan, S.E., et al.: The prisma 2020 statement: an updated guideline for reporting systematic reviews. *bmj* **372** (2021)
- Parmigiani, S., Mikulan, E., Russo, S., Sarasso, S., Zauli, F., Rubino, A., Cattani, A., Fecchio, M., Giampiccolo, D., Lanzone, J., *et al.*: Simultaneous stereo-eeG and high-density scalp eeg recordings to study the effects of intracerebral stimulation parameters. *Brain Stimulation* **15**(3), 664–675 (2022)
- Peters, B., Niculae, V., Martins, A.F.: Sparse sequence-to-sequence models. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1504–1519 (2019)
- Prakash, A.J., Patro, K.K., Saunak, S., Sasmal, P., Kumari, P.L., Geetamma, T.: A new approach of transparent and explainable artificial intelligence technique for patient-specific eeg beat classification. *IEEE Sensors Letters* **7**(5), 1–4 (2023)
- Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., *et al.*: The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific data* **7**(1), 225 (2020)
- Park, Y., Tong, A., Lee, S., Seong, J., Xie, Q., Choi, J.: Towards transparent time series analysis: Exploring methods and enhancing interpretability. *ACM Computing Surveys* **58**(9), 1–36 (2026)
- Park, J., Yoo, J., Kim, T., Ha, J.M., Youn, B.D.: Multi-head de-noising autoencoder-based multi-task model for fault diagnosis of rolling element bearings under various speed conditions. *Journal of Computational Design and Engineering* **10**(4), 1804–1820 (2023)
- Qin, D., Peng, Z., Wu, L.: Deep attention fuzzy cognitive maps for interpretable

- multivariate time series prediction. *Knowledge-Based Systems* **275**, 110700 (2023)
- Quinlan, J.R.: Induction of decision trees. *Machine learning* **1**(1), 81–106 (1986)
- Quinlan, J.R.: Combining instance-based and model-based learning. In: *Proceedings of the Tenth International Conference on Machine Learning*, pp. 236–243 (1993)
- Ravanelli, M., Bengio, Y.: Speaker recognition from raw waveform with sincnet. In: *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1021–1028 (2018). IEEE
- Robinson, N., Chouhan, T., Mihelj, E., Kratka, P., Debraine, F., Wenderoth, N., Guan, C., Lehner, R.: Design considerations for long term non-invasive brain computer interface training with tetraplegic cybathlon pilot. *Frontiers in human neuroscience* **15**, 648275 (2021)
- Ronchieri, E., Giommi, L., Scarponi, L.B., Torzi, L., Costantini, A., Duma, D.C., Salomoni, D.: Anomaly detection in data center it & physical infrastructure. In: *EPJ Web of Conferences*, vol. 295, p. 07004 (2024). EDP Sciences
- Romeu-Guallart, P., Zamora-Martinez, F.: SML2010. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C5RS3S> (2014)
- Räuker, T., Ho, A., Casper, S., Hadfield-Menell, D.: Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. In: *2023 IEEE Conference on Secure and Trustworthy Machine Learning (satml)*, pp. 464–483 (2023). IEEE
- Rojat, T., Puget, R., Filliat, D., Del Ser, J., Gelin, R., Díaz-Rodríguez, N.: Explainable artificial intelligence (xai) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950* (2021)

- Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence* **1**(5), 206–215 (2019)
- Shalaeva, V., Alkhoury, S., Marinescu, J., Amblard, C., Bisson, G.: Multi-operator decision trees for explainable time-series classification. In: *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pp. 86–99 (2018). Springer
- Sahbani, S., Bejaoui, B., Ottaviani, E., Benabdallah, S., Riccomagno, E., Prampolini, E., Canu, D.M., Missaoui, H., Solidoro, C.: Performance of different modeling techniques in testing the impact of environmental variables on eel landing in ichkeul lake, a ramsar wetland and unesco biosphere reserve. *Regional Studies in Marine Science* **77**, 103587 (2024)
- Stieger, J.R., Engel, S.A., He, B.: Continuous sensorimotor rhythm based brain computer interface learning in a large population. *Scientific Data* **8**(1), 98 (2021) <https://doi.org/10.1038/s41597-021-00883-1>
- Stieger, J., Engel, S., Jiang, H., Cline, C., Kreitzer, M., He, B.: Mindfulness improves brain-computer interface performance by increasing control over neural activity in the alpha band. *Cerebral Cortex* **31**(1), 426–438 (2020) <https://doi.org/10.1093/cercor/bhaa234> . Publisher Copyright: © 2020 The Author(s) 2020. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.
- Scarselli, F., Gori, M., Tsoi, A.C., Hagenbuchner, M., Monfardini, G.: The graph neural network model. *IEEE transactions on neural networks* **20**(1), 61–80 (2008)
- Shoeb, A.H.: Application of machine learning to epileptic seizure onset detection and

- treatment. PhD thesis, Massachusetts Institute of Technology (2009)
- Song, J., Han, H., Wu, X., Lu, J., Wang, X., Lu, S.: Multisource deep feature fusion of optimized symmetrized dot patterns for srm fault diagnosis. *IEEE Transactions on Instrumentation and Measurement* **73**, 1–13 (2024)
- Stubbs, A., Kotfila, C., Uzuner, Ö.: Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/uthealth shared task track 1. *Journal of biomedical informatics* **58**, 11–19 (2015)
- Senin, P., Malinchik, S.: Sax-vsm: Interpretable time series classification using sax and vector space model. In: 2013 IEEE 13th International Conference on Data Mining, pp. 1175–1180 (2013). IEEE
- Silva, I., Moody, G.B., Celi, L.: Improving the quality of ecgs collected using mobile phones: The physionet/computing in cardiology challenge 2011. In: 2011 Computing in Cardiology, pp. 273–276 (2011). IEEE
- Sahbani, S., Ottaviani, E., Missaoui, H., Béjaoui, B., Riccomagno, E., Benabdallah, S., Prampolini, E.: Machine learning techniques for forecasting the effect of climate change and anthropogenic pressures on coastal wetlands (ichkeul lake, ramsar site). In: *Innovation and Technological Advances for Sustainability*, pp. 88–96. CRC Press, ??? (2024)
- Schirrmeister, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping* **38**(11), 5391–5420 (2017) <https://doi.org/10.1002/hbm.23730>
- Shi, X., Wang, J., Li, Z., Zhang, W.: Efficient interpretable wind speed prediction system. *Energy*, 138102 (2025)

- Samek, W., Wiegand, T., Müller, K.-R.: Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296 (2017)
- Sun, Y., Yao, J., Chen, T., Xia, K., Jiang, Y., Qian, P.: Adaptive fuzzy-convolution and task-guided attention for interpretable eeg mi decoding. *IEEE Transactions on Fuzzy Systems* (2025)
- Su, Y., Zhao, Y., Niu, C., Liu, R., Sun, W., Pei, D.: Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2828–2837 (2019)
- She, X., Zhao, X., Yang, H., Cui, X.: Development and temporal validation of a nomogram for predicting icu 28-day mortality in middle-aged and elderly sepsis patients: An icu database study. *Plos one* **20**(7), 0328701 (2025)
- Tang, D., Donnelly, J., Barnett, A.J., Semenova, L., Jing, J., Hadar, P., Karakis, I., Selioutski, O., Zhao, K., Westover, M.B., *et al.*: This eeg looks like these eegs: Interpretable interictal epileptiform discharge detection with protoeeg-knn. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 615–625 (2025). Springer
- Tian, X., Deng, Z., Ying, W., Choi, K.-S., Wu, D., Qin, B., Wang, J., Shen, H., Wang, S.: Deep multi-view feature learning for eeg-based epileptic seizure detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **27**(10), 1962–1972 (2019)
- Tangemann, M., Müller, K.-R., Aertsen, A., Birbaumer, N., Braun, C., Brunner, C., Leeb, R., Mehring, C., Müller, K.J., Müller-Putz, G.R., Nolte, G., Pfurtscheller,

- G., Preissl, H., Schalk, G., Schlögl, A., Vidaurre, C., Waldert, S., Blankertz, B.: Review of the BCI competition IV. *Frontiers in Neuroscience* **6**, 55 (2012) <https://doi.org/10.3389/fnins.2012.00055>
- Trindade, A.: ElectricityLoadDiagrams20112014. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C58C86> (2015)
- Takagi, T., Sugeno, M.: Fuzzy identification of systems and its applications to modeling and control. *IEEE transactions on systems, man, and cybernetics* (1), 116–132 (1985)
- Tutek, M., Šnajder, J.: Toward practical usage of the attention mechanism as a tool for interpretability. *IEEE access* **10**, 47011–47030 (2022)
- Theissler, A., Spinnato, F., Schlegel, U., Guidotti, R.: Explainable ai for time series classification: a review, taxonomy and research directions. *Ieee Access* **10**, 100700–100724 (2022)
- Tan, Q., Ye, M., Ma, A.J., Yang, B., Yip, T.C.-F., Wong, G.L.-H., Yuen, P.C.: Explainable uncertainty-aware convolutional recurrent neural network for irregular medical time series. *IEEE Transactions on Neural Networks and Learning Systems* **32**(10), 4665–4679 (2020)
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. *arXiv preprint arXiv:1710.10903* (2017)
- Vito, S.: Air Quality. UCI Machine Learning Repository. DOI: <https://doi.org/10.24432/C59K5F> (2008)
- Vilone, G., Longo, L.: Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093* (2020)

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
- Wijesena, S., Pradhan, B.: Enhancing weather index insurance through surrogate models: leveraging machine learning techniques and remote sensing data. *Environmental Research Communications* **7**(4), 045009 (2025)
- Wang, J., Peng, Z., Wang, X., Li, C., Wu, J.: Deep fuzzy cognitive maps for interpretable multivariate time series prediction. *IEEE Transactions on Fuzzy Systems* **29**(9), 2647–2660 (2021) <https://doi.org/10.1109/TFUZZ.2020.3005293>
- Wagner, P., Strodthoff, N., Bousseljot, R.-D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data* **7**(1), 154 (2020)
- Wang, C., Ye, S., Lei, X., An, D., Xi, H.: Battery joint state estimation with uncertainty based on feature independent representation and re-weighted. *Measurement* **265**, 120409 (2026)
- Wu, X., Zhan, J., Ding, W., Qian, Y.: A linear fuzzy granulation framework leveraging multi-head attention and self-calibrating k-means for precise short-horizon time-series forecasting. *Fuzzy Sets and Systems*, 109899 (2026)
- Xu, Y., Cohen, S.B.: Stock movement prediction from tweets and historical prices. In: Gurevych, I., Miyao, Y. (eds.) *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1970–1979. Association for Computational Linguistics, Melbourne, Australia (2018). <https://doi.org/10.18653/v1/P18-1183> . <https://aclanthology.org/P18-1183/>
- Xu, C., Chen, G.: Interpretable transformer-based model for probabilistic short-term

- forecasting of residential net load. *International Journal of Electrical Power & Energy Systems* **155**, 109515 (2024)
- Xin, Q., Hu, S., Liu, S., Zhao, L., Wang, S.: Wtrpnet: an explainable graph feature convolutional neural network for epileptic eeg classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* **17**(3s), 1–18 (2021)
- Xiaolin, L., Qingyuan, W., Panicker, R.C., Cardiff, B., John, D.: Binary ecg classification using explainable boosting machines for iot edge devices. In: 2022 29th IEEE International Conference on Electronics, Circuits and Systems (ICECS), pp. 1–4 (2022). IEEE
- Yu, Y., Li, Y., Zhou, Y., Wang, Y., Wang, J.: A learnable and explainable wavelet neural network for eeg artifacts detection and classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **32**, 3358–3368 (2024)
- Yaseen, Son, G.-Y., Kwon, S.: Classification of heart sound signal using multiple features. *Applied Sciences* **8**(12), 2344 (2018)
- Yu, B., Yin, H., Zhu, Z.: Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. *arXiv preprint arXiv:1709.04875* (2017)
- Yao, F., Zhao, W., Forshaw, M., Song, Y.: A holistic power optimization approach for microgrid control based on deep reinforcement learning. *arXiv preprint arXiv:2403.01013* (2024)
- Yao, F., Zhao, W., Forshaw, M., Song, Y.: A self-organizing interval type-2 fuzzy neural network for multi-step time series prediction. *Applied Soft Computing*, 113221 (2025)

- Zadeh, L.A.: Fuzzy sets. *Information and control* **8**(3), 338–353 (1965)
- Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning—i. *Information sciences* **8**(3), 199–249 (1975)
- Zadeh, L.A.: Fuzzy sets as a basis for a theory of possibility. *Fuzzy sets and systems* **1**(1), 3–28 (1978)
- Zhao, J., Chang, D., Cao, B., Liu, X., Lyu, Z.: Multiobjective evolution of the deep fuzzy rough neural network. *IEEE Transactions on Fuzzy Systems* **33**(1), 242–254 (2025) <https://doi.org/10.1109/TFUZZ.2024.3397728>
- Zheng, J., Fu, G., Anderson, K., Chu, H., Rakovski, C.: A 12-lead ecg database to identify origins of idiopathic ventricular arrhythmia containing 334 patients. *Scientific data* **7**(1), 98 (2020)
- Zhang, Q., Feng, H., Ding, L., Liu, T., Sun, C., Zhang, J., Nie, J.: Dynamic estimation of mean skin temperature during physical exercises using an explainable attention-based neural network: A data-driven alternative to segmental weighting methods. *Engineering Applications of Artificial Intelligence* **175**, 114604 (2026)
- Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., Chen, S.X.: Cautionary tales on air-quality improvement in beijing. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* **473**(2205), 20170457 (2017)
- Zhu, T., Meng, P., Zhou, Y.: An interpretable deep learning approach for hypertension risk assessment in the elderly. *Biomedical Signal Processing and Control* **112**, 108479 (2026)
- Zhu, Y., Niu, X., Yan, T., Xu, L.: A robust displacement monitoring model for high-arch dams integrating signal dimensionality reduction and deep learning-based

- residual correction. *Structural Control and Health Monitoring* **2024**(1), 3330769 (2024)
- Zhang, D., Peng, Y., Xu, Y., Du, C., Zhang, Y., Wang, N., Chong, Y., Wang, H., Wu, D., Liu, J., *et al.*: A high-speed railway network dataset from train operation records and weather data. *Scientific data* **9**(1), 244 (2022)
- Zurlo, G., Ronchieri, E., Cornali, R., Scarponi, L., Cesini, D., Dell’Agnello, L.: Graph neural network solutions for interpretable anomaly detection in it infrastructure monitoring time series. *Journal of Business Analytics*, 1–20 (2026)
- Zhang, Y., Tang, Q., Zhang, Y., Wang, J., Stimming, U., Lee, A.A.: Identifying degradation patterns of lithium ion batteries from impedance spectroscopy using machine learning. *Nature communications* **11**(1), 1706 (2020)
- Zhang, Q., Wu, Y.N., Zhu, S.-C.: Interpretable convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836 (2018)
- Zhang, D., Xu, Y., Peng, Y., Du, C., Wang, N., Tang, M., Lu, L., Liu, J.: An interpretable station delay prediction model based on graph community neural network and time-series fuzzy decision tree. *IEEE Transactions on Fuzzy Systems* **31**(2), 421–433 (2022)
- Zheng, G., Yue, X., Yi, W., Jia, R.: Establishment, interpretation and application of logistic regression models for predicting thermal sensation of elderly people. *Energy and Buildings* **315**, 114318 (2024)
- Zhao, W., Yuan, G., Zhang, Y., Liu, X., Liu, S., Zhang, L.: An interpretable and efficient multi-scale spatio-temporal neural network for traffic flow forecasting. *Expert Systems with Applications* **296**, 128961 (2026)

Zheng, J., Zhang, J., Danioko, S., Yao, H., Guo, H., Rakovski, C.: A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data* **7**(1), 48 (2020) <https://doi.org/10.1038/s41597-020-0386-x>