Lyu, Qi and Wu, Shaomin (2025) *An explainable machine learning framework for recurrent event data analysis.* European Journal of Operational Research, 328 (2). pp. 591-606. ISSN 0377-2217.

# An explainable machine learning framework for recurrent event data analysis

Qi Lyu, Shaomin Wu*

*Kent Business School, University of Kent, Canterbury, Kent CT2 7FS, UK*

**Abstract:** This paper introduces a novel explainable temporal point process (TPP) model, Stratified Hawkes Point Process (SHPP), for modelling recurrent event data (RED). Unlike existing approaches that treat temporal influence as a black box or rely on post-hoc explanations, SHPP structurally decomposes event intensities into semantically meaningful components for describing self-, Markovian, and joint influences. This decomposition enables direct quantification of how past events contribute to future event risks, termed as influence values. We further provide a sufficient condition for mean-square stability based on kernel decay, ensuring long-term boundedness of intensities and realistic behavioural predictions. Experiments and an e-commerce case study demonstrate SHPP's ability to deliver accurate, interpretable, and stable modelling of complex event-driven systems.

**Keywords:** (R) explainable machine learning; counting process; Hawkes process; stability; explainable artificial intelligence

# 1 Introduction

## 1.1 Motivation

In many practical applications, events occur in a recurring form. For example, patients with chronic conditions may accept repeated treatments from their hospitals due to recurring illnesses or complications (Watson et al., 2020; Chen et al., 2015); product users may recurrently claim warranty for repairing or replacing a product item under the terms of its warranty (Wu, 2012); social media users repeatedly create and share content like text, images, and videos with others via online platforms, customers in online shopping applications intermittently pick up items (Hu et al., 2022). These events are referred to as "recurrent events", and times between the occurrences of recurrent events are therefore called recurrent event data (RED).

RED analysis has been a key area of research in survival data analysis. Both statistical models and machine learning models are developed (Cook et al., 2007; Amorim and Cai, 2015; Du et al., 2016). Statistical models are for the scenarios where the size of the datasets is typically not very large. Examples include the Andersen-Gill (AG) model (which is an extension of the proportional hazards model) (Andersen and Gill, 1982), the Prentice-Williams-Peterson (PWP) models (Prentice et al., 1981), the marginal mean/rates model (Cook et al., 2007), the frailty model (Kelly and Lim, 2000), and multi-state models (Andersen and Keiding, 2002).

While traditional statistical models have laid the foundation for RED analysis, their strict assumptions—such as linearity and proportional hazards—limit their applicability to modelling complex data with high-dimensional covariates. These assumptions may be violated in emerging applications like social media and e-commerce, where RED shows complex temporal patterns and heterogeneity across subjects. As such, there is a need for developing more flexible and interpretable models to relax these assumptions and capture these dynamics.

[*] *E-mail: s.m.wu@kent.ac.uk.*

Recent advances in artificial intelligence (AI) offer promising alternatives. Deep learning methods explicitly model temporal dynamics through mechanisms like recurrent neural networks (RNNs) and attention-based transformers. For instance, Cai et al. (2020) introduced a multi-mechanism temporal framework that disentangles periodic, decaying, and persistent influences in multivariate event sequences, outperforming classical models. Gupta et al. (2019) developed a deep survival framework that jointly addresses competing risks and recurrent events by learning latent representations of time-varying risk interactions. These AI models demonstrate superior capability in capturing complex temporal patterns that defy traditional parametric assumptions.

However, the predictive ability of AI models comes at a cost: their inherent opacity. Complex neural networks, often labelled as 'black boxes', obscure the reasoning behind predictions—a critical barrier in high-risk domains like healthcare and industrial safety. For example, clinicians cannot act on a model's prediction of cancer recurrence without understanding how time-varying biomarkers (e.g., dynamic gene expression profiles) interact with prior treatment history to drive risk fluctuations (Rajpal et al., 2023). Similarly, engineers require explainable fault forecasts to prioritise maintenance actions in multi-component systems (Gashi et al., 2023).

Explainable AI (XAI) provides insights into how and why models make predictions, which is crucial for understanding complex temporal behaviours and for deploying AI systems in sensitive domains like healthcare and e-commerce. While XAI is effective for some data types such as panel data and time series data, it fails to address the temporal gap and event interdependency inherent in RED analysis. Most post-hoc methods (e.g., SHAP (Lundberg and Lee, 2017), LIME (Ribeiro et al., 2016a)) provide snapshot explanations that ignore temporal dependencies. While attention mechanisms in sequence models often combine short-term noise with long-term risk factors (Li et al., 2023). Although there is a rich literature on RED analysis, little has considered quantifying and understanding how the occurrences of historical events influence future customer behaviours. For example, in an e-commerce scenario, a customer's final action is influenced by a sequence of historical behaviours—such as repeatedly viewing an item and adding it to the cart. These behaviours correspond to three different types of historical behaviour influences, as shown in Figure  1.

- *Self-influence*: A customer views an item at time $t_1$ and returns to view it again at time $t$ (the last event with $t > t_1$). The dashed blue arrow from *View* at $t_1$ to *View* at $t$ captures this repeated behaviour, where viewing an item is regarded as a marker. That is, a marker that occurs earlier increases the probability that the marker will occur in the future.

- *Markovian influence*: The sequential path from *View* to *Cart*, and from *Cart* to *Buy*, as shown by red arrows, represents direct influence between different types of markers. For example, viewing an item may increase the chance of carting it, and carting an item may increase the chance of buying it. That is, a marker (i.e., view) directly influences the next marker (i.e., cart), and a marker (i.e., cart) directly influences the next marker (i.e., buy).

- *Joint influence*: The blue brace between *View* and *Cart* ($t_2$, jointly pointing to the *Buy* event, illustrates a combined influence. While each action alone may contribute modestly, together they significantly increase the likelihood of purchase—capturing a joint dependency that cannot be attributed to either event in isolation. That is, a marker (i.e., view) indirectly influences the next-but-one marker (i.e., buy).

This example demonstrates how different types of influence—repetition, inter-type triggering, and combinatorial influences—interact to shape a user's future decision, providing a concrete motivation for structured influence modelling in RED analysis.

However, existing models ignore these historical influences, let alone these three different influences, making users lose trust for decisions made by AI models. Motivated by this need, this paper aims to develop novel XAI methods for RED analysis, enabling an explainable and understandable framework for RED, considering the temporal historical information and three influences of events.

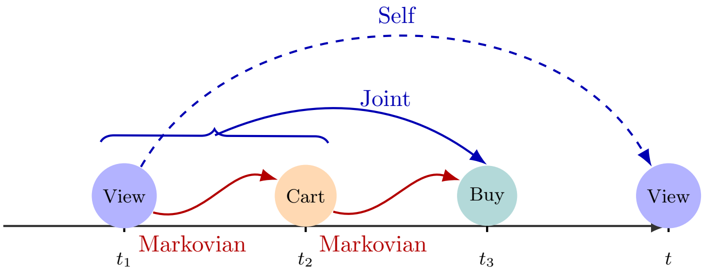

Figure 1: Illustration of self-, Markovian, and joint influences in e-commerce user behaviour

## 1.2 Related Work

### 1.2.1 RED Analysis

The literature on RED analysis has expanded rapidly, leading to the development of a diverse range of models and methodologies. RED analysis has evolved through two perspectives: statistical methods (e.g., Cook et al. (2007)) and machine learning methods (e.g., Du et al. (2016)).

*Statistical models*: Traditional approaches include the AG model, PWP models, frailty models and multi-state models. The AG model generalises the Cox proportional hazards model, which is expressed as increments in the number of events along a timeline, where the outcome of interest is the time from randomisation to treatment (or other exposure) to the event, that is, the time since the beginning of the study, also known as the total time scale (Andersen and Gill, 1982). The PWP model analyses multiple events in strata according to the number of events that occurred during follow-up, where all participants are at risk in the first stratum, but only participants who had an event in the previous stratum are at risk in the subsequent stratum (Prentice et al., 1981). The core idea of the random effects approach, also known as frailty models, is to introduce random covariates into a model, thus inducing dependencies between the times of RED (Kelly and Lim, 2000). Specifically, random effects describe the excess risk or frailty of different individuals while considering unmeasured heterogeneity that cannot be explained by observed covariates alone. The simplest multi-state model (MSM) is defined as two states: alive (a transient state) and dead (an absorbed state). A special case of MSM occurs when individuals transition from one state to another over time and intermediate states are identified. These states can be viewed as recurring events of the same marker (Andersen and Keiding, 2002). Oyamada et al. (2022) evaluated the performance of these statistical models using an open cohort design with Monte Carlo simulation in various settings and their application using an actual example. Lintu and Kamath (2022) illustrated the usefulness of RED models in the context of defect proneness analysis in software quality assessment. In addition to previous methods, some new statistical methods developed, for instance, Oganisian et al. (2024) proposed a Bayesian framework for causal analysis of recurrent events with timing misalignment. Overall, these statistical models are well-established and offer robust tools for understanding recurrent events based on probabilistic and time-dependent frameworks; more can be seen in Amorim and Cai (2015)

*Machine learning models*: Recently, machine learning has been used to analyse data from recurrent events. For example, Gupta et al. (2019) proposed a deep learning based flexible probabilistic framework for cause-specific recurrent survival analysis for both single-risk scenarios and multi-risk scenarios. Murris et al. (2024) introduced an extension of random forests tailored for RED, leveraging principles from survival analysis and ensemble learning, and evaluates their methods on both simulated and open-source data. This proposed method provides a valuable addition to the analytical toolbox in this domain.

In addition to traditional statistical models and the machine learning models that have emerged for RED, the Temporal Point Process (TPP) is another widely applied modelling method for modelling RED (Shchur et al., 2021). TPP combines the theoretical rigour of statistical methods with the ability of deep learning models to process complex high-dimensional data, becoming an important tool for RED analysis research.

In the field of statistics, classic TPPs such as the Poisson process (Dewanji and Moolgavkar, 2000) and the Hawkes process (Hawkes, 1971; Ketelbuters and Bersini, 2022) are often used in RED analysis. These models rely on explicit probabilistic assumptions and can infer the frequency and timing of events. The Hawkes process, in particular, allows for modelling both self- and mutual excitation between events, making it interpretable in terms of temporal influence structures (Xu et al., 2016).

Otherwise, in the field of deep learning, TPP has been further extended to deep learning models. Du et al. (2016) firstly proposed Recurrent Marked Temporal Point Process (RMTPP) model for RED analysis, applies a recurrent neural network to automatically learn a representation of influences from the event history. Lin et al. (2022) estimated the gap times using a generative model for TPP and revised the attentive models to improve prediction performance. There are a lot of research about TPP with neural network, and Shchur et al. (2021) summarised the existing body of knowledge on neural TPP, and provide an overview of application areas commonly considered in the literature.

However, as models become more complex, particularly in cases where non-linear or high-dimensional covariates are involved, the interpretability of models for RED analysis is decreasing. For instance, non-parametric methods and deep learning-based TPP methods excel at capturing complex relationships but often result in black-box models that lack clear interpretability. Balancing complexity with transparency remains a significant challenge, motivating continued research into explainable artificial intelligence (XAI) models, which strive to achieve both.

### 1.2.2 Explainable Artificial Intelligence (XAI)

The development of XAI has gained significant attention in recent years, especially in applications requiring both high predictive performance and transparency/interpretability (Lyu and Wu, 2025; Stevens and De Smedt, 2024; de Bock et al., 2024). This section reviews key methods that aim to balance these two aspects, progressing from traditional generalised additive models to neural extensions and specialised adaptations.

Generally, XAI methods can be categorised by their application stages, including ante-hoc and post-hoc methods (Speith, 2022; Arrieta et al., 2020). The ante-hoc methods focus on enhancing transparency and fairness during model development, for instance, developing generalised additive models (GAMs) (Chang et al., 2021) and attention branch network (ABN) (Fukui et al., 2019), both of which are explainable. While the post-hoc methods interpret or explain predictions after an AI model has been trained. Such methods include SHAP (SHapley additive exPlanations) (Lundberg and Lee, 2017) and LIME (Local Interpretable MA Explanations) (Ribeiro et al., 2016a), which attribute predictions to input features by perturbing local data points. Attention mechanisms in transformers (Wiegreffe and Pinter, 2019) provide built-in explanations by highlighting influential features/factors. More broadly, Shapley-value explanations have been extensively surveyed in the OR literature (Borgonovo et al., 2024), providing theoretical background for post-hoc baselines. Topuz et al. (2024) proposed a model utilising the inner mechanics of Markovian theory to achieve explainability and obtain interpretable scores for evaluating the performance of healthcare.

However, these methods face significant limitations when applied to RED analysis. SHAP values, for instance, treat temporal sequences as static feature vectors, ignoring the time-varying structure of event dependencies (e.g., how a prior hospitalisation alters future risk trajectories). Even if time encodings such as event indices are added, the resulting feature space does not reflect time-dependent changes, and influence attributions remain insensitive to when an event occurred.

Recent effort to adapt XAI for RED analysis and temporal data include TimeSHAP (Bento et al., 2021), which extends SHAP to RNNs by aggregating feature attributions over sliding time windows, and dynamic counterfactual explanations (Tsirtsis et al., 2021) that simulate "what-if" scenarios across event histories. While TimeSHAP captures the influence of features at a snapshot in time, it aggregates importance across fixed windows and does not decompose model predictions into individual event attributions in continuous time, which will be discussed in this work.

Transformer attention mechanisms offer another form of explanation. However, attention weights are not guaranteed to reflect true causal influence (Wiegreffe and Pinter, 2019), and they are normalised (via softmax) rather than aligned with intensity values. Attention may highlight relevant past tokens, but cannot quantify their additive contribution to a predicted event intensity.

While XAI methods can improve transparency, they also come with potential risks in high-stakes applications such as healthcare, criminal justice, and finance. As pointed out by Rudin (2019), post-hoc explanation methods like SHAP or LIME can be misleading or overly simplified. This can lead people to place too much trust in a model, even if it is incorrect. Furthermore, XAI models do not automatically gain user trust unless the quality of explanations is well-calibrated and evaluated. This challenge highlight the importance of evaluating the quality of explanations in practice.

XAI evaluation helps build consumer trust, meet demands, reduce bias, and enable more ethical and informed decision making. As AI becomes more integrated into business and the economy, XAI assessments will be increasingly crucial, promoting the responsible and effective use of AI. Lozano-Murcia et al. (2023) compared different kinds of evaluation methods on several datasets, and gave corresponding evaluation methods for feature importance, consistency, stability and robustness, computation time and efficiency, fairness and bias and regulatory compliance. Recently, the OR community has begun to systematise XAI under an "XAIOR" framework (de Bock et al., 2024), outlining design principles and evaluation criteria, which will be followed in this paper.

In summary, XAI techniques have made significant progress in static settings and sequence modelling. However, when applied to RED, these techniques still have several limitations:

- *Lack of temporal sensitivity*: Most XAI methods treat events as isolated points, ignoring how the influence of past events decays or accumulates over time. This leads to temporally myopic explanations that miss long-term dependencies crucial in domains like healthcare or e-commerce.
- *Inability to attribute historical influence*: Existing methods fail to quantify how specific past events contribute to current risks. For example, a history of product returns may signal declining purchase intent, but snapshot explanations cannot trace or assign influence to such patterns.
- *Predictive–interpretability trade-off*: Traditional statistical models (e.g., Cox models) offer interpretability but struggle with complex event dynamics. In contrast, high-capacity models (e.g., neural TPPs) perform well in predictive performance but lack built-in interpretability, often relying on unreliable post-hoc explanations.

These gaps motivate us towards *XAI for RED analysis*—a challenge we address with our proposed method in this paper. Our proposed framework clearly models temporal influence—decomposing it into self-, Markovian, and joint influences—and provides interpretability through influence values.

### 1.3 Overview

The remainder of this paper is organised as follows. Section 2 introduces a novel explainable temporal point process (TPP) model, Stratified Hawkes Point Process (SHPP), for modelling RED. Section 3 discusses the experimental design and their applications in practical scenarios. Section 5 concludes the research conclusions and proposes future research directions.

## 2 Methodology

Let $\{t_i\}_{i\geq 1}$ denote the occurrence times of events with $0 < t_1 < t_2 < \cdots$, and $t_0 (= 0)$ denote the starting time. The associated *counting process* is defined by $N(t) = \sup\{n \geq 0 \colon t_n \leq t\}$, representing the total number of events by time $t$, as illustrated in Fig. 2. Suppose that each occurrence has a marker associated with it and $p$ covariates. Denote the marker at the $i$-th event occurrence as $m_i$, where $m_i \in \mathcal{M}$ with $\mathcal{M} = \{1, 2, \ldots, K\}$, and $K$ is the number of marker types. Denote the covariates as $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top \in \mathbb{R}^p$, where $\boldsymbol{x}_i$ can be variable in time or static. The $i$-th event is characterised by the tuple $c_i = (t_i,\, m_i,\, \boldsymbol{x}_i)$.

Denote the gap time between the $i$-th and $(i-1)$-th events as $\tau_i = t_i - t_{i-1}$ for $i \geq 1$. For any time $t > 0$, the observed history up to $t$ is

$$\mathcal{H}_{[0,t)} = (c_k \colon t_k < t)_{k=1}^{N(t^-)}, \tag{1}$$

where $N(t^-) = \lim\limits_{s\to t} N(s)$ ensures exclusion of events exactly at $t$.

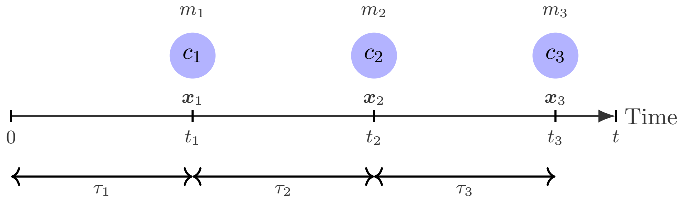

Figure 2: Recurrent event data structure.

Real-world recurrent event data typically arise from multiple interacting events rather than isolated event. To illustrate the proposed influence mechanism, consider an e-commerce user browsing and purchasing items (e.g., smartphones). An e-commerce firm would like to understand their customers' behaviour by modelling times between views or purchases. To this end, they need to know the exact times when the actions are taken, where each action is marked with a marker such as *View*, *Cart*, *Purchase*, or *Return* (that is, $K = 4$ and the associated covariates $\boldsymbol{x_i}$ may include user profile (e.g., age, VIP level), product attributes (e.g., discount, rating), or behavioural features (e.g., time spent, browsing frequency). To build a model for depicting the times between events, we need to consider the association between the markers from the three perspectives: self-influence, Markovian influence, and joint influence, as discussed in Section 1.1. However, existing models either neglect these three types of influences, or oversimplify them by only considering temporal gaps $\tau_i$ between events. They fail to capture the influence from historical markers $m_i$ and covariates $\boldsymbol{x_i}$. To solve these problems, this paper aims to model RED by considering the markers, the covariates, and the three types of influences. To characterise the logical of decision making based on RED rigorously, we propose an interpretable framework with the three types of *influence*, which capture temporal dependencies and interactions among events:

(i). *Self-influence*: Historical occurrences of the same marker modify the likelihood of similar events that will occur in the future.

(ii). *Markovian influence*: Direct interactions between different event markers where one marker explicitly influences another.

(iii). *Joint influence*: The joint influence of multiple past event marker sequences collectively influence future event occurrences.

We give the definition of *influence* in this work.

**Definition 1** (Influence)**.** *Influence is a term that describes the temporal association or interaction from a set of past events* $\{c_i\}_{t_i<t}$ *towards one or multiple subsequent events* $\{c_j\}_{j\geq t}$.

This term captures the extent to which earlier events collectively relate to or predict future event occurrences, acknowledging that multiple historical factors may shape these temporal associations.

Building on the influence framework from the previous description, we further formalise the analysis of RED through TPP. A TPP is a stochastic model characterizing event sequences $\{t_i\}_{i=1}^{n}$ (Rizoiu et al., 2017) and can be modelled by a *conditional intensity* function:

**Definition 2** (Conditional Intensity (Daley and Vere-Jones, 2006))**.** *Given history* $\mathcal{H}_{[0,t)}$, *a conditional intensity* $\lambda(t|\mathcal{H}_{[0,t)})$ *is defined by:*

$$\lambda(t|\mathcal{H}_{[0,t)}) = \lim_{\Delta\to 0^+} \frac{\mathbb{P}\left(N([t,t+\Delta)) \geq 1 \mid \mathcal{H}_{[0,t)}\right)}{\Delta}, \tag{2}$$

*where* $N([t,t+\Delta))$ *counts the number of events in interval* $[t,t+\Delta)$.

A conditional intensity function can fully specify a TPP through two fundamental components:

- *Event probability* can be defined by:

$$\mathbb{P}(\text{Occurrence of an event in } [t,t+dt) \mid \mathcal{H}_{[0,t)}) = \lambda(t|\mathcal{H}_{[0,t)})dt + o(dt), \tag{3}$$

  where $o(dt)$ satisfies $\lim\limits_{dt\to 0} o(dt)/dt = 0$.

- The *survival function* of the $i$-th occurrence can be defined by:

$$S(t|\mathcal{H}_{[0,t)}) = \exp\left(-\int_{t_i}^{t} \lambda(\tau|\mathcal{H}_{[0,\tau)})d\tau\right), \quad t > t_i. \tag{4}$$

The established notations in Eqs (1)-(4) provide a general framework for RED analysis. However, widely used TPPs such as the Poisson process and the renewal process cannot model the aforementioned influences because they fail to explain the connections between events. This is where *mutual point process* come into play—it accounts for the excitatory influences between events. For instance, purchasing item A can stimulate subsequent purchases, creating a chain of influence reaction.

The self-exciting process, aka Hawkes' process, and the mutual exciting process (MEP) represent special cases of point processes that model event occurrences conditioned on historical information. Formally, these intensities are expressed using conditional intensity functions, given the event history $\mathcal{H}_{[0,t)]}$.

- *Hawkes process*: Historical events of a single marker increase the likelihood of future occurrences of the same marker. Its conditional intensity function is defined as (Hawkes, 1971):

$$\lambda(t|\mathcal{H}_{[0,t)}) = \mu + \sum_{r=1}^{N(t)} \gamma(t-T_r), \tag{5}$$

  where $T_r$ denotes the occurrence of the $r$-th event, and $\mu$ represents base intensity.

- *Mutual-exciting process* generalizes the Hawkes process to multiple occurrences of events with markers. The MEP models how occurrences of one type of marker influence an event with a different type of marker, and the conditional intensity is (Daley and Vere-Jones, 2006):

$$\lambda_i(t|\mathcal{H}_{[0,t)}) = \mu_i + \sum_{j=1}^{K} \sum_{r=1}^{N_j(t)} \gamma_{ij}(t - T_{jr}), \tag{6}$$

  where $T_{jr}$ denotes the $r$-th occurrence time of event with marker of type $j$, $\mu_i$ represents a base intensity for marker of type $i$, and $N_j(t)$ is the number of the occurrences of events with marker of type $j$ by time $t$. The kernel $\gamma_{ij} \colon \mathbb{R}^+ \to \mathbb{R}^+$ quantifies how an event with a marker of type $j$ excites future events with a marker of type $i$.

However, the MEP assumes that each past event contributes independently and additively to the future event intensity. Nevertheless, both self- and Markovian influences can exhibit not only excitatory influences but also inhibitory behaviours, which cannot be modelled by Hawkes' process or the MEP. Furthermore, the MEP cannot model joint influence, which requires non-additive interactions among multiple events. To overcome these limitations, we propose a new TPP, as shown in Section 2.1.

## 2.1 Stratified Hawkes Point Processes

This section proposes a new TPP: *stratified Hawkes point process* (SHPP), which can model self- and Markovian influences comprehensively.

**Definition 3** (Stratified Intensity)**.** *For recurrent events with $K$ type of markers, the intensity of events with marker of type $i$ is:*

$$\lambda_i(t|\mathcal{H}_{[0,t)]}) = \exp\left( \underbrace{\mu_i}_{Base\ Rate} + \sum_{j=1}^{K} \gamma_{ij}\Big(\{t - T_{jk}\}_{k=1}^{N_j(t)} |\mathcal{H}_{[0,t)]}\Big)\right), \tag{7}$$

*where $\gamma_{ij} \colon \mathbb{R}^{N_j(t)} \to \mathbb{R}$ encodes the stratified influence from events with marker of type $j$ to events with marker of type $i$, considering all historical $\{T_{jk}\}_{k=1}^{N_j(t)}$.*

The word *stratified* highlights that the influence from past events is decomposed by marker types: for each event of marker of type $i$, its intensity $\lambda_i(t)$ considers contributions from each marker of type $j$ through a specific kernel $\gamma_{ij}$. In particular, it supports *self-influence* for the case of $i = j$ (e.g., repeated views reinforcing future views), and *Markovian influence* for the case of $i \neq j$ (e.g., cart actions increasing purchase likelihood).

The stratified intensity function in Eq.(7) provides a mathematical foundation for RED analysis. However, to fully characterise the stochastic process governing these events, we must define the probabilistic structure that links the intensity function to the actual event occurrences. This leads us to the following definition of a stratified Hawkes point process (SHPP):

**Definition 4** (Stratified Hawkes Point Process)**.** *A collection $\{N_i(t)\}_{i=1}^{K}$ forms a stratified Hawkes point process if:*

$$\mathbb{P}(N_i(t+\Delta) - N_i(t) = 1|\mathcal{H}_t) = \lambda_i(t)\Delta + o(\Delta), \tag{8}$$

$$\mathbb{P}(N_i(t+\Delta) - N_i(t) > 1|\mathcal{H}_t) = o(\Delta), \tag{9}$$

*where $\mathcal{H}_t = \sigma(\{N_j(s)\}_{j=1}^{M} \colon s \leq t)$ contains the complete history, and $\lambda_i(t)$ follows Eq. (7).*

Compared with the MEP, past events $T_{jr}$ in Eq. (6) contribute independently to the intensity in an additive manner. The function $\gamma_{ij}(\cdot)$ typically depends only on the time difference $(t - T_{jr})$, limiting its ability to capture higher-order dependencies. In contrast, $\gamma_{ij}\colon \mathbb{R}^{N_j(t)} \to \mathbb{R}$ in our proposed model (shown in Eq. (7)) encodes the stratified influence from events with marker of type $j$ to events with marker of type $i$, considering the entire history $\mathcal{H}_{[0,t)}$. This structure introduces two key differences: (i). The exponential transformation enables multiplicative interactions rather than additive influences; (ii). The function $\gamma_{ij}$ operates on the entire historical sequence rather than individual time gaps $t - T_{jr}$ in equation (6). These two differences enhance the model's ability to represent complex, higher-order dependencies, capturing intricate patterns such as combined excitation and inhibition influences in RED.

To assess the influence from historical events, SHPP introduces a kernel function $\gamma_{ij}$, which modulates the impact of marker of type $j$ events on marker of type $i$. This design enables several distinctive properties:

**Remark 1** (Key Properties of SHPP)**.** *The proposed SHPP has the following properties:*

- *Nonlinear Coupling: SHPP adopts an exponential link function that combines event influences multiplicatively, enabling the model to capture nonlinear accumulation effects beyond additive frameworks.*
- *History-Aware Kernels: Unlike traditional Hawkes models that treat events independently via time gap functions, SHPP's kernel $\gamma_{ij}$ can incorporate the full historical context, including temporal features and covariates, allowing it to model complex sequential dependencies.*
- *Flexible Influence Semantics: SHPP supports both excitation ($\gamma_{ij} > 0$) and inhibition ($\gamma_{ij} < 0$) effects, and can model mixed patterns, which is not possible under classical Hawkes assumptions where all influences are positive.*

**Example 1.** *Consider the task of predicting whether a customer will make a purchase on an e-commerce platform. The three core properties of the proposed SHPP model work together to capture the complexity of real customer behaviour:*

*(i). Nonlinear coupling models how multiple factors—such as repeated product views, recent promotions, and prior purchases—can jointly amplify the likelihood of a purchase. This goes beyond simple additive influences by capturing interactions between events.*

*(ii). History-aware kernels consider the customer's entire browsing and interaction history, not just recent actions. This allows the model to recognise long-term patterns that may signal sustained interest or disengagement.*

*(iii). Flexible influence enables the model to represent both positive influences (e.g., increasing interest through discounts) and negative influences (e.g., repeated poor reviews reducing likelihood).*

To better understand the generality of our framework, we now show that SHPP can reduce to several classical models under specific parameterisations.

**Remark 2** (Connections to other processes)**.** *The proposed SHPP framework generalises several other point processes:*

- *Hawkes process: If only self-influence is retained (i.e., $i = j$ for all $i$), and Markovian/joint influences are absent, that is, $\lambda_i(t) = \mu_i + \sum_{r=1}^{N_i(t)} \gamma_{ii}(t - T_{ir})$, then, SHPP reduces to a Hawkes process,*
- *Mutual exciting process: When the kernel depends only on individual time gaps and influences are additive, i.e., $\gamma_{ij}(t) = \sum_{r=1}^{N_j(t)} \gamma_{ij}(t - T_{jr})$, then SHPP reduces to an MEP, and*
- *Homogeneous Poisson process: If all influence terms vanish, i.e., $\gamma_{ij} \equiv 0$, the intensity becomes constant: $\lambda_i(t) = \mu_i$, then SHPP reduces to a homogeneous Poisson process.*

*These reductions show that SHPP improves modelling flexibility while remaining compatible with classical models.*

## 2.2 Stability Analysis

The SHPP captures how different events influence each other over time. However, to ensure the model's predictions stay realistic and reliable, especially over long periods, we need to ensure dynamic stability. Without this property, the model may output meaningless results, like predicting infinite medication doses in healthcare scenario or vanishing user actions in e-commerce scenario. Thus, this section defines the concept of dynamic stability, discusses its importance for practical applications, and explains how the SHPP framework is designed to ensure it.

- *Dynamic stability* (Hawkes, 1971): Mathematically, $\exists C > 0$ such that:

$$\mathbb{P}\left(\sup_{t>0} \lambda_i(t) \leq C\right) = 1 \quad \forall i \in \mathcal{M}. \tag{10}$$

  This ensures the model does not predict impossible scenarios—like a patient taking infinite medication doses in a short period.

**Example 2.** *To illustrate the importance of dynamic stability, consider an e-commerce platform analysing two key user actions:*

- *Event-A: product impressions (system recommends or displays a product), and*
- *Event-B: user clicks (user clicks on the product).*

*Suppose the model learns that impressions strongly increase the likelihood of clicks, and clicks in turn induce more impressions (e.g., via a recommender system loop).*

- *If this mutual exciting is not properly controlled, the model may predict a runaway feedback loop: infinite impressions and clicks in a short time, which is an unrealistic and undesirable scenario,*
- *Conversely, if negative feedback is too strong (e.g., assuming that users become completely uninterested after a single impression), the model may predict that users never interact again, which also contradicts real-world behaviour, where users often return after delays.*

*These outcomes reflect a lack of dynamic stability, where the model fails to keep event intensities within realistic bounds over time. Ensuring stability helps prevent such unreal behaviour and ensures the model remains reliable in long-term forecasting.*

To rigorously analyse the dynamic stability of the proposed SHPP, we first establish the probabilistic framework. Let $(\Omega, \mathcal{H}, \mathbb{P})$ represent the filtered probability space supporting all counting processes $\{N_i(t)\}_{i=1}^{K}$, where the filtration $\mathcal{H}_t$ encodes historical event information.

We define the *intensity vector* $\mathbf{\Lambda}(t) = (\lambda_1(t), ..., \lambda_K(t))^\top \in \mathbb{R}_+^K$, where $\lambda_i(t)$ is the conditional intensity of event of type $i$. The evolution of this system is governed by a differential equation derived from the SHPP formulation:

$$\dot{\mathbf{\Lambda}}(t) = F(\mathbf{\Lambda}(t)), \quad F_i(\mathbf{\Lambda}) = \lambda_i(t) \sum_{j=1}^{K} \sum_{k=1}^{N_j(t)} \frac{\partial \gamma_{ij}(t - T_{jk})}{\partial t}. \tag{11}$$

This dynamical formulation allows us to analyse the stability of SHPP using tools from stochastic process theory and dynamical systems. In the context of RED analysis, the mean-square stability ensures that the expected intensity remains bounded over time, preventing unrealistic behaviours. We adopt the following definition adapted from Higham (2000):

**Definition 5** (Mean-Square Stability (Higham, 2000)). *A stochastic intensity process $\mathbf{\Lambda}(t)$ is said to be mean-square stable if*

$$\limsup_{t\to\infty} \mathbb{E}\left[\|\mathbf{\Lambda}(t)\|^2\right] < \infty,$$

*where $\|\mathbf{\Lambda}(t)\|^2 := \sum_{i=1}^K \lambda_i^2(t)$ measures the total fluctuation in intensity.*

We now establish a sufficient condition under which SHPP satisfies this mean-square stability criterion.

**Theorem 1** (Sufficient Condition for Mean-Square Stability of SHPP)**.** *Consider a Stratified Hawkes Point Process (SHPP) with intensity vector $\mathbf{\Lambda}(t)$. If all kernel functions $\gamma_{ij}(\tau)$ are non-increasing in $\tau > 0$, i.e.,*

$$\frac{\partial \gamma_{ij}(\tau)}{\partial \tau} \leq 0, \quad \forall \tau > 0,\ \forall i, j \in \mathcal{M}, \tag{12}$$

*then the process is mean-square stable:*

$$\limsup_{t\to\infty} \mathbb{E}\left[\|\mathbf{\Lambda}(t)\|^2\right] < \infty, \tag{13}$$

*where $\|\mathbf{\Lambda}(t)\|^2 := \sum_{i=1}^K \lambda_i^2(t)$ measures total intensity fluctuations. This condition guarantees that intensities remain bounded over time to prevent explosion.*

*Proof.* To analyse the long-term boundedness of the intensity process, we adopt Lyapunov's second method from stochastic stability theory (Khasminskii, 2011). Let the Lyapunov candidate function be:

$$V(\mathbf{\Lambda}) = \|\mathbf{\Lambda}(t)\|^2 = \sum_{i=1}^K \lambda_i^2(t), \tag{14}$$

While the choice $V(\mathbf{\Lambda}) = \sum_{i=1}^K \lambda_i^2(t)$ represents a specific Lyapunov candidate, where the quadratic form captures intensity variance, and the kernel decay condition ensures its monotonic decrease. More general Lyapunov functions exist but would complicate interpretation without strengthening results.

This function satisfies:

- *Radial unboundedness*: $V(\mathbf{\Lambda}) \geq 0$ and grows without bound as $\|\mathbf{\Lambda}\| \to \infty$,
- *Monotonic decay*: The kernel condition $\partial\gamma_{ij}/\partial\tau \leq 0$ ensures that the cumulative contribution from past events is non-increasing.

Applying the infinitesimal generator $\mathcal{L}$, we compute:

$$\begin{aligned}
\mathcal{L}V &= 2\sum_{i=1}^K \lambda_i(t)\dot{\lambda}_i(t) \\
&= 2\sum_{i=1}^K \lambda_i^2(t) \sum_{j=1}^K \sum_{k=1}^{N_j(t)} \frac{\partial \gamma_{ij}(t - T_{jk})}{\partial t} \\
&\leq 2\sum_{i=1}^K \lambda_i^2(t) \sum_{j=1}^K \left( \sum_k \frac{\partial \gamma_{ij}(t - T_{jk})}{\partial t} \right) \\
&\leq -2\sum_{i=1}^K \lambda_i^2(t)\Gamma_i, \quad \text{where } \Gamma_i := -\max_j \sum_k \frac{\partial \gamma_{ij}}{\partial t} > 0 \\
&\leq -2\Gamma V(t), \quad \text{where } \Gamma := \min_i \Gamma_i > 0.
\end{aligned}$$

By Lyapunov's stability theorem, this exponential decay yields:

$$\mathbb{E}[V(t)] \leq V(0)e^{-2\Gamma t} \quad \Rightarrow \quad \limsup_{t\to\infty} \mathbb{E}[V(t)] = 0. \tag{15}$$

Therefore, $\limsup_{t\to\infty} \mathbb{E}[\|\mathbf{\Lambda}(t)\|^2] < \infty$, which completes the proof of mean-square stability. ■

To clarify the meaning of our stability condition, we provide an example from an e-commerce scenario.

**Example 3.** *Consider an e-commerce platform with two key user actions:*

- *Action 1 – Views ($m_1$): Users tend to revisit or re-explore products they have viewed before. For example, it can be modelled via a self-influence kernel: $\gamma_{11}(\tau) = \alpha_1 e^{-\beta_1 \tau}$, where $\alpha_1, \beta_1 > 0$. The decay term $\beta_1$ ensures that earlier views gradually lose influence.*
- *Action 2 – Cart Adds ($m_2$): Cart behaviour is influenced by recent browsing activity, which may both trigger and suppress add-to-cart actions depending on user intent. For example, a Markovian influence can be modelled by: $\gamma_{21}(\tau) = -\alpha_2 e^{-\beta_2 \tau}$, where $\alpha_2 > 0$.*

*Now consider two bad cases from the mean-square stability condition:*

- *Case 1 – Unstable Browsing: If the kernel for product views increases over time, e.g., $\gamma_{11}(\tau) = \alpha_1 \tau$, the intensity $\lambda_1(t)$ may grow uncontrollably, leading to unrealistic predictions such as infinite browsing behaviour.*
- *Case 2 – Over-Inhibition of Cart Adds: If the inhibition from views to cart adds grows with time (e.g., $\gamma_{21}(\tau) = -\alpha_2 e^{\beta_2 \tau}$), the model may predict that users stop adding items to carts altogether, contradicting typical return-to-cart behaviour seen in real-world platforms.*

The above examples show how violating the kernel decay condition $\partial\gamma/\partial\tau \leq 0$ results in unstable or unreal system behaviour. Complying with the stability condition ensures that user activity evolves in a bounded and interpretable manner.

For a summary of which kernel types satisfy the stability condition, please refer to Appendix C..

## 2.3 Interpretable Kernel Design

### 2.3.1 Interpretable Decomposition

Section 2.1 introduced the definition of the SHPP to model self- and Markovian influences. However, joint influence, where multiple historical events interact to affect future outcomes, cannot be modelled unless the kernel function $\gamma_{ij}$ is designed to capture such higher-order dependencies. In this section, we introduce a kernel function, $\gamma_{ij}(\cdot)$, to capture joint influence.

In discussed in Section 1.1, we need to ensure an AI model be explainable. To this end, we can use an interpretable kernel function in the SHPP, where the term *interpretable* refers to the model's ability to attribute the intensity of an event to specific historical events and influence types. For example, in an e-commerce scenario, the predicted likelihood of a purchase event can be broken down into influences such as repeated product views (self-influence), recent cart additions (Markovian influence), and the joint effect of viewing and carting together.

The interpretability of the kernel relies on Theorem 2, which shows how any continuous multivariate function can be decomposed into a finite sum of univariate functions, which is a result that is widely adopted for functional decomposition.

**Theorem 2** (Kolmogorov-Arnold Representation Theorem (Schmidt-Hieber, 2021))**.** *For any continuous multivariate function $f\colon \mathbb{R}^n \to \mathbb{R}$, then $f$ can be written as a finite composition of continuous functions of a single variable and the binary operation of addition. More specifically,*

$$f(x_1, ..., x_n) = \sum_{q=0}^{2n} \Phi_q \left( \sum_{p=1}^{n} \phi_{q,p}(x_p) \right), \tag{16}$$

*where $\Phi_q\colon \mathbb{R} \to \mathbb{R}$ and $\phi_{q,p} : [0,1] \to \mathbb{R}$ are continuous univariate functions.*

Based on the Kolmogorov–Arnold representation theorem, we construct our kernel using a combination of univariate functions that support both expressiveness and interpretability. This motivates the following structured decomposition.

The kernel $\gamma_{ij} \colon \mathcal{Z} \to \mathbb{R}_+$ in our SHPP is defined over the input space $\mathcal{Z} = \mathbb{R}^{p+1}$, which consists of covariates $\mathbf{x} \in \mathbb{R}^p$ and a temporal feature $\tau \in \mathbb{R}_+$. According to Theorem 2, any continuous multivariate function defined over this domain can be expressed as a finite sum of outer univariate functions applied to inner univariate transformations. Specifically:

$$\gamma_{ij}(\mathbf{z}) = \sum_{q=0}^{2d} \Phi_q \left( \sum_{k=1}^{d} \phi_{q,k}(z_k) \right), \quad d = p + 1. \tag{17}$$

To retain interpretability while ensuring sufficient expressiveness, we retain only the first two components and assume linear outer functions: $\Phi_0(y) = y$, $\Phi_1(y) = y$. This simplification preserves the additivity of influence contributions, allowing for clear attribution. Under this design, the kernel is structured as follows:

$$\gamma_{ij}(\mathbf{z}) = \Phi_0 \left( \sum_{k=1}^{N_j(t)} \phi_{0,k}^{(ij)}(\mathbf{z}_k) \right) + \Phi_1 \left( \sum_{k=1}^{N_j(t)} \sum_{s \neq k} \phi_{1,ks}^{(ij)}(\mathbf{z}_k, \mathbf{z}_s) \right) = \underbrace{\sum_{k=1}^{N_j(t)} \phi_{0,k}^{(ij)}(\mathbf{z}_k)}_{\text{Self-/Markovian Influence}} + \underbrace{\sum_{k=1}^{N_j(t)} \sum_{s \neq k} \phi_{1,ks}^{(ij)}(\mathbf{z}_k, \mathbf{z}_s)}_{\text{Joint Influence}}, \tag{18}$$

where we use superscripts $(ij)$ to indicate dependence on the target-output marker pair. And the first term quantifies the impact of each historical event $\mathbf{z}_k$, capturing both self-influence (when $i = j$) and Markovian influence (when $i \neq j$). The second term models higher-order interactions among multiple historical events, thereby enabling the expression of joint influence.

To enhance interpretability, we adopt base function expansions for both components:

- *Self-/Markovian Influence*:

$$\phi_{0,k}^{(ij)}(\mathbf{z}_k) = \sum_{r=1}^{p+1} \beta_{ij,r}^{(k)} g_r(z_{kr}), \tag{19}$$

- *Joint Influence*:

$$\phi_{1,ks}^{(ij)}(\mathbf{z}_k, \mathbf{z}_s) = \sum_{r,u=1}^{p+1} \theta_{ij,ru}^{(ks)} h_{ru}(z_{kr}, z_{su}), \tag{20}$$

where $g_r(\cdot)$ and $h_{ru}(\cdot, \cdot)$ are chosen as interpretable functions such as decision trees or generalised linear models (Quinlan, 1986; Ribeiro et al., 2016b).

Thus, the final interpretable kernel is:

$$\gamma_{ij}(\mathbf{z}) = \sum_{k=1}^{N_j(t)} \sum_{r=1}^{p+1} \beta_{ij,r}^{(k)} g_r(z_{kr}) + \sum_{k=1}^{N_j(t)} \sum_{s>k} \sum_{r,u=1}^{p+1} \theta_{ij,ru}^{(ks)} h_{ru}(z_{kr}, z_{su}). \tag{21}$$

This formulation enables explicit decomposition into self/Markovian and joint influences, ensuring interpretability while capturing complex dependencies in RED.

To operationalize temporal influence, we adopt exponential-shaped kernels of the form $\kappa(t - t_k) = \alpha \exp(-\beta(t - t_k))$. These kernels are interpretable, capturing decaying influence over time and analytically tractable, with decay parameters directly controlling long-term behaviour. While our framework supports alternative kernels (e.g., power-law), we default to exponential forms for their simplicity and stability guarantees. A comparative overview of common kernels and their theoretical properties is provided in Table 6 in  Appendix C..

For the interpretable kernel basis, we use *logistic units* for $g_r(\cdot)$ and *bilinear forms* for $h_{ru}(\cdot,\cdot)$. This choice balances interpretability with expressiveness: logistic functions yield bounded, smooth attribution, while bilinear terms naturally capture pairwise covariate interactions. More base functions are provided in Table 8.

### 2.3.2 Interpretability Mechanism

The stratified architecture of SHPP enables explicit attribution of intensities to three types of influence. By design, each perspective corresponds to an observable mechanism, allowing the model to quantify "why" an event is likely to occur. We refer to each quantifiable component of the kernel $\gamma_{ij}(\cdot)$ as an influence value, representing the importance and direction of impact from specific historical events on intensity.

**Corollary 1** ( Influence Values)**.** *For any intensity $\lambda_i(t)$ and interpretable kernel $\gamma_{ij}(\mathbf{z})$ given by SHPP, the following influence values can be extracted:*

- *Self-influence value: $\sum_{r=1}^{p+1} \beta_{ij,r}^{(k)} g_r(z_{kr})$ in Eq.* (21) *with $i = j$ quantifies the self-influence of the $k$-th historical event,*
- *Markovian influence value: $\sum_{r=1}^{p+1} \beta_{ij,r}^{(k)} g_r(z_{kr})$ in Eq.* (21) *with $i \neq j$ captures the influence from marker of type $j$ to marker of type $i$, and*
- *Joint influence value: $\sum_{r,u=1}^{p+1} \theta_{ij,ru}^{(ks)} h_{ru}(z_{kr}, z_{su})$ models within-marker dependencies based on interactions between pairs of historical events.*

*Each influence value represents a quantifiable contribution to the intensity $\lambda_i(t)$, enabling interpretable tracing of event-to-event temporal influence.*

For self-influence and Markovian influence, the influence value has $n$ dimensions, where each element represents the influence of a past event $e_k = (t_k, m_k)$ on the subsequent event $e_{k+1} = (t_{k+1}, m_{k+1})$, with $k \in 1, 2, \ldots, n$. In other words, each value quantifies how much a specific historical event contributes to the occurrence of the next event, and it can be represented by:

$$\mathcal{I}(e_k) = \begin{cases} \sum_{r=1}^{p+1} \beta_{ij,r}^{(k)} g_r(z_{kr}), \text{if } m_k = m_{n+1} (\text{self-influence}), \\ \sum_{r=1}^{p+1} \beta_{ij,r}^{(k)} g_r(z_{kr}), \text{if } m_k \neq m_{n+1} (\text{Markovian influence}), \end{cases} \tag{22}$$

where $k \in \{1, 2, \cdots, n\}$ are the occurrence of events.

For joint influence, the influence values are organised as a matrix, where each element corresponds to the influence between a pair of markers $(i, j)$:

$$\mathcal{I}(i, j) = \sum_{r,u=1}^{p+1} \theta_{ij,ru}^{(ks)} h_{ru}(z_{kr}, z_{su}), \tag{23}$$

where $i, j \in \mathcal{M}$ are type of markers. Specifically, each element captures how events with marker of type $i$ influence events with marker of type $j$ across the entire sequence.

We emphasise that covariate effects are explicitly modelled by the base functions $g_r(z_{kr})$ and $h_{ru}(z_{kr}, z_{su})$ in the SHPP framework, where $z_{kr}$ encodes covariate values from historical events. As such, attribution values naturally reflect both temporal positioning and covariate influence. This enables users to assess not only when, but also under what contextual conditions (e.g., product category, user demographics) historical events exert influence on future ones.

While self-/Markovian influence in Eq. (22) captures local temporal dependencies and joint influence in Eq. (23) models global interactions, each of them offers a limited, single-perspective influence.

To better integrate the contributions of different influence mechanisms, we propose a dynamic hybrid weighting strategy to combine self-/Markovian and joint influences into a unified influence value. Specifically, we define a combined influence value for the $k$-th event as:

$$\mathcal{I}^{(k)}_{\text{combined}} = \alpha \underbrace{\left( \frac{\mathcal{I}^{(k)}_{\text{self/Markovian}}}{\max_j |\mathcal{I}^{(j)}_{\text{self/Markovian}}| + \epsilon} \right)^2}_{\text{Normalised self/Markovian importance}} + (1-\alpha) \underbrace{\left( \frac{\mathcal{I}^{(k)}_{\text{joint}}}{\max_j |\mathcal{I}^{(j)}_{\text{joint}}| + \epsilon} \right)^2}_{\text{Normalised joint importance}}, \tag{24}$$

where:

- $\mathcal{I}^{(k)}_{\text{self/Markovian}}$: the individual influence value of event $e_k$ derived from self- or Markovian influence,
- $\mathcal{I}^{(k)}_{\text{joint}}$: the aggregated joint influence involving event $e_k$ as part of pairwise interactions,
- $\alpha \in [0, 1]$: a learnable coefficient that adaptively balances the two values,
- $\epsilon > 0$: a small constant added to avoid division by zero during normalisation.

The adaptive weight $\alpha$ allows the model to shift emphasis based on scenario. For example, in repetitive behaviour scenarios, $\alpha \to 1$ prioritizes self-influence, while in combinatorial conditions, $\alpha \to 0$ emphasises joint patterns. The final influence value $\mathcal{I}^{(k)}_{\text{combined}}$ keeps the interpretability of three influences while showing how they work together to influence events.

## 2.4 Evaluation Metrics

In this section, we will compare our proposed SHPP model with existing TPP models. To this end, we will use EasyTPP, a user-friendly framework for developing and benchmarking temporal point process (TPP) models (Xue et al., 2024). We evaluate each model from two main aspects:

(i). *Predictive performance*, we need to measure the performance of our proposed method, as explained below.

- *Marker prediction*: Given a sequence of historical events up to time $t$, the model predicts the next event with marker-$m_{i+1}$. To measure the performance of the prediction, we use some metrics for measuring the performance of classification models. Such classification metrics include *Accuracy*, *F1-score*, or *Top-k Precision*, depending on the number of event markers (Novaković et al., 2017),
- *Time forecasting*: The model predicts the time $t_{i+1}$ at which the next event will occur. This is evaluated using the *mean absolute error (MAE)* or the *root mean squared error (RMSE)* between predicted and actual event times, reflecting how well the model captures temporal dynamics (Armstrong, 2001).

For consistency, we report RMSE for event time forecasting and accuracy for marker prediction across all models.

(ii). *Interpretability*, we will focus on *fidelity*—the degree to which the explanation reflects the true behaviour of the model (Miró-Nicolau et al., 2025). High fidelity indicates that explanations closely match the model's actual predictions.

Since fidelity lacks a standardised definition (Miró-Nicolau et al., 2024), we assess it from two perspectives:

- *Internal consistency*: whether the explanation aligns with the model's own decision-making,
- *Fidelity to real data*: whether the explanation reasonably supports the model's outputs with respect to actual event outcomes.

### 2.4.1 Internal Consistency

To assess internal consistency, we first design perturbation strategies that test whether the model's explanations align with its own predictive behaviour. The central idea is that if certain events are truly important—i.e., assigned high influence values $\mathcal{I}(e_i)$—then perturbing them should cause meaningful changes in model outputs. Conversely, if perturbing low-influence events has little changes, the explanation is considered consistent.

Given an event sequence $S = \{(t_1, m_1), \ldots, (t_n, m_n)\}$, where $e_i = (t_i, m_i)$ and $\mathcal{I}(e_i)$ denotes the influence value of $e_i$ for predicting the next event $(t_{n+1}, m_{n+1})$, we propose the following three perturbation strategies:

- *Event deletion*: Remove top-$k$ events with highest $\mathcal{I}(e_i)$: $S_{\text{masked}} = S \setminus \{e_j \mid \mathcal{I}(e_j) \in \text{Top}_k(\mathcal{I}(S))\}$.
- *Time shifting*: Add Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ to timestamps: $t_j^{'} = t_j + \epsilon \quad \text{for} \quad e_j \in \text{Top}_k(\mathcal{I}(S))$.
- *Marker flipping*: Alter markers to random markers $m' \in \mathcal{M} \setminus \{m_j\}$.

Building upon the perturbation strategies, we now formalise the concept of *internal consistency*—whether the model's explanation is faithful to its own predicted behaviour. To evaluate internal consistency, we define two evaluation metrics: *Rank correlation* and *Directional agreement.* Let $f(S)$ denote the model's original prediction and $f(S_{\text{pert}})$ the prediction after perturbation.

- *Rank correlation*: quantifies whether the influence ranking $\mathcal{I}(e_i)$ is aligned with the actual impact that each event $e_i$ has on the model's prediction when perturbed. Specifically, for each event, we compute $\Delta f(S_i) = f(S) - f(S_{pert})$ (e.g., event deletion, time shifting or marker flipping).

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}\big(\mathcal{I}(e_i) - \mathcal{I}(e_j)\big)\text{sgn}\big(\Delta f(S_i) - \Delta f(S_j)\big), \tag{25}$$

  where sgn$(\cdot)$ is the signum function. A high correlation value $\tau$ indicates that events have more influence and causes larger prediction shifts when perturbed—demonstrating internal consistency.

- *Directional agreement (DA)*: verifies whether masking high-influence events reliably leads to a decrease in predictive accuracy. This metric ensures that explanations align with the model's actual behaviour.

$$\text{DA} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[f(S_{\text{masked}}^{(i)}) < f(S^{(i)}) - \delta\right], \tag{26}$$

  where $\delta$ is a predefined significance threshold that accounts for minor prediction fluctuations due to randomness or noise. It ensures that only meaningful prediction drops—those that exceed $\delta$—are counted as valid directional changes. In practice, $\delta$ can be set based on a small proportion of the standard deviation of prediction scores across the dataset (e.g., $\delta = 0.01$ or $\delta = 5\%$). This avoids over-sensitivity to small variations and improves robustness of the directional agreement metric.

### 2.4.2 Fidelity to Real Data

To evaluate the fidelity of a model estimated on a real RED dataset, we propose two distinct evaluation methods: one for marker prediction and another for time forecasting.

Each method captures different aspects of alignment between the model's predictions and actual data, ensuring a comprehensive assessment of fidelity.

- *Marker prediction*: Logistic regression accuracy:

$$\text{Acc}_{\text{marker}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left[m_{n+1}^{I}t(i) = \text{argmax}\left(\sum_{j=1}^{n} w_j \mathcal{I}(e_j^{(i)})\right)\right]. \tag{27}$$

It measures how well the importance scores $\mathcal{I}(e_j^{(i)})$ can construct the actual markers through a logistic regression. A higher accuracy indicates stronger fidelity to real data, meaning the model's ability to capture meaningful patterns.

- *Time forecasting*: The Spearman correlation is defined as

$$\rho_t = \text{Spearman}\left(t_{n+1}, \sum_{j=1}^{n} \mathcal{I}(e_j)(t_{n+1} - t_j)\right). \tag{28}$$

It measures the rank correlation between the actual time of the next event $t_{n+1}$ and the aggregated influence-weighted time gaps $\sum_{j=1}^{n} \mathcal{I}(e_j)(t_{n+1} - t_j)$ [*]. A higher $\rho_t$ indicates that more influential past events tend to be temporally closer or more relevant to the future event.

Beyond these metrics, it is helpful to clarify the distinction between fidelity and interpretability, which reflect different goals of explanation. Fidelity evaluates whether attribution scores align with the model's actual behaviour under perturbations, while interpretability concerns how easily humans can understand the explanations (Lozano-Murcia et al., 2023). In our work, fidelity is assessed quantitatively through perturbation based metrics, while interpretability is illustrated qualitatively via the case study in Section 3.4.

With these evaluation metrics, we propose the following algorithm, as shown in Algorithm 1, to assess the interpretability of our SHPP model.

**Algorithm 1** Interpretability Evaluation Algorithm

```
Require: Event sequences $\{S^{(i)}\}_{i=1}^{N}$, model $f$, influence value $\mathcal{I}$
for each sequence $S^{(i)}$ do
    Compute influence values $\mathcal{I}(S^{(i)})$
    for perturbation $p \in \{\text{Delete}, \text{Shift}, \text{Flip}\}$ do
        Generate $S_{\text{pert}}^{(i)} \leftarrow p(S^{(i)})$
        Compute $\Delta f^{(i)} \leftarrow |f(S^{(i)}) - f(S_{\text{pert}}^{(i)})|$
    end for
    Compute reconstruction metrics $\text{Acc}_{\text{marker}}$ and $\rho_t$
    Aggregate $\tau$, DA, $\text{Acc}_{\text{marker}}$, and $\rho_t$
end for
```

# 3 Experimental Design and Results

We evaluate the proposed SHPP model across a wide range of RED from diverse domains, including environmental events, healthcare, e-commerce, and business processes. Our experiments assess both predictive performance (event time and marker) and the quality of influence-based explanations. We also perform ablation studies, statistical significance testing, and case-specific analysis.

Each dataset provides sequences of timestamped events labeled with categorical markers. See Table 5 for details on marker counts and domains. All datasets are split into 60% training, 20% validation, and 20% test sets.

## 3.1 Predictive Performance

We compare our proposed SHPP model with three representative neural TPP baselines from predictive performance perspective:

- *A-G*: A classical counting-process extension of the Cox proportional-hazards model for RED. It treats every RED as a new start–stop interval and estimates a common baseline hazard while

[*] Spearman's correlation captures the monotonic relationship between influence-weighted time gaps and actual event times. Formally, given two sequences $\{x_i\}$ and $\{y_i\}$, the Spearman correlation is computed as the Pearson correlation between their rank variables: $\rho = \frac{\text{Cov}(\text{rank}(x), \text{rank}(y))}{\sigma_{\text{rank}(x)}\sigma_{\text{rank}(y)}}$.

allowing time-varying covariates, thereby capturing event intensity without specifying self-excitation kernels (Andersen and Gill, 1982).

- *PWP*: A stratified Cox framework that orders RED by introducing one stratum per event number (gap-time or total-time variants). By conditioning on prior events within each stratum, PWP accounts for event order–specific baseline hazards and provides greater flexibility than A–G when event risk changes after each occurrence (Prentice et al., 1981).
- *RMTPP (Recurrent Marked Temporal Point Process)*: The first neural TPP model that uses recurrent neural networks (RNNs) to encode event history and predict both event time and marker. It captures sequential dependencies through hidden states and serves as a foundational deep learning-based TPP baseline (Du et al., 2016),
- *NHP (Neural Hawkes Process)* : An extension of Hawkes processes with continuous-time LSTM architecture, which extends RMTPP with a continuous-time LSTM and model time intervals better (Mei and Eisner, 2017), and
- *THP (Transformer Hawkes Process)*: A Transformer-based TPP model that employs self-attention to capture long-range dependencies across events. It supports flexible modelling of temporal influence patterns and has achieved state-of-the-art performance on several TPP benchmarks (Yang et al., 2021).

Table 1 displays the predictive performance measures of our proposed model SHPP against three other models[†]. The proposed SHPP model demonstrates competitive performance across multiple datasets in joint marker prediction and time forecasting tasks. As shown in Table 1, SHPP achieves the highest marker prediction accuracy (Acc) on 7 out of 10 datasets including Earthquake (0.481), Synthetic (0.409), ContTime (0.400), Mutual (0.646), Taxi (0.926), Amazon (0.362), and MIMIC-Diab. (0.826). These results highlight SHPP's ability in classification tasks across both scientific and operational domains.

Table 1: Predictive performance across datasets.

| Dataset | A–G | PWP | RMTPP | | NHP | | THP | | SHPP | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RMSE | RMSE | Acc | RMSE | Acc | RMSE | Acc | RMSE | Acc | RMSE |
| Earthquake | 10.214 | 8.107E3 | 0.441 | 1.742 | 0.472 | 1.988 | 0.472 | 1.863 | 0.481 | 1.838 |
| Synthetic | 15.183 | 3.447E3 | 0.381 | 0.612 | 0.381 | 0.606 | 0.382 | 0.557 | 0.409 | 0.604 |
| ContTime | 10.213 | 3.566E4 | 0.388 | 0.353 | 0.390 | 0.342 | 0.351 | 0.344 | 0.400 | 0.343 |
| Mutual | 15.519 | 6.831E3 | 0.379 | 1.702 | 0.633 | 1.219 | 0.628 | 1.164 | 0.646 | 1.504 |
| Taxi | 4.732 | 8.321E3 | 0.897 | 0.358 | 0.891 | 0.376 | 0.883 | 0.361 | 0.926 | 0.365 |
| Taobao | 1.512E5 | 1.454E5 | 0.436 | 0.269 | 0.512 | 0.332 | 0.436 | 0.297 | 0.436 | 0.259 |
| Amazon | 1.033E1 | 4.851E4 | 0.301 | 0.598 | 0.331 | 0.620 | 0.333 | 0.629 | 0.362 | 0.479 |
| BPIC | 3.483E2 | 8.251E4 | 0.435 | 6.788E1 | 0.662 | 8.497E2 | 0.412 | 6.956E1 | 0.413 | 6.835E1 |
| MIMIC-ICU | 3.982E4 | 4.943E4 | 0.502 | 1.736E3 | 0.881 | 1.734E3 | 0.894 | 2.293E3 | 0.882 | 1.736E3 |
| MIMIC-Diab. | 4.051E4 | 4.436E4 | 0.548 | 2.204E3 | 0.361 | 2.304E3 | 0.378 | 2.141E3 | 0.826 | 2.140E3 |

*Note: Acc* = marker classification accuracy (%, higher is better); *RMSE* = root mean squared error for timestamp prediction. A–G and PWP do not model markers explicitly, thus only RMSE is reported, and $aEb = a \times 10^b$.

In terms of time prediction (RMSE), SHPP outperforms all neural baselines on Taobao (0.259), Amazon (0.479), and MIMIC-Diabetes (2.140E3), and achieves competitive results on Mutual (1.504), where NHP and THP tend to suffer from instabilEity. On many datasets (e.g., Earthquake, Taxi), RMTPP achieves slightly lower RMSE, but with considerably worse marker accuracy, reflecting a trade-off.

SHPP achieves the most balanced performance on the Amazon dataset, attaining both the highest

[†] Accuracy not applicable to A–G and PWP as they do not support marker prediction

Accuracy (0.362) and the lowest RMSE (0.479), better than classical methods like A-G (RMSE: 10.325) and PWP (RMSE: 4.850E4) by a large margin.

On large-scale datasets such as BPIC, MIMIC-ICU, and MIMIC-Diab, SHPP remains competitive and stable, while classical models like PWP yield high RMSEs (e.g., BPIC: 8.251E4), indicating limited scalability of traditional statistical frameworks.

This performance comparison suggests that SHPP effectively balances event time prediction with marker classification. The consistent advantage in Accuracy across diverse domains indicates SHPP's enhanced modelling of marker-specific temporal dependencies and generalisation across heterogeneous datasets.

### 3.2 Attribution Analysis

We evaluate the internal consistency of SHPP and TimeSHAP (TimeS) across ten datasets in terms of marker attribution and event time attribution, using Kendall's $\tau$ rank correlation and Directional Agreement (DA), as shown in Table 2a, SHPP consistently outperforms TimeS in Kendall $\tau$ on both marker and time dimensions across most datasets.

Table 2: Comparison of SHPP and TimeSHAP on internal consistency (left) and fidelity (right).

| | (a) Internal consistency | | | | | | (b) Fidelity | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | Kendall $\tau$ (Marker) | | Kendall $\tau$ (Time) | | DA (%) | | Marker Acc (%) | | Time Spearman $\rho$ | |
| | SHPP | TimeS | SHPP | TimeS | SHPP | TimeS | SHPP | TimeS | SHPP | TimeS |
| Earthquake | 0.18±0.01 | 0.08 | 0.69±0.03 | 0.41 | 88.2±0.6 | 92.4 | 83.4±0.5 | 44.1 | 0.71±0.02 | 0.74 |
| Synthetic | 0.13±0.02 | 0.18 | 0.71±0.04 | 0.58 | 83.4±0.8 | 90.2 | 84.0±0.8 | 79.3 | 0.32±0.04 | 0.25 |
| ContTime | 0.40±0.01 | 0.12 | 0.43±0.02 | 0.29 | 98.0±0.2 | 92.7 | 59.2±0.1 | 54.8 | 0.37±0.05 | 0.29 |
| Mutual | 0.19±0.01 | 0.16 | 0.21±0.03 | 0.11 | 88.7±0.04 | 92.1 | 56.3±0.7 | 55.9 | 0.44±0.03 | 0.28 |
| Taxi | 0.62±0.01 | 0.29 | 0.65±0.02 | 0.20 | 82.1±0.4 | 95.3 | 89.2±0.4 | 83.7 | 0.78±0.03 | 0.69 |
| Taobao | 0.49±0.02 | 0.16 | 0.56±0.03 | 0.17 | 88.0±0.5 | 90.9 | 97.3±0.2 | 92.5 | 0.95±0.01 | 0.88 |
| Amazon | 0.17±0.03 | 0.09 | 0.21±0.04 | 0.25 | 64.0±1.0 | 49.2 | 84.1±0.6 | 79.9 | 0.32±0.03 | 0.26 |
| BPIC | 0.47±0.01 | 0.46 | 0.17±0.01 | 0.16 | 64.1±0.2 | 81.2 | 63.9±0.2 | 73.1 | 0.64±0.02 | 0.66 |
| MIMIC-ICU | 0.67±0.03 | 0.62 | 0.73±0.02 | 0.73 | 87.6±0.4 | 83.7 | 94.0±0.5 | 89.6 | 0.34±0.03 | 0.26 |
| MIMIC-Diab. | 0.76±0.03 | 0.72 | 0.74±0.03 | 0.76 | 88.0±0.3 | 90.1 | 95.1±0.3 | 74.6 | 0.43±0.04 | 0.47 |

*Note: Kendall* $\tau$ evaluates the rank correlation between original and perturbed importance rankings (higher is better); *DA* (Direction Agreement) indicates the consistency in influence direction after perturbation; *Acc* is classification accuracy of predicted event type (%); *Spearman* $\rho$ measures rank correlation on event timestamps (higher is better).

Compared to TimeSHAP, SHPP achieves higher Kendall $\tau$ for marker attribution in 9 out of 10 datasets and outperforms in time attribution in 9 out of 10 datasets as well. For example, on the Taxi dataset, SHPP attains a Kendall $\tau$ of 0.62 (marker) and 0.65 (time), significantly higher than TimeSHAP (0.29 and 0.20 respectively). Similarly, on the ContTime dataset, SHPP obtains $\tau = 0.40$ (marker) and $\tau = 0.43$ (time), while TimeSHAP only achieves 0.12 and 0.29. An exception is the Amazon dataset, where TimeSHAP slightly outperforms SHPP in time attribution ($\tau = 0.25$ vs. 0.21), suggesting that TimeSHAP can be more effective under sparse or low-signal settings.

Directional Agreement (DA) further supports the robustness of SHPP. On 3 of the 10 datasets, SHPP achieves significantly higher DA scores than TimeSHAP. Notably, on Mutual, SHPP maintains a DA of 88.7% vs. TimeS's 92.1%, while on Amazon, SHPP's DA is 64.0%, still higher than TimeSHAP (49.2%), despite the weaker $\tau$ score.

From a DA perspective, SHPP performs better than TimeSHAP on 3 of the 10 datasets: ContTime (98.0 % vs. 92.7 %), Amazon (64.0 % vs. 49.2 %), and MIMIC-ICU (87.6 % vs. 83.7 %). On the remaining datasets, TimeSHAP attains a higher DA, indicating that its attributions switch direction less often under perturbation. Notably, SHPP's advantage on Amazon arises despite a lower $\tau$ score, suggesting that even when rank correlation is weaker, its influence directions remain more coherent than those of TimeSHAP. These mixed outcomes highlight a trade-off: SHPP offers stronger direction consistency in certain domains, while TimeSHAP proves more robust in others.

Table 2b presents the fidelity evaluation results for SHPP and TimeSHAP (TimeS), focusing on two key dimensions: marker prediction accuracy and time attribution fidelity (Spearman's $\rho$). The results consistently demonstrate SHPP's ability to reproduce model behaviour under input perturbations across diverse datasets.

SHPP achieves notably high marker attribution fidelity, with accuracy ranging from 56.3% (Mutual) to 97.3% (Taobao), outperforming TimeSHAP on 9 out of 10 datasets. For example, on the Taxi dataset, SHPP achieves 89.2% accuracy versus TimeSHAP's 83.7%, and on Amazon, SHPP reaches 84.1% versus 79.9%. An exception is BPIC, where TimeSHAP slightly outperforms SHPP in marker accuracy (73.1% vs. 63.9%), potentially due to variance in process noise or annotation sparsity.

In terms of time attribution fidelity, SHPP also shows a consistent advantage, obtaining higher Spearman $\rho$ values in most datasets. Notably, on the Taobao dataset, SHPP achieves $\rho = 0.95$, exceeding TimeSHAP's $\rho = 0.88$, and on Taxi, SHPP records $\rho = 0.78$ versus TimeSHAP's $\rho = 0.69$. On MIMIC-Diab., however, TimeSHAP outperforms SHPP ($\rho = 0.47$ vs. 0.43), indicating marginally better alignment in medical event timing.

Overall, SHPP demonstrates robust fidelity across both attribution types, especially in datasets with strong sequential or behavioural signals (e.g., e-commerce and transportation). These results validate SHPP's effectiveness in approximating the model's true behaviour and underline its utility in high-stakes temporal modelling tasks.

### 3.3 Ablation Analysis

To assess how each component in SHPP contributes to both predictive performance and explanation ability, we perform an ablation analysis. Table 3 presents a detailed ablation study of the SHPP model across three representative datasets: Mutual, Taxi, and MIMIC-ICU. We examine the contribution of three influence components: *Self*-, *Markovian*, and *Joint*, by selectively removing each and measuring the impact on predictive performance, internal consistency, and fidelity.

Table 3: Ablation study on influence components across datasets.

| Dataset | Variant | Predictive | Internal Consistency | | Fidelity | |
|---|---|---|---|---|---|---|
| | | Acc / RMSE | $\tau$ (M / T) | DA (%) | Acc (%) | $\rho$ |
| Mutual | Full (S+M+J) | 0.646 / 15.450 | 0.19 / 0.21 | 88.7 | 56.3 | 0.44 |
| | –Self (M+J) | 0.378 / 15.448 | 0.08 / 0.06 | 57.5 | 52.1 | 0.28 |
| | –Markov (S+J) | 0.623 / 15.450 | 0.21 / 0.11 | 88.6 | 46.5 | 0.18 |
| | –Joint (S+M) | 0.623 / 15.450 | 0.09 / 0.00 | 88.6 | 54.8 | 0.25 |
| Taxi | Full (S+M+J) | 0.332 / 4.654 | 0.32 / 0.35 | 92.1 | 46.5 | 0.51 |
| | –Self (M+J) | 0.364 / 4.674 | 0.08 / 0.06 | 57.5 | 52.1 | 0.18 |
| | –Markov (S+J) | 0.133 / 4.655 | 0.21 / 0.11 | 88.6 | 46.5 | 0.08 |
| | –Joint (S+M) | 0.133 / 4.655 | 0.09 / 0.00 | 88.6 | 54.8 | 0.15 |
| MIMIC-ICU | Full (S+M+J) | 0.211 / 11.991 | 0.34 / 0.34 | 98.6 | 52.1 | 0.65 |
| | –Self (M+J) | 0.256 / 11.996 | 0.28 / 0.36 | 57.5 | 52.1 | 0.18 |
| | –Markov (S+J) | 0.111 / 11.991 | 0.21 / 0.41 | 98.6 | 46.5 | 0.08 |
| | –Joint (S+M) | 0.011 / 11.991 | 0.09 / 0.20 | 98.6 | 54.8 | 0.15 |

*Note:* Full = SHPP with all three influence components: Self (S), Markov (M), and Joint (J). –Self = without self-influence; –Markov = without Markovian influence –Joint = without joint influence.

Firstly, for predictive performance, the full model (S+M+J) consistently achieves the best or near-best accuracy and RMSE across datasets, indicating the importance of incorporating all three influence types. Removing the *Self* component (*–Self*) causes the most significant drop in accuracy (e.g., from 0.646 to 0.378 in *Muautl*), underscoring the critical role of self-influence in modelling event dependencies.

The impact of removing *Markovian* or *Joint* components is less severe in terms of accuracy, but still non-negligible.

Then, for internal consistency, the Kendall's $\tau$ scores and Directional Agreement (DA) show that eliminating *Self* or *Joint* components leads to degraded consistency in influence ranking. Notably, DA drops drastically to 57.5% in all datasets when *Self* is removed, confirming its central role in preserving stable influence attribution.

Finally, for fidelity, removing the *Joint* component (*–Joint*) slightly improves fidelity accuracy in some cases (e.g., 54.8% vs. 56.3% in *Mutual*), but this comes at the cost of reduced Spearman's $\rho$ (e.g., 0.44 to 0.25), suggesting temporal degradation. The *–Self* variant again performs the worst across all fidelity metrics, highlighting the importance of self-influence for both accurate and faithful explanations.

Overall, these findings demonstrate that: *Self-influence* is the most influential component for both prediction and explanation; *Markovian influence* improves consistency, particularly in recent interactions; *Joint influence* enhances the expressiveness of attributions, especially for capturing pairwise marker dependencies. The joint modelling of all three components enables SHPP to strike a desirable balance between predictive performance and interpretability.

## 3.4 Case Study: E-commerce Behaviour Analysis

In this section, we use the E-commerce dataset (Alibaba group, 2018; Zhuo et al., 2020), which contains time-stamped user click behaviours on Taobao.com from November 25 to December 03, 2017.

There are four marker types in the dataset:

- *pv*: Page view of an item's detail page (i.e., item click),
- *buy*: Purchase of an item,
- *cart*: Add an item to the shopping cart, and
- *fav*: Favor (bookmark) an item.

Each user has a sequence of events, with each event containing a timestamp and the item's category. To reduce the level of noise, we keep only the top 53 most frequent item categories. We then select a subset of 309,312 active users. After preprocessing, we retain $K = 4$ marker types. The dataset is split into training, development, and test sets with 68,950, 19,700, and 9,851 sequences, respectively.

Table 4: Predictive and interpretability metrics of SHPP for the case study.

| Perf. (Acc/RMSE) | $\tau$ (Marker/Time) | DA | Fid. Acc | Fid. $\rho$ |
|---|---|---|---|---|
| 92.02% / 181.99 | 0.624 / 0.638 | 0.980 | 94.00% | 0.648 |

Table 4 summarises the performance of SHPP across two key dimensions: prediction accuracy and temporal modelling fidelity, and explanation consistency under perturbations. The model achieves high marker classification accuracy (92.02%) and reasonably low timestamp error (RMSE = 181.99), demonstrating strong predictive performance. In terms of explanation quality, rank correlation $\tau$ and directional agreement show that the influence values are consistent with the model's predictive behaviour under perturbations. Furthermore, high marker reconstruction accuracy and Spearman correlation $\rho$ validate the fidelity of the learned representations in capturing true RED.

A specific case study is provided in the next section to illustrate the model's effectiveness on a real user sequence.

### 3.4.1 Understanding Behaviour Importance Value

To better understand how the model interprets user behaviours and identifies key decision points, we conduct a case study analysis on different user action routes, supported by influence value proposed

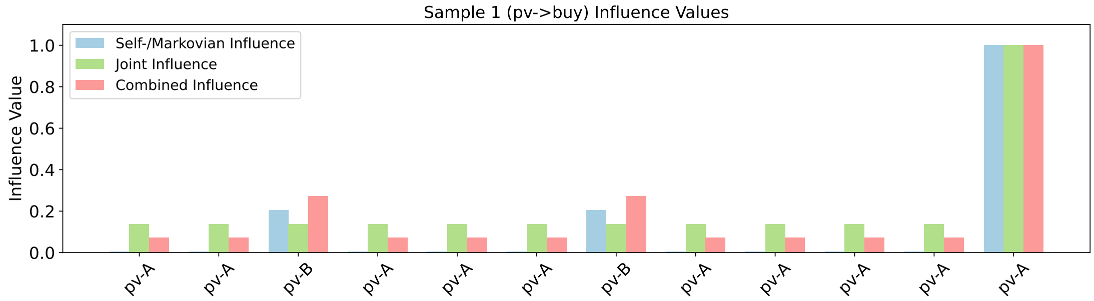

Figure 3: Sample 1 (pv → buy) influence values.

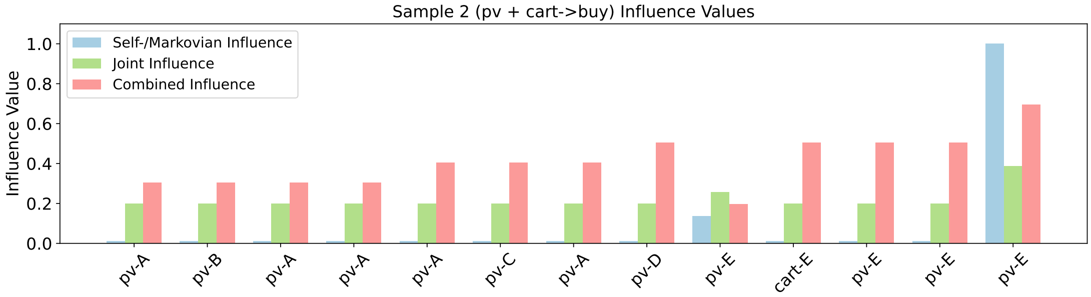

Figure 4: Sample 2 (pv + cart→buy) influence values.

in Section 2.3.2. We select three representative behaviour-to-purchase paths and analyse how the model assigns importance scores based on our three influences mechanism:

- *pv → buy:* This path represents users who make a purchase without any *fav* or *cart* actions,
- *pv + cart → buy:* Here, users directly add an item to the cart and later proceed to purchase, bypassing favoriting, and
- *pv + fav + cart → buy:* In this path, both *fav* and *cart* behaviours precede the final purchase.

We select several users with the previous representative behaviour-to-purchase path, which means the last behaviour is *buy*.

From Fig. 3, we observe that the final view of item A receives the highest influence value, while the views of item B also hold high influence value. This suggests that the user made the purchase decision through a comparative evaluation of similar items, and the last view of item A has the highest importance value, which influence most of the final decision: *buy* item A.

As shown in Fig. 4, the last browsing behaviour before purchase receives the highest influence value from three perspectives. During the user's ongoing comparison of similar products (e.g., item A, B, C, D, E), the combined influence value gradually increases. Notably, the *cart* action of item E itself does not carry the highest influence value; instead, it is the subsequent *post-cart* browsing behaviours that are more influential in the final purchase decision of item E.

Fig. 5 shows that the purchase of item A was influenced by recent views of similar items (e.g., item

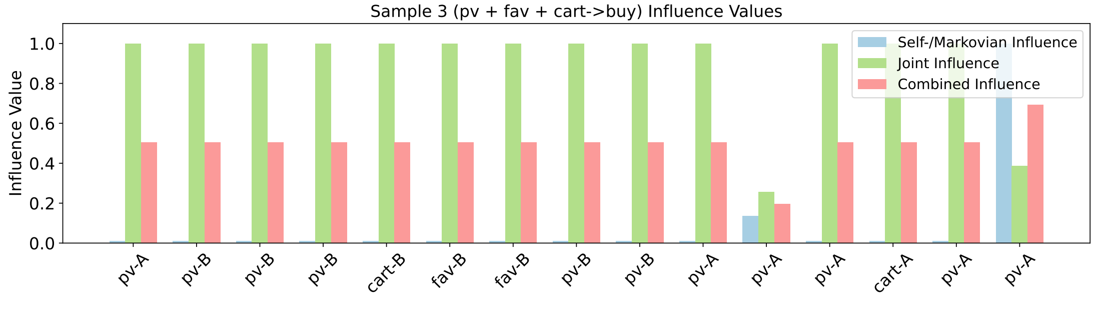

Figure 5: Sample 3 (pv + fav + cart→buy) influence value.

B). The influence value starts to vary only in the last five steps, with earlier actions having minimal influence. And the joint influence values from item B are almost the same (except 'pv-A', which is self-influence), which means the joint influence from B has no difference in the previous steps and have no influence for the final decision. The last self-influence from view of item A (e.g., self-influence) highly influence the final decision: *buy* item A.

In summary, the case studies illustrate that SHPP can generate user-level explanations that are not only interpretable but also actionable. This opens the door for more personalised recommendation strategies tailored to individual behavioural patterns—achieving the goal of customised recommendation for everyone.

## 4 Discussion and Limitations

Despite providing a structured and interpretable view of RED, SHPP still faces several practical limitations:

- *Data sparsity and scalability:* SHPP assumes moderately dense event histories. In scenarios with short sequences or, conversely, very long traces and many marker types, the model may underfit or suffer from sparse and noisy interactions. Pretraining, sequence augmentation, or marker grouping could help address these challenges.
- *Kernel flexibility:* The logistic–bilinear kernel is easy to interpret, yet its performance hinges on sensible basis-function choices and initialisation. Future work could adopt estimable monotone kernels or Bayesian priors that adapt shape while preserving interpretability.
- *Explainability coverage:* We report internal consistency and fidelity scores against *TimeSHAP*. A better method would require quantitative head-to-head tests with other XAI methods (e.g. attention heat-maps, Integrated Gradients) along XAI dimensions such as stability, completeness and robustness. Reducing this gap calls for a public benchmark for RED explainability—currently absent in the literature.
- *Modelling assumptions:* SHPP factorises an intensity into additive and pairwise terms. Domains with strong latent confounders or higher-order interactions may violate this assumption. Extending SHPP with latent variables, hierarchical strata, or graph priors could improve realism.
- *Computational efficiency:* We analyse SHPP's theoretical cost in Appendix A., but do not report running time and memory usage due to variability across environments. Potential optimisions in future work for large datasets and inference may includes: (i) history truncation beyond a temporal horizon, (ii) sparsification by pruning weak kernel entries, and (iii) low-rank compression of the joint influence matrix.

In our future work we plan to (i) introduce sparsity-aware regularisers to handle extremely sparse RED, (ii) build a unified benchmark that scores interpretability across multiple XAI metrics and baselines, including attention-based transformers, and (iii) develop online and multi-agent variants of SHPP for RED.

## 5 Conclusion

This paper introduced the *Stratified Hawkes Point Process (SHPP)*, an explainable temporal point process framework for modelling and interpreting recurrent event data. SHPP decomposes event dynamics into self-, Markovian, and joint influence components, enabling attribution of temporal dependencies across multiple event types.

By designing interpretable influence kernels and establishing sufficient stability conditions, SHPP balances predictive power with theoretical soundness and practical transparency. Extensive experiments demonstrate the model's effectiveness in both prediction and explainability tasks across diverse domains.

Overall, SHPP contributes a unified, interpretable, and extensible framework for explainable risk modelling, with potential applications in personalised recommendation, clinical monitoring, user behaviour analysis, and beyond.

# References

Alibaba group (2018). User Behavior Data from Taobao for Recommendation. `https://tianchi.aliyun.com/dataset/649`. [Online; accessed 03-April-2025].

Amorim, L. D. and Cai, J. (2015). Modelling recurrent events: a tutorial for analysis in epidemiology. *International Journal of Epidemiology*, 44(1):324–333.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, pages 1100–1120.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research*, 11(2):91–115.

Armstrong, J. S. (2001). Evaluating forecasting methods. *Principles of forecasting: A handbook for Researchers and Practitioners*, pages 443–472.

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. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115.

Bento, J., Saleiro, P., Cruz, A. F., Figueiredo, M. A., and Bizarro, P. (2021). Timeshap: Explaining recurrent models through sequence perturbations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2565–2573.

Borgonovo, E., Plischke, E., and Rabitti, G. (2024). The many Shapley values for explainable artificial intelligence: A sensitivity analysis perspective. *European Journal of Operational Research*, 318(3):911–926.

Cai, H., Nguyen, T. T., Li, Y., Zheng, V. W., Chen, B., Cong, G., and Li, X. (2020). Modeling marked temporal point process using multi-relation structure rnn. *Cognitive Computation*, 12:499–512.

Chang, C.-H., Tan, S., Lengerich, B., Goldenberg, A., and Caruana, R. (2021). How interpretable and trustworthy are gams? In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 95–105.

Chen, C.-M., Chuang, Y.-W., and Shen, P.-S. (2015). Two-stage estimation for multivariate recurrent event data with a dependent terminal event. *Biometrical Journal*, 57(2):215–233.

Cook, R. J., Lawless, J. F., et al. (2007). *The statistical analysis of recurrent events*. Springer.

Daley, D. J. and Vere-Jones, D. (2006). *An introduction to the theory of point processes: volume I: elementary theory and methods*. Springer Science & Business Media.

de Bock, K. W., Coussement, K., De Caigny, A., Słowiński, R., Baesens, B., Boute, R. N., Choi, T.-M., Delen, D., Kraus, M., Lessmann, S., et al. (2024). Explainable AI for operational research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*, 317(2):249–272.

Dewanji, A. and Moolgavkar, S. H. (2000). A poisson process approach for recurrent event data with environmental covariates. *Environmetrics: The Official Journal of the International Environmetrics Society*, 11(6):665–673.

Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., and Song, L. (2016). Recurrent marked temporal point processes: Embedding event history to vector. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1555–

1564.

Farajtabar, M., Du, N., Rodriguez, M., Valera, I., Zha, H., and Song, L. (2015). Shaping social activity by incentivizing users. In *Advances in Neural Information Processing Systems*, volume 28.

Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2019). Attention branch network: Learning of attention mechanism for visual explanation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10705–10714.

Gashi, M., Mutlu, B., and Thalmann, S. (2023). Impact of interdependencies: Multi-component system perspective toward predictive maintenance based on machine learning and xai. *Applied Sciences*, 13(5):3088.

Gupta, G., Sunder, V., Prasad, R., and Shroff, G. (2019). Cresa: a deep learning approach to competing risks, recurrent event survival analysis. In *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part II 23*, pages 108–122. Springer.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.

Higham, D. J. (2000). A-stability and stochastic mean-square stability. *BIT Numerical Mathematics*, 40:404–409.

Hu, X., Ma, W., Chen, C., Wen, S., Zhang, J., Xiang, Y., and Fei, G. (2022). Event detection in online social network: Methodologies, state-of-art, and evolution. *Computer Science Review*, 46:100500.

Kelly, P. J. and Lim, L. L.-Y. (2000). Survival analysis for recurrent event data: an application to childhood infectious diseases. *Statistics in Medicine*, 19(1):13–33.

Ketelbuters, L. and Bersini, H. (2022). Cds-hawkes: A causality-based hawkes process for event modeling and prediction. *European Journal of Operational Research*, 299(2):663–677.

Khasminskii, R. (2011). *Stochastic Stability of Differential Equations*. Springer.

Kobayashi, R. and Lambiotte, R. (2016). Tideh: Time-dependent hawkes process for predicting retweet dynamics. *Proceedings of the Tenth International AAAI Conference on Web and Social Media.*

Li, P., Bahri, O., Boubrahimi, S. F., and Hamdi, S. M. (2023). Attention-based counterfactual explanation for multivariate time series. In *International Conference on Big Data Analytics and Knowledge Discovery*, pages 287–293. Springer.

Lin, H., Wu, L., Zhao, G., Liu, P., and Li, S. Z. (2022). Exploring generative neural temporal point process. *arXiv preprint arXiv:2208.01874*.

Lintu, M. and Kamath, A. (2022). Performance of recurrent event models on defect proneness data. *Annals of Operations Research*, 315(2):2209–2218.

Lozano-Murcia, C., Romero, F. P., Serrano-Guerrero, J., and Olivas, J. A. (2023). paparison Between Explainable Machine Learning Methods for Classification and Regression Problems in the Actuarial Context. *Mathematics*, 11(14):3088.

Lundberg, S. M. and Lee, S. (2017). A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Lyu, Q. and Wu, S. (2025). Explainable artificial intelligence for business and economics: Methods, applications and challenges. *Expert Systems*, 42(4):e70017.

Mei, H. and Eisner, J. M. (2017). The neural hawkes process: A neurally self-modulating multivariate point process. *Advances in Neural Information Processing Systems*, 30.

Miró-Nicolau, M., Jaume-i Capó, A., and Moyà-Alcover, G. (2024). Assessing fidelity in xai post-hoc techniques: A comparative study with ground truth explanations datasets. *Artificial Intelligence*, 335:104179.

Miró-Nicolau, M., Jaume-i Capó, A., and Moyà-Alcover, G. (2025). A Comprehensive Study on Fidelity Metrics for XAI. *Information Processing & Management*, 62(1):103900.

Murris, J., Bouaziz, O., Jakubczak, M., Katsahian, S., and Lavenu, A. (2024). Random survival forests for the analysis of recurrent events for right-censored data, with or without a terminal event. working paper or preprint.

Narteau, C., Shebalin, P., and Holschneider, M. (2002). Temporal limits of the power law aftershock decay rate. *Journal of Geophysical Research: Solid Earth*, 107(B12):ESE–12.

Novaković, J. D., Veljović, A., Ilić, S. S., Papić, Ž., and Tomović, M. (2017). Evaluation of classification mod els in machine learning. *Theory and Applications of Mathematics & Computer Science*, 7(1):39.

Oganisian, A., Girard, A., Steingrimsson, J. A., and Moyo, P. (2024). A bayesian framework for causal analysis of recurrent events with timing misalignment. *Biometrics*, 80(4):ujae145.

Oyamada, S., Chiu, S.-W., and Yamaguchi, T. (2022). Comparison of statistical models for estimating intervention effects based on time-to-recurrent-event in stepped wedge cluster randomized trial using open cohort design. *BMC Medical Research Methodology*, 22(1):123.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379.

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning*, 1:81–106.

Rajpal, S., Rajpal, A., Saggar, A., Vaid, A. K., Kumar, V., Agarwal, M., and Kumar, N. (2023). Xai-methylmarker: Explainable AI approach for biomarker discovery for breast cancer subtype classification using methylation data. *Expert Systems with Applications*, 225:120130.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016a). " Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016b). Model-Agnostic Interpretability of Machine Learning. *arXiv: 1606.05386.*

Rizoiu, M.-A., Lee, Y., Mishra, S., and Xie, L. (2017). *Hawkes processes for events in social media.* ACM Books.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Schmidt-Hieber, J. (2021). The kolmogorov–arnold representation theorem revisited. *Neural Networks*, 137:119–126.

Shchur, O., Türkmen, A. C., Januschowski, T., and Günnemann, S. (2021). Neural temporal point processes: A review. *arXiv preprint arXiv:2104.03528.*

Speith, T. (2022). A Review of Taxonomies of Explainable Artificial Intelligence (XAI) Methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 2239–2250.

Stevens, A. and De Smedt, J. (2024). Explainability in process outcome prediction: Guidelines to obtain interpretable and faithful models. *European Journal of Operational Research*, 317(2):317–329.

Topuz, K., Urban, T. L., and Yildirim, M. B. (2024). A markovian score model for evaluating provider performance for continuity of care—an explainable analytics approach. *European Journal of Operational Research*, 317(2):341–351.

Tsirtsis, S., De, A., and Rodriguez, M. (2021). Counterfactual explanations in sequential decision making under uncertainty. *Advances in Neural Information Processing Systems*, 34:30127–30139.

Watson, V., Tudur Smith, C., and Bonnett, L. (2020). Protocol for a systematic review of prognostic models for recurrent events in chronic conditions. *Diagnostic and Prognostic Research*, 4:1–6.

Wiegreffe, S. and Pinter, Y. (2019). Attention Is Not Not Explanation. 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)*, pages 11–20. Association for Computational Linguistics.

Wu, S. (2012). Warranty data analysis: A review. *Quality and Reliability Engineering International*, 28(8):795–805.

Xu, H., Farajtabar, M., and Zha, H. (2016). Learning granger causality for hawkes processes. In *International Conference On Machine Learning*, pages 1717–1726. PMLR.

Xue, S., Shi, X., Chu, Z., Wang, Y., Hao, H., Zhou, F., Jiang, C., Pan, C., Zhang, J. Y., Wen, Q., Zhou, J., and Mei, H. (2024). Easytpp: Towards open benchmarking temporal point processes. In *International Conference on Learning Representations (ICLR)*.

Xue, S., Shi, X., Zhang, J., and Mei, H. (2022). Hypro: A hybridly normalized probabilistic model for long-horizon prediction of event sequences. *Advances in Neural Information Processing Systems*, 35:34641–34650.

Yang, C., Mei, H., and Eisner, J. (2021). Transformer embeddings of irregularly spaced events and their participants. *arXiv preprint arXiv:2201.00044*.

Zhou, K., Zha, H., and Song, L. (2013). Learning triggering kernels for multi-dimensional hawkes processes. In *International Conference on Machine Learning*, pages 1301–1309. PMLR.

Zhuo, J., Xu, Z., Dai, W., Zhu, H., Li, H., Xu, J., and Gai, K. (2020). Learning optimal tree models under beam search. In *International Conference on Machine Learning*, pages 11650–11659. PMLR.

## Appendix A. Optimisation Framework

Having established the theoretical foundations of SHPP with interpretable kernels, we now turn to the parameter estimation problem.

Let $\{t_i\}_{i=1}^N$ be the event times in observation window $[0,T]$ with associated markers $\{m_i\}_{i=1}^N$. Define that $\Delta N_j(t_{i-1},t) \overset{\text{def}}{=} |\{s : t_{i-1} < t_s \leq t, m_s = j\}|$, which represents event with marker of type $j$ count in $(t_{i-1},t]$. The conditional intensity function can be decomposed as: $\lambda(t) = \sum_{j=1}^M \lambda_j(t \mid \mathcal{H}_t)$. Thus, the distributions for event time are:

$$F(t \mid \mathcal{H}_{t_{i-1}}) = 1 - \exp\left(-\int_{t_{i-1}}^t \lambda(s)ds\right), \quad f(t \mid \mathcal{H}_{t_{i-1}}) = \lambda(t)\exp\left(-\int_{t_{i-1}}^t \lambda(s)ds\right). \tag{29}$$

The marker type's distribution satisfies: $\mathbb{P}(M_i = j \mid T_i = t) = \frac{\lambda_j(t)}{\lambda(t)}$. Then, the joint likelihood over $[0,T]$ decomposes as:

$$\mathcal{L}(\Theta) = \prod_{i=1}^N f(t_i)\mathbb{P}(m_i \mid t_i) = \prod_{i=1}^N \lambda_{m_i}(t_i)\exp\left(-\int_{t_{i-1}}^{t_i} \lambda(s)ds\right). \tag{30}$$

We can then obtain the log-likelihood:

$$
\begin{aligned}
\ell(\Theta) &= \sum_{i=1}^{N} \log \lambda_{m_i}(t_i) - \int_0^T \lambda(t)dt \\
&= \underbrace{\sum_{i=1}^{N} \left[ \mu_{m_i} + \sum_{j=1}^{M} \gamma_{m_i j}(\{t_i - T_{js}\}) \right]}_{\text{Event Term}} - \underbrace{\int_0^T \exp\left( \sum_{j=1}^{M} \mu_j + \sum_{\ell=1}^{M} \gamma_{j\ell}(\{t - T_{\ell m}\}) \right) dt}_{\text{Non-Event Term}}.
\end{aligned} \tag{31}
$$

To improve the computation efficiency of the non-event integral term, which is often computationally expensive due to its dependence on the entire event history, we propose an adaptive Monte Carlo optimisation approach that leverages adaptive sampling to reduce variance in non-event integral estimation while maintaining computational efficiency.

**Algorithm 2** Adaptive Monte Carlo Optimisation

```
Initialize parameters $\Theta^{(0)} = \{\mu_j, \gamma_{jk}\}$
for epoch = 1 to $E$ do
    Shuffle event sequences
    for each mini-batch $\mathcal{B}$ do
        Compute event term: $\ell_{\text{event}} = \sum_{(t_i,m_i)\in\mathcal{B}} \log \lambda_{m_i}(t_i)$
        Estimate non-event term: $\ell_{\text{non-event}} \approx \frac{T}{S} \sum_{s=1}^{S} \lambda(t_s)$ where $t_s \sim \text{AdaptiveSampler}(\lambda)$
        Compute gradient: $\nabla\Theta = \nabla(\ell_{\text{event}} - \ell_{\text{non-event}})$
        Update: $\Theta^{(\text{new})} \leftarrow \Theta^{(\text{old})} + \eta\text{Adam}(\nabla\Theta)$
    end for
end for
```

We implement SHPP using PyTorch and optimise it using the Adam optimizer with a learning rate of $10^{-3}$ and batch size of 64. The kernel functions $\gamma_{ij}$ are parameterised by neural basis expansions (see Eq. (21)), and all parameters including coefficients $\beta$, $\theta$, and adaptive weight $\alpha$ are jointly learned via backpropagation. Regularisation is applied via $\ell_2$-norm penalties to avoid overfitting in sparse regimes. Training typically converges within 50 epochs.

The computational complexity of SHPP depends on the number of historical events and the complexity of the kernel evaluations. Specifically, the per-event computation cost is $\mathcal{O}(N_j(t)\cdot d + N_j(t)^2 \cdot d^2)$, where $N_j(t)$ is the number of historical events of type $j$, and $d = p + 1$ is the feature dimension. The first term accounts for self-/Markovian type influence, while the second corresponds to joint influence over all event pairs.

## Appendix B. Experimental Datasets and Setup

There are several recurrent event datasets that have been prepared by our proposed SHPP, as shown in Table 5

Table 5: Overview of recurrent event datasets used in experiments.

| Data name | Scenario | Data description | Artificial? | Marker # | Size | Resource |
|---|---|---|---|---|---|---|
| Earthquake | Environmental | Timestamped earthquake events over the U.S. (1996–2023) | No | 1 | 49363 | USGS |
| Synthetic | Generic simulation | Data simulated based on Hawkes process | No | 1 | 8000 | Tick library |
| Conttime | Generic simulation | Data simulated based on continuous-time Hawkes process | No | 1 | 8000 | Tick library |
| Mutual | Generic simulation | Data simulated based on mutual-exciting process | No | 2 | 8000 | tick library |
| Taxi | Transportation | Timestamped taxi pick-up events | Yes | 10 | 51854 | NYC FOIL |
| Taobao | E-commerce | User online shopping behaviour on Taobao.com | Yes | 17 | 75205 | Xue et al. (2022) |
| Amazon | E-commerce | User product review behaviour (2008–2018) | Yes | 16 | 6454 | Amazon data |
| BPIC | Finance | Business process logs from Dutch financial institution | Yes | 26 | 10000 | BPIC2017 |
| MIMIC-Diab. | Healthcare | Hospital events for people with diabetes from MIMIC-IV | Yes | 11 | 25593 | MIMIC-IV |
| MIMIC-ICU | Healthcare | Hospital events for people in ICU from MIMIC-IV | Yes | 11 | 65366 | MIMIC-IV |

All experiments are developed in PyTorch and run on a machine with NVIDIA A40 GPU. We use a standard train-validation-test split of 60%-20%-20% across all datasets unless otherwise specified. For each experiment, we run 5 different random seeds and report mean ± standard deviation.

For SHPP, we set the maximum number of past events $K = 10$, kernel function $\kappa(\tau) = \alpha e^{-\beta\tau}$, and use default base functions $g_r(\cdot)$ (logistic) and $h_{ru}(\cdot,\cdot)$ (bilinear). The learning rate is set to $10^{-3}$, batch size 64, and we use the Adam optimizer with early stopping on validation loss.

Evaluation metrics include prediction accuracy (marker and time), Kendall's $\tau$, direction agreement (DA), and fidelity scores. See Appendix C. for kernel stability assumptions.

## Appendix C. Kernel Types and Stability Conditions

Here, we summarise commonly used temporal kernels for point processes and discuss whether they satisfy the stability condition proposed in Theorem 1.

Let $\tau(= t - t_k)$ denote the time gap between the current and historical events. The kernels listed in Table 6 are widely used in temporal modelling: Theorem 1 states that a sufficient condition for mean-square stability is that the kernel function $\gamma(\tau)$ satisfies $\partial\gamma/\partial\tau \le 0$ for all $\tau > 0$. This guarantees that the cumulative influence does not diverge over time.

In our implementation, we use exponential-based kernels for both excitation and inhibition due to their stability and analytical simplicity.

Table 6: Common kernel types, properties, and stability under Theorem 1.

| **Kernel Type** | **Form $\gamma(\tau)$** | **Monotonic?** | **Stable?** | **Reference** |
|---:|---|---|---|---|
| Exponential decay | $\alpha e^{-\beta\tau},\ \alpha{>}0$ | Yes | Yes | Hawkes (1971) |
| Gaussian-shaped | $\alpha e^{-\beta(\tau-\mu)^2}$ | No | No | Zhou et al. (2013) |
| Rayleigh | $\alpha\tau e^{-\beta\tau^2}$ | No | No | Farajtabar et al. (2015) |
| Power-law | $\frac{\alpha}{(\tau+c)^\delta},\ \delta{>}1$ | Yes | Yes | Narteau et al. (2002) |
| Signed exponential | $\alpha e^{-\beta\tau},\ \alpha{<}0$ | Yes | Yes | Kobayashi and Lambiotte (2016) |

## Appendix D. Sensitive Analysis

To assess the robustness and flexibility of SHPP, we conduct a series of sensitivity analyses using synthetic datasets. Specifically, we investigate: (i) The impact of the influence balance parameter $\alpha$, which balances historical events influences (see Table 7), (ii) The role of different types of base functions in the interpretable kernel (see Table 8), and (iii) The effect of varying the number of event marker types on performance and explanation performance (see Table 9).

We provide a synthetic data generation algorithm for the marker types sensitive analysis (Algorithm 3).

Table 7: Sensitivity of SHPP to the influence weight $\alpha$.

| $\alpha$ | Kendall $\tau$ | | DA (%) | Fidelity | |
|---|---|---|---|---|---|
| | Marker | Time | Value | Acc (%) | $\rho$ |
| 0.1 | 0.22 | 0.05 | 88.1 | 17.3 | 0.29 |
| 0.3 | 0.21 | 0.11 | 88.2 | 17.3 | 0.38 |
| 0.5 | 0.23 | 0.12 | 87.9 | 17.3 | 0.33 |
| 0.7 | 0.14 | 0.11 | 89.0 | 19.3 | 0.33 |
| 0.9 | 0.18 | 0.05 | 87.5 | 17.3 | 0.38 |

From Table 7, we observe that internal consistency metrics (Kendall's $\tau$) improve as $\alpha$ increases from 0.1 to 0.5, suggesting that a moderate emphasis on influence structure helps stabilise importance estimation. Beyond $\alpha = 0.5$, the consistency drops slightly, possibly due to over-regularisation. Directional agreement (DA) remains stable across all settings, while fidelity (Acc and $\rho$) peaks near $\alpha = 0.7$, indicating an optimal trade-off between self- and pairwise contributions.

From Table 8, the combination of Logistic encoding with Bilinear interaction has the best overall fidelity and consistency scores. Decision Stump + Bilinear performs competitively, while shallow neural

**Algorithm 3** Simulated RED Generation Algorithm

```
Require: Number of sequences $N$, event types $K$, max time $T_{\max}$, baseline intensity $\mu_i$, kernel $\gamma_{ij}(\cdot)$, noise level $\sigma_t$, perturbation probability $p$
for $n = 1$ to $N$ do
    Initialise event list $S^{(n)} \leftarrow \emptyset$
    Set current time $t \leftarrow 0$
    while $t < T_{\max}$ do
        Compute intensity $\lambda_i(t) = \exp(\mu_i + \sum_{j=1}^{K} \sum_{t_k<t} \gamma_{ij}(t - t_k))$
        Sample next time gap $\Delta t \sim \sum_i \lambda_i(t)$
        Update time: $t \leftarrow t + \Delta t$
        Sample event type $m \sim \text{Multinomial}(\lambda_1(t), \ldots, \lambda_K(t))$
        Add $(t, m)$ to $S^{(n)}$
    end while
    /* Add perturbations */
    for $(t_i, m_i) \in S^{(n)}$ do
        $t_i \leftarrow t_i + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_t^2)$          ▷ Timestamp noise
        if Rand() $< p$ then
            $m_i \leftarrow \text{UniformRandom}(1, 2, \ldots, K)$          ▷ Marker flipping
        end if
    end for
end for
```

Table 8: Sensitivity to the choices of base functions.

| Basis Function | Kendall $\tau$ | | DA (%) | Fidelity | |
|---|---|---|---|---|---|
| | Marker | Time | Value | Acc (%) | $\rho$ |
| Logistic + Bilinear | 0.12 | 0.22 | 87.9 | 20.3 | 0.31 |
| Decision Stump + Bilinear | 0.24 | 0.23 | 83.4 | 19.9 | 0.27 |
| Logistic + Shallow NN | 0.11 | 0.21 | 77.0 | 18.8 | 0.27 |
| Shallow Tree + Tree Interact. | 0.08 | 0.12 | 74.8 | 16.8 | 0.28 |

*Note:* Basis functions used in SHPP are defined as follows: (1) *Logistic*: $\phi(x) = \frac{1}{1+\exp(-w^\top x)}$; (2) *Decision Stump*: binary indicator $\phi(x) = \mathbb{I}(x_j > \theta)$ for some feature $j$ and threshold $\theta$; (3) *Shallow NN*: one hidden layer neural network $\phi(x) = \sigma(W_2 \cdot \sigma(W_1 x + b_1) + b_2)$; (4) *Tree Interaction*: pairwise indicator features from a shallow decision tree. Bilinear or additive forms are used for modelling interactions among events.

Table 9: Sensitivity to number of marker types.

| # Markers | Kendall $\tau$ | | DA (%) | Fidelity | |
|---|---|---|---|---|---|
| | Marker | Time | Value | Acc (%) | $\rho$ |
| 5 | 0.22 | 0.21 | 82.6 | 33.6 | 0.21 |
| 10 | 0.22 | 0.21 | 90.0 | 17.4 | 0.36 |
| 20 | 0.11 | 0.22 | 94.8 | 8.3 | 0.24 |
| 40 | 0.13 | 0.22 | 98.0 | 4.3 | 0.33 |

nets and tree-based designs slightly reduce interpretability metrics. This confirms that simple yet expressive base functions align better with SHPP's structured assumptions.

When selecting a base function, we suggest starting with a small number of logistic units plus bilinear terms. If the application needs rule-level transparency, switching the logistic units to a small number of decision stumps provides clearer if–then statements at the cost of less fidelity. Only when data are large enough and highly non-linear interactions are expected should one consider shallow neural or tree-interaction bases, while Directional Agreement will drop.

From Table 9, fidelity metrics (especially Fid. and Acc) degrade noticeably, though DA improves, as the number of marker types increases from 5 to 40. This suggests that SHPP maintains relative ordering of influences even under complex event marker types, but the absolute attribution becomes less precise. These results highlight the challenge of interpretability under high-dimensional settings, motivating future work on scalable regularisation or clustering-based summarisation.