Deviation-based Marked Temporal Point Process for Marker Prediction

Anand Vir Singh Chauhan, Shivshankar Reddy, Maneet Singh (✉), Karamjit Singh, and Tanmoy Bhowmik

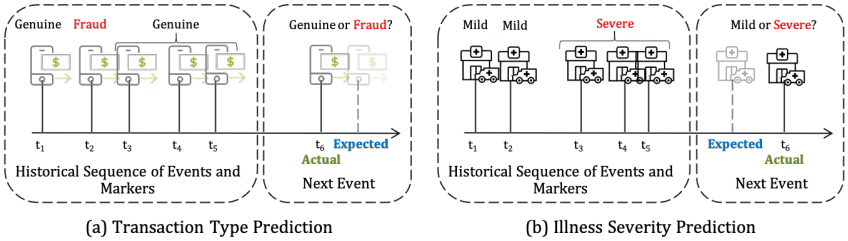
AI Garage, Mastercard, India
{anandvirsingh.chauhan, shivshankar.reddy, maneet.singh, karamjit.singh, tanmoy.bhowmik}@mastercard.com

Abstract. Temporal Point Processes (TPPs) are useful for modeling event sequences which do not occur at regular time intervals. For example, TPPs can be used to model the occurrence of earthquakes, social media activity, financial transactions, etc. Owing to their flexible nature and applicability in several real-world scenarios, TPPs have gained wide attention from the research community. In literature, TPPs have mostly been used to predict the occurrence of the next event (time) with limited focus on the *type/category* of the event, termed as the *marker*. Further, limited focus has been given to model the inter-dependency of the event time and marker information for more accurate predictions. To this effect, this research proposes a novel Deviation-based Marked Temporal Point Process (DMTPP) algorithm focused on predicting the marker corresponding to the next event. Specifically, the deviation between the estimated and actual occurrence of the event is modeled for predicting the event marker. The DMTPP model is explicitly useful in scenarios where the marker information is not known immediately with the event occurrence, but is instead obtained after some time. DMTPP utilizes a Recurrent Neural Network (RNN) as its backbone for encoding the historical sequence pattern, and models the dependence between the marker and event time prediction. Experiments have been performed on three publicly available datasets for different tasks, where the proposed DMTPP model demonstrates state-of-the-art performance. For example, an accuracy of 91.76% is obtained on the MIMIC-II dataset, demonstrating an improvement of over 6% from the state-of-the-art model.

Keywords: Temporal Point Processes · Marker Prediction · Recurrent Neural Network.

1 Introduction

The developments in technology and fast-paced lifestyle have resulted in the generation of large amount of temporal data containing *events* spanned across irregular time intervals. For example, activity on social media such as uploading images, post reacts, interactions with other users; utilizing public transportation



(a) Transaction Type Prediction (b) Illness Severity Prediction

Fig. 1: Event marker prediction has wide-spread applicability in various real-world scenarios. The proposed Deviation-based Marked Temporal Point Process model focuses on predicting the marker of an event in real-time, while modeling the inter-dependence between the expected event time and the actual event time.

such as cabs, taxis, or buses; financial activity such as buying/selling stocks, online purchases; and dining out at restaurants or reviewing eating joints. Coupled with the advent of Machine Learning and the day-to-day usage of different deployed applications, developing algorithms for automated event prediction has garnered substantial research attention. Traditionally, event prediction referred to determining *when* the next event would happen. With several recent real-world applications, research has also focused on predicting the *type* of the event referred to as the *marker* corresponding to an event. Fig. 1 presents sample real-world applications requiring event type prediction (often in real-time). Fig. 1(a) presents a sample scenario where banks could utilize algorithms to identify whether the current transaction (event) was fraudulent or not (marker), and Fig. 1(b) presents another scenario where hospitals could identify the duration or severity (marker) of a patient's visit (event) based on their historical information. Event marker prediction thus has wide applicability in real-world scenarios across different domains, demanding dedicated attention.

Initial research on event prediction [21,27] utilized statistical techniques [2], followed by modeling the sequences as time series [12]. While earlier research focused primarily on events spaced evenly in time, as discussed previously, most of the above mentioned activities are uneven or irregular in terms of the inter-event time. The uneven characteristic of event sequences makes it appropriate to model them as Temporal Point Processes (TPP) [10,18], often defined by an intensity function modeling the inter-event duration. Generally, historical sequences are modeled to predict the occurrence of the next event, and a categorical value associated with it, referred to as the event marker. While event time prediction has been well studied in the past few years, limited attention has been given to the task of marker prediction. To this effect, this research proposes a novel *Deviation-based Marked Temporal Point Process (DMTPP)* model for predicting the event marker. The proposed model is specifically applicable in scenarios where the event marker is not available immediately after the event occurrence, but is instead computed/obtained after some time. For example, a fraudulent transaction (marker) might be reported after some time of the trans-

action (event) by the concerned person (Fig. 1(a)), the severity of an illness (marker) is not known upon the immediate admission of a patient (event) into a hospital (Fig. 1(b)), and the impact of an online advertisement (event) on the subsequent sales (marker) is known after some time. By utilizing the real-time event occurrence, the DMTPP model presents high applicability in such scenarios, where the marker prediction is also performed in real-time.

The proposed DMTPP model focuses on explicitly modeling the inter-dependence between the marker prediction and the variation observed in the expected behavior. This enables the model to capture *anomalous* behavior with respect to the event occurrence in real-time, while utilizing the representation from the historical event sequence. Therefore, the contributions of this research are:

- A novel Deviation-based Marked Temporal Point Process (DMTPP) model has been proposed for marker prediction in real-time. The DMTPP algorithm models the dependence of the marker on the expected and actual time occurrence of the next event. To the best of our knowledge, this is the first-of-a-kind model operating at real time, which explicitly models the dependence of the marker on the deviation in the expected and actual event time.
- The DMTPP model utilizes a Recurrent Neural Network (RNN) as its backbone architecture. The RNN learns an embedding based on the past sequence of events and markers, while modeling the intensity function of the TPP as a non-linear function. The choice of RNN as the backbone architecture provides more flexibility during sequence modeling, and also prevents learning of *user-specific* models/representations. Thus enabling the proposed DMTPP model to be useful in real-world scenarios of unseen test users as well.
- The proposed model has been evaluated on three marker prediction tasks: (i) retweet prediction on the Retweet dataset [28], (ii) illness type prediction on the MIMIC-II dataset [16], and (iii) badge prediction on the StackOverflow dataset [4]. Comparison has been performed with recent state-of-the-art methods, where the proposed model demonstrates improved performance. For example, it achieves a classification accuracy of 91.76% on the challenging MIMIC-II dataset. The improved performance promotes the utility of the proposed model for real-time marker prediction tasks.

2 Related Work

This section analyzes the related concepts and research in the area of marked temporal point process. Marked temporal point processes build upon the traditional temporal point processes by associating a *marker* with the occurrence of each event. Here, *marker* can refer to the category of the event or some additional information of the event that is mostly categorical in nature. Research in marked temporal point processes has focused on the next event and marker prediction based on the sequence of historical events which is measured by an intensity function. The intensity function measures the number of events that can be expected in a specific time interval.

The traditional methods in temporal point processes like Poisson [11], Self-exciting [6], and Self-correcting [9] processes estimate the conditional intensity function by making parametric assumptions. Such assumptions limit the flexibility of the model, thus making it challenging to apply in different real-world scenarios. The more recent models which utilize deep learning algorithms train the models by maximizing the log-likelihood of the desired loss function. One of the seminal algorithms at the intersection of marked temporal point processes and deep learning is the Recurrent Marked Temporal Point Processes (RMTPP) model [3], which uses a neural network to predict both next event time and event marker independently using a Recurrent Neural Network (RNN). Following this, Wang et al. [22] proposed an RNN network to build a marker-specific intensity function that considers the inter-dependency between the marker and time of the next event. Beyond RNNs, in 2018, Decoupled Learning for Factorial Marked Temporal Point Processes [23] was proposed, where a decoupling approach is presented for learning the factorial marked temporal point process, in which each event is represented by multiple markers. Recently, Türkmen et al. [20] leveraged both Hawkes processes and RNN to capture local and global temporal relationships. Shchur et al. [19] proposed a novel approach of using neural density estimation to estimate the conditional density instead of modeling the conditional intensity function. Further, in 2020, Transformer Hawkes Process (THP) [29] and Self-Attentive Hawkes Process (SAHP) [26] addressed the problem of long-term dependencies by using a self-attention mechanism to capture short-term and long-term dependencies in the past sequence of the event.

As demonstrated above, the field of marked temporal point processes has recently garnered substantial attention. TPPs have shown applicability in several real-time applications, and are successful in capturing the influence of past historical sequence information for the prediction of the next event time and marker. However, in most of the existing literature (Fig. 2(a)-(b)), the event time and marker are assumed to be independent, which might not be true in real-time applications where the event time and marker are inter-dependent. For example, as shown in Fig. 1(a), in scenarios of fraudulent transaction detection, a given transaction is required to be classified as fraudulent or not. In this scenario, unusual occurrence (time) of the actual event as compared to the predicted time by the learned model can help in identifying a fraudulent transaction. Similarly, in other domain applications such as predicting visits to the Intensive Care Unit (ICU) (Fig. 1(b)), illness severity prediction can be dependent on the deviation between the next predicted event time and the actual event time. Similar trend can be observed in the scenario of online advertisements, where variation between the actual and expected time of posting can often result in variation of the next marker type (impact of advertisement measured by subsequent sales).

Based on the above intuition, this research proposes utilizing the deviation in the predicted event time and actual event time for predicting the event marker. A novel Deviation-based Marked Temporal Point Process (DMTPP) model is proposed, which incorporates the inter-dependence between the event time and

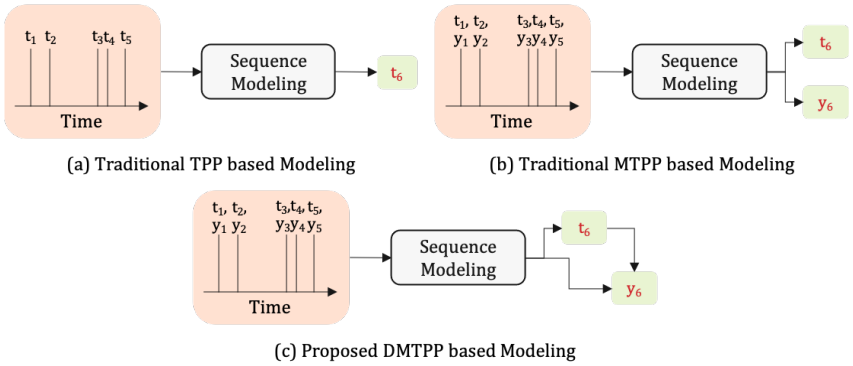


Fig. 2: Existing literature in TPP based modeling has focused mostly on (a) predicting the next event time (t_6), or (b) predicting the next event and marker information (t_6, y_6) independently. (c) The proposed DMTPP based model learns the dependence of the marker information on the time prediction.

marker (Fig. 2(c)) by considering the real-time event occurrence information for predicting the next event marker.

3 Proposed Algorithm

Fig. 3 presents a diagrammatic overview of the proposed Deviation-based Marked Temporal Point Process (DMTPP) model. The proposed model takes the historical sequence of time and marker, and predicts the marker for the next event. Further, the model also utilizes the actual time of the next event for predicting the corresponding marker. The proposed model thus demonstrates high applicability in scenarios where the marker is computed/derived after some time as opposed to being simultaneously available with the event occurrence. As shown in Fig. 3, a RNN based architecture is used for modeling the relationship between the lists of past event times and markers, which learns a non-linear hidden representation based on the past sequence. The next event time and marker are predicted by utilizing the hidden representation. The time deviation measures the variation between the predicted event time and actual event time, which is passed to a dense layer along with the RNN hidden representation to predict the next event marker. The following subsections elaborate upon the mathematical problem formulation, preliminaries for the proposed model, and the in-depth explanation of the DMTPP model, followed by the implementation details.

3.1 Problem Definition

The problem setting involves a sequence of events denoted by their time of occurrence and corresponding markers. Mathematically, each sequence is represented by $S = \{(t_1, y_1), (t_2, y_2), \ ..., \ (t_n, y_n)\}$, where n refers to the total sequence

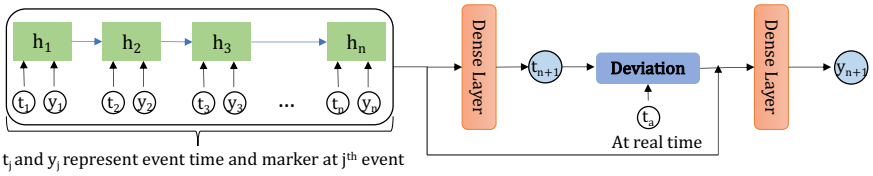


Fig. 3: Diagrammatic representation of the proposed DMTPP model. A sequence of past events $((t_j, y_j),\ j \in (1, n))$ is provided as input to the RNN model, which outputs the learned embedding for the sequence, which is then used for predicting the next event time (t_{n+1}). The deviation between the expected and the actual time event is calculated, followed by the combination of the embedding and the deviation for marker prediction (y_{n+1}).

length. Here, (t_j, y_j) refers to the j^{th} event represented by the time of the event (t_j) and the corresponding marker (y_j). By default, the events are ordered in time, such that $t_{j+1} \geq t_j$. Given the sequence of last n events, often the task is to predict the next event time t_{n+1} and the corresponding marker y_{n+1}. In reference to Fig. 1(a), the event refers to a transaction, event time refers to the time of the transaction, and event marker refers to whether the transaction (event) was fraudulent or not.

3.2 Preliminaries

Temporal Point Process (TPP): A temporal point process is a stochastic process that models a sequence of discrete events occurring in a continuous-time interval [1]. Typically, a TPP is modeled by using a conditional intensity function, which measures the number of events that can be expected in a specific time interval, given the historical sequence of event information. Mathematically, the intensity function of a TPP is defined as the probability an event will occur in $[t, t + dt]$ time interval given the event history h_t till time t:

$$\lambda^*(t)dt = \lambda(t \mid h_t) = P(event\ in\ [t, t + dt] \mid h_t) \tag{1}$$

where, dt refers to a small window of time, and $P(.)$ refers to the probability function. As derived by Du et al. [3], the conditional density function $(f(t|h_t))$ of an event occurring at time t can thus be specified as:

$$f(t|h_t) = \lambda^*(t) \exp\big(-\int_{t_n}^{t} \lambda^*(\tau)d\tau\big) \tag{2}$$

where, t_n refers to the last event and τ corresponds to a very small value tending to zero. The conditional intensity function has been modelled using different parametric forms in the past. Some of the well known methods are:

– Poisson process [11]: Events are assumed to be independent of their history, such that $\lambda(t|h_t) = \lambda(t)$.

– Hawkes Process [6]: In the Hawkes process, the conditional intensity function constitutes a time decay kernel to take into consideration the events history. The intensity function is assumed to be a linear function ($\gamma(.)$) of the history along with the base intensity value (γ_0) and a weight parameter (α) as:

$$\lambda^*(t) = \gamma_0 + \alpha \sum_{t_j < t} \gamma\left(t, t_j\right) \tag{3}$$

As demonstrated above, the traditional temporal point process based techniques model the conditional intensity function by assuming that the data follows some parametric form, which can often be estimated using maximum likelihood estimation (MLE). The above assumption constraints the expressive power of the conditional intensity function, since the true form of the intensity function is unknown in real-time scenarios. This limitation often renders the above techniques unusable in several real-world applications having complex intensity functions.

Marked Temporal Point Process: A natural extension of the TPP based techniques is the inclusion of a marker information along with each event. In such scenarios, the model is expected to predict the next event time and corresponding marker, while having access to the history of past events. Therefore, the conditional intensity function for a marked temporal point process can be formulated as follows:

$$\lambda^*\big((t_j, y_j)\big) = \lambda\big((t_j, y_j) \mid h_t\big) \tag{4}$$

where, t_j and y_j refer to the event time and event marker, respectively. $\lambda^*\big((t_j, y_j)\big)$ can take multiple forms, however, for mathematical simplicity it is mostly assumed that the event time and marker are conditionally independent given the event history i.e. $\lambda^*((t_j, y_j)) = \lambda^*(t_j)\lambda^*(y_j)$. The above assumption assumes independent marker and time occurrence, which often limits the model performance in scenarios where the event time and marker are interdependent.

3.3 Proposed Deviation-based Marked Temporal Point Process

In the proposed Deviation-based Marked Temporal Point Process (DMTPP) model, the objective is to predict the next event marker given the history of the past events. The proposed DMTPP model addresses the discussed limitations by utilizing a universal approximator for learning the conditional intensity function, and by explicitly modeling the relationship between the event time and event marker. This is achieved by utilizing a Recurrent Neural Network (RNN) as the backbone model, and by incorporating the deviation between the actual event time and the predicted event time for marker prediction. Since RNNs are characterized by the property of being a universal approximator, they can thus be applied to model complex intensity functions, and the deviation component can be used to model the relationship between the event occurrence and the marker category. In literature, one of the seminal works involving the usage of

RNNs for marked temporal point processes was presented by Du et al. [3]. The proposed model extends the current body of literature by modeling the inter-dependence of the marker on the event time as well.

Fig 3 presents the diagrammatic representation of the proposed technique. The model takes as input the past time sequence and marker sequence. It utilizes an RNN as the backbone architecture and learns an embedding based on the past events, followed by the prediction of the next event time. The deviation between the predicted time and the actual time of the event is then concatenated with the previously learned embedding for predicting the marker corresponding to the next event. The RNN is thus used to model the intensity function for the given sequences of events.

Mathematically, the last n events are passed as one sequence to the model ($\mathcal{S} = \left((t_j, y_j)_{j=1}^{n}\right)$). For processing, instead of the absolute time stamps, the sequence of the inter-event duration is provided to the algorithm for learning a model invariant to the absolute time. For the time sequence t_j, t_{j-1}, the inter-event duration is calculated as $d_j = t_j - t_{j-1}$. The inter-event duration sequence is calculated for all consecutive events in the sequence $\mathcal{S}$ and is provided to the model $(d_1, d_2,\ ...,\ d_n)$ along with the previous marker sequence $(y_1, y_2,\ ...,\ y_n)$. The marker information is converted into a sparse one-hot encoding for better representation, followed by learning a feature vector using an embedding layer:

$$y_j^{em} = \mathbf{W}_{em}^{\top} y_j + b_{em} \tag{5}$$

where, $\mathbf{W}_{em}$ is the weight matrix for the embedding layer and b_{em} is the bias vector. Thus, the input sequence consisting of the historical inter-event duration (d_j) and the past marker sequence (y_j^{em}) are provided as input to the RNN for learning an embedding capturing the relation between the event sequence and marker. The RNN utilizes the past historical representation (h_{j-1}) along with the other inputs and returns the updated hidden representation as follows:

$$h_j = ReLU(\mathbf{W}^y y_j^{em} + \mathbf{W}^d d_j + \mathbf{W}^h h_{j-1} + b_h) \tag{6}$$

where $\mathbf{W}^y$, $\mathbf{W}^d$, $\mathbf{W}^h$ and b_h denote the marker weight matrix, time duration weight matrix, RNN's representation (history) weight matrix, and the bias weight vector, respectively. Based on the hidden representation h_j, the next inter-event time duration and marker can be calculated as follows:

$$p\left(d_{j+1} \mid h_j\right) = f_t\left(d_{j+1} \mid h_j\right);\ p\left(y_{j+1} \mid h_j\right) = f_y\left(y_{j+1} \mid h_j, \delta_{j+1}\right) \tag{7}$$

where δ_{j+1} is the deviation for the $(j+1)^{th}$ event which corresponds to the difference between the predicted time and the actual time:

$$\delta_{j+1} = t_{j+1} - t_a \tag{8}$$

where t_a and t_{j+1} is the actual and predicted time of the $(j+1)^{th}$ event, respectively. Therefore, the proposed DMTPP model utilizes the learned embedding from the past sequence for predicting the event time, and also incorporates the

deviation between the actual and expected time for marker prediction. Mathematically, for the time prediction, the conditional intensity function is calculated whereas for the marker prediction, the conditional distribution of the marker on hidden representation and deviation of time is calculated. Similar to Du et al. [3], the conditional intensity function for time prediction can be represented as:

$$\lambda^*(t) = \exp\left(w^{h^\top}\mathbf{h}_j + \beta\,(t - t_j) + b\right) \tag{9}$$

where w^h is a weight vector, while β and b are scalar values. The above equation ensures that the conditional intensity is dependent upon the inter-dependence of the past marker and time sequence obtained via the hidden representation from the RNN (first term), the influence of the current time (second term), and an offset base intensity value (third term). Given the above conditional intensity function, the conditional density function for TPPs (Eq. 2) can thus be updated. Therefore, the likelihood of the next event occurring at t_{j+1} given the history h_j can be given as:

$$f_t\left(t_{j+1} \mid h_j\right) = \lambda\left(t_{j+1} \mid h_j\right) \exp\left(-\int_{t_j}^{t_{j+1}} \lambda(t_j \mid h_j)dt\right) \tag{10}$$

The above equation is utilized for predicting the next event time, given the learned hidden representation from an RNN. Given the expected (predicted) and actual time occurrence, the corresponding marker can be predicted using the hidden representation h_j and the time deviation δ_{j+1}, by using the Softmax function on the conditional probability as:

$$f_y\left(y_{j+1} = k \mid h_j, \delta_{j+1}\right) = \frac{\exp\left(\mathbf{W}_k[h_j|\delta_{j+1}] + b_k^y\right)}{\sum_{i=1}^{K}\exp\left(\mathbf{W}_i[h_j|\delta_{j+1}] + b_i^y\right)} \tag{11}$$

where $[h_j|\delta_{j+1}]$ represents the concatenation of the embedding obtained via the RNN and the computed time deviation. Given a K class problem for marker prediction, $\mathbf{W}_i$ refers to the weight vector for the i^{th} marker, and k refers to the correct marker for the next event. The inclusion of the deviation parameter enables the model to learn the inter-dependence between the user behavior (for event occurrence) and the marker. This is especially useful in scenarios where the marker is not known at real-time, but is instead computed after some time of the event occurrence. For example, earthquake intensity or influence of a retweet/advertisement. In such scenarios, the deviation from the expected time of the event can often impact the outcome of the event (marker). By modeling the time deviation, the proposed DMTPP model is thus able to capture the corresponding variations in the marker in real-time. The model is trained by maximizing the joint log-likelihood of the event prediction and marker prediction loss functions as follows:

$$\mathcal{L}\left(\{\mathcal{S}\}\right) = \sum_{i=1}^{n}\sum_{j=r-1}^{\mathcal{S}_n^i-1}\left(\underbrace{\lambda_1 \log f_y\left(y_{j+1}^i \mid [h_j^i, \delta_{j+1}]\right)}_{Marker\ Prediction} + \underbrace{\lambda_2 \log f_t\left(t_{j+1}^i \mid h_j^i\right)}_{Time\ Prediction}\right) \tag{12}$$

Table 1: Details of the datasets used in this research demonstrating variability in terms of size, number of marks, number of events, and average sequence length.

Dataset	No. of Markers	No. of Events	Avg. Seq. Length
MIMIC-II [16]	75	2419	4
StackOverflow [4]	22	480414	72
Retweet [28]	3	2M	209

where, r refers to the first event that is being predicted for sequence i ($\mathcal{S}^i$). n refers to the number of sequences in the training set $\mathcal{S}$, $\mathcal{S}_n^i$ refers to the number of events in sequence $\mathcal{S}^i$. λ_1 and λ_2 refer to the weight given to each component.

3.4 Implementation Details

The proposed DMTPP model has been implemented in the Pytorch environment [17] with a NVIDIA Quadro RTX6000 GPU. As demonstrated in Fig. 3, the DMTPP model consists of an initial embedding layer for the marker sequence, a RNN model, and modules for predicting the event time and marker. The embedding layer is of dimension 10, while the RNN model consists of a single layer Long Short-Term Memory module [8]. A 32 dimension representation is obtained via the RNN model, which is provided to a dense-layer for predicting the time, followed by another dense-layer for marker prediction. Dropout [7] has also been applied as a regularizer after the RNN layer. The weight parameters in Eq. 12 are initialized as follows: $\lambda_1 = 0.15$ and $\lambda_2 = 0.05$. The model is trained using the Adam optimizer [13] for 100 epochs with 1024 batch-size.

4 Experiments and Protocols

In order to evaluate the effectiveness of the proposed Deviation-based Marked Temporal Point Process model, experiments have been performed on three datasets corresponding to different tasks. Table 1 presents the dataset statistics demonstrating high variability across different parameters. Details regarding the datasets and protocols are as follows:

(i) Disease Type Prediction on the MIMIC II Dataset [16]: The MIMIC II dataset is a subset of the Electrical Medical Records Dataset which contains a collection of clinical visit records of Intensive Care Unit (ICU) patients over a period of seven years. Each event contains the time when a patient visits the ICU along with their type of disease (75 disease types). For this dataset, marker prediction corresponds to predicting the disease type. Similar to the existing protocol [29], 90% of the data has been used for training the model, while the remaining 10% corresponds to the test set.

(ii) Badge Prediction on the StackOverflow Dataset [4]: StackOverflow[1] is a popular question answering website where users are awarded with different

[1] https://archive.org/details/stackexchange

Table 2: Marker prediction accuracy (%) on the MIMIC-II and StackOverflow datasets. Comparisons have been performed with the state-of-the-art algorithms. Owing to the same protocol, results have directly been taken from Zuo *et al.* [29].

Model	**MIMIC-II**	**StackOverflow**
Recurrent Marked Temporal Point Process [3]	81.2	45.9
Neural Hawkes Process [15]	83.2	46.3
Time Series Event Sequence [24]	83.0	46.2
Transformer Hawkes Process [29]	85.3	47.0
Deviation-based Marked TPP	**91.76**	**55.42**

badges (Guru, Great Answer, Stellar Question, etc.) for enhancing user engagement and popularity. The StackOverflow dataset contains sequence of badges (marker) received by a user along with the time when the badge is given. A similar protocol and pre-processing is followed as the existing manuscript [3]. The processed dataset contains 6,633 users and 480,414 events with total 22 badges. 90% of the data is used for training, and the remaining 10% is the test set.

(iii) Retweet Prediction on the Retweet Dataset [28]: The Retweet dataset is formed through the Seismic dataset[2]. A stream of retweets is available in which each event is a retweet with the time and number of followers of the user (who has retweeted). Markers are divided into three classes based on the number of followers (degree) of each user: (i) a normal user having degree lower than the median, (ii) an influencer having degree higher than or equal to the median and lower than 95 percentile, and (iii) a celebrity having degree higher than or equal to 95 percentile. Similar to the existing protocol [5], we randomly sample 10,000 streams of retweets and apply five-fold cross validation for experiments.

Consistent with the existing research, the following metrics have been used across different datasets:

- **Classification Accuracy/Micro-F1:** Micro-F1 measures the F1-score of the aggregated contributions of all classes. It is also defined as the overall accuracy which is the ratio of correctly classified samples out of all samples.
- **Macro-F1:** For a multi-class classification problem, Macro-Averaged F1-score or Macro-F1 score is defined as the average of F1-scores of each class.

5 Results and Analysis

Tables 2-3 present the performance of the proposed Deviation-based Marked Temporal Point Process model and other comparative techniques. Figs. 4-5 present the analysis performed on the DMTPP based model. The following paragraphs elaborate the results and analysis of the proposed model:

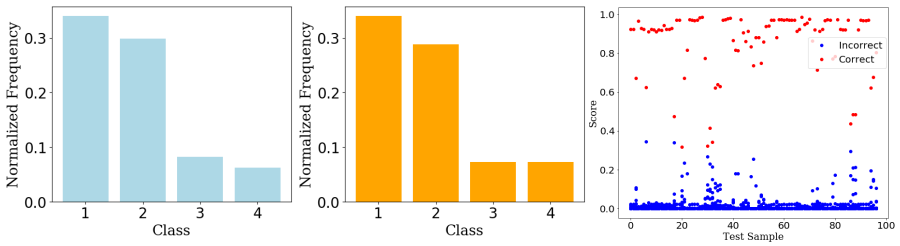
[2] http://snap.stanford.edu/seismic/

Table 3: Marker prediction performance on the Retweet Dataset for predicting the type of retweet. Owing to the same protocol, comparative results have directly been taken from the published manuscript [5].

Model	**Micro-F1**	**Macro-F1**
Noise-Contrastive Estimation Poisson (NCE-P)	0.52	0.28
Noise-Contrastive Estimation Gaussian (NCE-G)	0.49	0.30
MTPP with Discriminative Loss Function (DIS)	0.49	0.29
Maximum Likelihood Estimation (MLE)	0.50	0.29
Monte Carlo Maximum Likelihood Estimation (MCMLE)	0.49	0.28
INITIATOR [5]	0.57	0.35
Deviation-based Marked TPP	**0.58**	**0.45**

Comparison with State-of-The-Art Algorithms: Table 2 presents the performance on the MIMIC-II and StackOverflow datasets. It can be observed that the proposed technique achieves improved performance as compared to the state-of-the-art algorithm (Transformer Hawkes Process [29]). Due to the same protocol, results have directly been taken from the published manuscript. Specifically, on the MIMIC-II dataset, an improvement of at least 6% is observed from the existing results reported by the Transformer Hawkes Process. Further, comparison has also been made with the Recurrent Marked Temporal Point Process model (which forms the base for the proposed technique), where an improvement of over 10% is obtained. A similar improvement is observed on the StackOverflow dataset as well, wherein, the proposed technique achieves 55.42%, resulting in an improvement of over 8% from the current best results (Transformer Hawkes Process [29]). Further, Table 3 presents the performance obtained on the Retweet dataset, wherein results are reported using the standard Micro F-1 and Macro F-1 metrics. Due to the same protocol, results have directly been taken from the published manuscript [5]. From Table 3 it is observed that the proposed model's performance on both metrics is higher than the state-of-the-art model: the INITIATOR algorithm. Specifically, the proposed model achieves a Micro-F1 and Macro-F1 value of 0.58 and 0.45, respectively. Since the Retweet dataset is characterized by heavy class imbalance in the testing set, the above metrics provide a better understanding of the model performance as compared to traditional classification accuracy. We have also performed the Chi-Squared Statistic Test of Independence [14] on the Retweet dataset to evaluate the statistical association between the results obtained by the proposed model and the state-of-the-art INITIATOR model. A $p-$value of less than 0.01 is obtained between the Micro-F1 scores, which provides us with sufficient evidence to conclude that the models are disassociated. We believe that the explicit modeling of the dependence of the marker prediction on the time prediction enables the model to learn better features, thus resulting in improved marker prediction performance.

Analysis of the proposed Deviation-based Marked TPP Model: Fig. 4(a-b) presents the marker distribution of the (a) ground-truth labels and the



(a) Actual Labels (b) Predicted Labels (c) Score Distribution

Fig. 4: (a-b) Bar graphs demonstrating the distribution for the (a) ground-truth and (b) predicted marker labels. Distribution of the top four classes from the MIMIC-II dataset has been plotted. The predicted distribution follows a similar pattern as the ground-truth distribution, thus suggesting that the DMTPP model is able to capture the marker spread well. (c) Score distribution obtained from the DMTPP model on the test events of the MIMIC-II dataset. For each sample the actual (correct) class score and the other (incorrect) class scores obtained via the model have been plotted. For almost all samples, a clear distinction is seen between the correct and incorrect class scores.

(b) labels predicted by the DMTPP model on the MIMIC-II dataset. A similar distribution is observed across the two graphs, thus suggesting that the proposed model is able to learn and simulate the varying occurrence of different marker types. Further, Fig. 4(c) presents the score distribution for different samples of the MIMIC-II dataset, where clear distinction can be observed between the scores of the correct class versus the scores of the incorrect class. A large number of incorrect class scores fall below the range of 0.1 which demonstrates the discriminative nature of the learned classifier. Experiments have also been performed to analyze the different components and hyper-parameters of the Deviation-based Marked TPP model. Discussions regarding different aspects are as follows:

(i) Effect of Sequence Length: Experiments have been performed on the MIMIC-II dataset for understanding the effect of the input sequence length on the model's performance. The sequence length determines the relevance of the length of the user's history for predicting the next marker. We observe that on higher sequence length the model's performance decreases. Specifically, the model achieves 72.3% and 86.1% with a sequence length of 5 and 4, respectively, while achieving 91.76% with sequence length 3. Reducing the length further to 2 results in an accuracy of 88.2%, thus demonstrating a slight drop, while still achieving improved results from the state-of-the-art technique.

(ii) Effect of Weight Hyper-parameters: Experiments have also been performed to understand the effect of the weight hyper-parameters (λ_1 and λ_2 in Eq. 12). Specifically, the MIMIC-II dataset has been used to see the impact of varying weights of time loss and marker loss on the model's performance. As mentioned earlier, best performance of 91.76% is achieved with the following

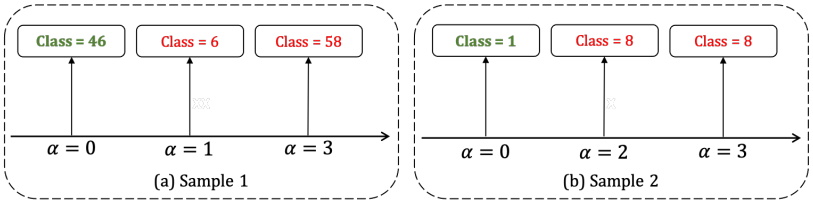


Fig. 5: Sample simulation demonstrating the effect of the deviation parameter for predicting the marker (class). Two sequences are presented where an alpha value (α) was added to the deviation, followed by marker prediction. Variation in the class prediction suggests dependency on the deviation parameter.

combination: (λ_1=0.05 and λ_2=0.15). A drop in performance is observed upon varying the value of λ_1 or λ_2, respectively. For example, a classification accuracy of 90.73% is obtained with the weight pairs (0.05, 1), (0.05, 2), and (0.01, 1). The slight drop in performance suggests the model's robustness to variations in the hyper-parameter selection.

(iii) Performance on Time Prediction: While the aim of the DMTPP model is to perform accurate marker prediction, analysis has also been performed on the MIMIC-II dataset to understand its performance for time prediction. The proposed model obtains a Root Mean Squared Error (RMSE) value of 0.89 for time prediction, which is the second best in comparison to the state-of-the-art performance of 0.82 obtained by the Transformer Hawked Process [29]. Further, the proposed model performs better as compared to the other reported results, specifically, Recurrent Marked Temporal Point Process [3]: 6.12, Neural Hawkes Process [15]: 6.13, Time Series Event Sequence [24]: 4.70, Self-Attentive Hawkes Process [25]: 3.89. The accurate time prediction further supports the high marker prediction performance of the proposed model.

(iv) Effect of Deviation: Experiments have also been performed to understand the effect of the deviation parameter, and whether it contributes to the marker prediction or not. A simulation was performed on the MIMIC-II dataset, where, a small value (α) was added to the deviation obtained after the time prediction. The updated or perturbed deviation value was then provided with the learned embedding for predicting the corresponding marker. Fig. 5 presents the output obtained on two sample sequences. In both the cases, the predicted marker was updated when the deviation was changed by an α value. Similar behavior was observed across different sequences as well. The simulation suggests that the DMTPP based model is able to learn the inter-dependence between the user behavior (event occurrence) and the corresponding marker, and updates its prediction based on the variability between the expected and actual event time.

6 Conclusion and Discussion

In various real-world scenarios, the marker information is not immediately known after the occurrence of an event. For example, the impact of an advertisement

retweet on increased sales or online traffic generation, intensity of an earthquake, or identifying fraudulent transactions. In such scenarios, the marker information (advertisement impact, earthquake intensity, and fraud event) is known after some time of the event occurrence. It is our hypothesis that in such scenarios, the variation between the expected and actual event occurrence also impacts the corresponding marker. Therefore, in this research, a novel Deviation based Marked Temporal Point Process (DMTPP) model is proposed. The proposed model focuses on learning the dependency between the event time and marker information for predicting accurate markers. The DMTPP model builds upon the existing literature in the field of Marked Temporal Point Processes which has focused majorly on predicting the next event time and corresponding marker information without explicitly modeling the relationship between the two. The efficacy of the proposed model has been demonstrated on three different tasks and datasets (Table 2 and Table 3), where it achieves state-of-the-art performance. Further analysis on the model demonstrates its high performance for time prediction and impact of the deviation component as well. We believe that the research performed in this paper can act as a stepping stone to further explore the possibilities that Temporal Point Processes hold in terms of high accuracy of event type prediction and not just event time prediction. One of the key highlights of the DMTPP model is the requirement of real-time event occurrence (time) for marker prediction, which can further be improved in future algorithms. As part of future work, the proposed DMTPP model can also be extended to incorporate additional meta-information during training which can further boost the marker prediction performance.

References

1. Daley, D.J., Vere-Jones, D.: An introduction to the theory of point processes: Volume I: Elementary theory and methods. Springer (2003)
2. Diggle, P.J.: Statistical analysis of spatial and spatio-temporal point patterns. CRC press (2013)
3. Du, N., Dai, H., Trivedi, R., Upadhyay, U., Gomez-Rodriguez, M., Song, L.: Recurrent marked temporal point processes: Embedding event history to vector. In: Proceedings of ACM KDD. pp. 1555–1564 (2016)
4. Grant, S., Betts, B.: Encouraging user behaviour with achievements: an empirical study. In: Proceedings of MSR. pp. 65–68 (2013)
5. Guo, R., Li, J., Liu, H.: Initiator: Noise-contrastive estimation for marked temporal point process. In: Proceedings of IJCAI. pp. 2191–2197 (2018)
6. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika **58**(1), 83–90 (1971)
7. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580 (2012)
8. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
9. Isham, V., Westcott, M.: A self-correcting point process. Stochastic Processes and their Applications **8**(3), 335–347 (1979)

10. Ji, Y., Yin, M., Fang, Y., Yang, H., Wang, X., Jia, T., Shi, C.: Temporal heterogeneous interaction graph embedding for next-item recommendation. Proceedings of ECML-PKDD (2020)
11. Kemp, A.: Poisson Processes. Wiley Online Library (1994)
12. Keogh, E., Chu, S., Hart, D., Pazzani, M.: Segmenting time series: A survey and novel approach. In: Data Mining in Time Series Databases, pp. 1–21 (2004)
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. McHugh, M.L.: The chi-square test of independence. Biochemia medica **23**(2), 143–149 (2013)
15. Mei, H., Eisner, J.: The neural hawkes process: A neurally self-modulating multivariate point process. arXiv preprint arXiv:1612.09328 (2016)
16. Pan, Z., Du, H., Ngiam, K.Y., Wang, F., Shum, P., Feng, M.: A self-correcting deep learning approach to predict acute conditions in critical care. arXiv preprint arXiv:1901.04364 (2019)
17. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Proceedings of NeurIPS, pp. 8024–8035 (2019)
18. Rasmussen, J.G.: Temporal Point Processes: The Conditional Intensity Function. Lecture Notes, Jan (2011)
19. Shchur, O., Biloš, M., Günnemann, S.: Intensity-free learning of temporal point processes. arXiv preprint arXiv:1909.12127 (2019)
20. Türkmen, A.C., Wang, Y., Smola, A.J.: Fastpoint: Scalable deep point processes. In: Proceedings of ECML-PKDD. pp. 465–480 (2019)
21. Verenich, I., Dumas, M., Rosa, M.L., Maggi, F.M., Teinemaa, I.: Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring. ACM TIST **10**(4), 1–34 (2019)
22. Wang, Y., Liu, S., Shen, H., Gao, J., Cheng, X.: Marked temporal dynamics modeling based on recurrent neural network. In: Proceedings of PAKDD. pp. 786–798 (2017)
23. Wu, W., Yan, J., Yang, X., Zha, H.: Decoupled learning for factorial marked temporal point processes. In: Proceedings of ACM KDD. pp. 2516–2525 (2018)
24. Xiao, S., Yan, J., Yang, X., Zha, H., Chu, S.: Modeling the intensity function of point process via recurrent neural networks. In: Proceedings of AAAI. vol. 31 (2017)
25. Zhang, Q., Lipani, A., Kirnap, O., Yilmaz, E.: Self-attentive hawkes processes. arXiv preprint arXiv:1907.07561 (2019)
26. Zhang, Q., Lipani, A., Kirnap, O., Yilmaz, E.: Self-attentive hawkes process. In: Proceedings of ICML. pp. 11183–11193 (2020)
27. Zhao, L.: Event Prediction in Big Data Era: A Systematic Survey. arXiv preprint arXiv:2007.09815 (2020)
28. Zhao, Q., Erdogdu, M.A., He, H.Y., Rajaraman, A., Leskovec, J.: Seismic: A self-exciting point process model for predicting tweet popularity. In: Proceedings of KDD. pp. 1513–1522 (2015)
29. Zuo, S., Jiang, H., Li, Z., Zhao, T., Zha, H.: Transformer hawkes process. In: Proceedings of ICML. pp. 11692–11702 (2020)