Event Uncertainty using Ensemble Neural Hawkes Process

Manisha Dubey*
Indian Institute of Technology
Hyderabad, India
cs17resch11003@iith.ac.in

Ragja Palakkadavath†
Deakin University
Melbourne, Australia
ragjapk@gmail.com

P.K. Srijith ‡
Indian Institute of Technology
Hyderabad, India
srijith@cse.iith.ac.in

ABSTRACT

Various real world applications in science and industry are often recorded over time as asynchronous event sequences. These event sequences comprise of the time of occurrence of events. Different applications including such event sequences are crime analysis, earthquake prediction, neural spiking train study, infectious disease prediction etc. A principled framework for modeling asynchronous event sequences is temporal point process. Recent works on neural temporal point process have combined the theoretical foundation of point process with universal approximation ability of neural networks. However, the predictions made by these models are uncertain due to incorrect model inference. Therefore, it is highly desirable to associate uncertainty with the predictions as well. In this paper, we propose a novel model, Ensemble Neural Hawkes Process, which is capable of predicting event occurrence time along with uncertainty, hence improving the generalization capability. We also propose evaluation metric which captures the uncertainty modelling capability for event prediction. The efficacy of proposed model is demonstrated using various simulated and real world datasets.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Event modeling, Time-to-event prediction, Hawkes process

ACM Reference Format:
Manisha Dubey, Ragja Palakkadavath, and P.K. Srijith . 2023. Event Uncertainty using Ensemble Neural Hawkes Process. In *6th Joint International Conference on Data Science and Management of Data (10th ACM IKDD CODS and 28th COMAD) (CODS-COMAD 2023), January 4–7, 2023, Mumbai, India.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3570991.3571002

1 INTRODUCTION

Many often real-world activities are recorded over time as asynchronous event sequences, which essentially means that time interval between events are as important as order of the events. These discrete and irregular event sequences are modeled using a principled framework known as temporal point process (TPP) [20]. A keytool of temporal point process is the conditional intensity function which models the instantaneous occurrence of event. A commonly used point process known as Hawkes process [12] has a self-triggering property, i.e occurrence of the previous events trigger occurrences of future events. Hawkes process has been used in earthquake modelling [11], crime forecasting [14], finance [3, 8] and epidemic forecasting [5, 7]. Hawkes process assumes that past events always bring positive influence on current events. However, such a strong assumption might lead to model misspecification. Therefore, recent works use neural network to model conditional intensity function.

Recent works of neural temporal point process have brought point process to mainstream machine learning. Despite this, real world implementation of TPP is still scarce. One of the plausible reasons could be that current literature doesn't consider uncertainty quantification over the predictions. Uncertainty often emerges due to limitation of modeling capabilities. Studies have shown that DNNs are overconfident in their prediction results and produce miscalibrated softmax output probabilities for classification [10]. This may lead to lack of confidence in the predictions, hence limiting its use in real-world scenarios, e.g in critical applications like seismology, finance and epidemiology etc. Due to this criticality, the applications must be associated with uncertainty values along with the predictions in order to reduce the risk. Applications of uncertainty quantification for time-to-event modeling are promising and can aid in better decision making. For example, it will help in better risk assessment if we can obtain uncertainty over predicted time of occurrence of earthquake to have a better understanding of seismic vulnerability and hence, earthquake risk assessment. Similarly, it is quintessential to model the high-frequency trading events such that model can predict uncertainty associated with the time of occurrence of buy-sell orders for limit book order transactions. Awareness of such uncertainty in such settings can help to manage resources, react to market changes accordingly and understand market microstructure in a better way. In above examples, we can observe that overconfident incorrect predictions can be deleterious. Therefore, uncertainty quantification is often crucial for assessing trustworthiness of the predictions of the time-to-event model.

Bayesian Neural Networks (BNNs) [9, 16] are widely used framework to find uncertainty estimates for deep models. However, ensembles [13] created from different initializations have been shown to outperform various approximate Bayesian neural network models. Deep ensembles [6] are considered to be fully congruous with Bayesian model averaging which estimates posterior distribution by marginalizing the parameters [23]. Several works have been done in this direction for establishing effectiveness of deep ensembles in a theoretical manner [2] and to improve the effectiveness of deep ensembles [1, 15, 19, 26, 27]. Here, uncertainty is captured due to optimization to different local minima. However, this setting might

CODS-COMAD 2023, January 4–7, 2023, Mumbai, India

ACM ISBN 978-1-4503-9797-1/23/01…$15.00
https://doi.org/10.1145/3570991.3571002

limit uncertainty estimation ability. [22] considers different hyper-parameters to bring diversity during uncertainty quantification.

In this work, we propose a novel yet less studied problem of uncertainty prediction over time-to-event occurrence. We base our contribution on the insight that ensemble neural networks, which utilizes multiple independently trained neural networks, can be used for uncertainty modeling. Further, we propose multiple variants to bring diversity to the network. Our work discusses possibilities to bring diversity in time-to-event models. The proposed methods form a set of simple yet effective ways to obtain uncertainty along with time of occurrence of events for event modeling.

There are few works [21, 28, 29] where uncertainty is captured over the parameters of parametric hawkes process. In [4] authors have proposed an adversarial non-parametric model that accounts for calibration and uncertainty for time to event and uncertainty prediction. Another work [24] has introduced a model for uncertainty in Accelerated Failure Time (AFT) models using a combination of RNN and sparse Gaussian process. However, neural point processes are a widely adopted model for event modeling due to its theoretical and practical effectiveness in various applications. Therefore, the goal of this paper is to augment the capabilities of the neural Hawkes process to predict future events with the uncertainty estimation capability. To achieve this, we propose a novel approach which combines neural Hawkes process and ensembles for performing uncertainty estimation for event-time prediction. In summary, our contributions are as follows -

- We propose a novel simple and effective model, Ensemble Neural Hawkes Process, which combines the advantages of the neural Hawkes process and ensembles for modelling uncertainty over time of occurrence of events.
- We develop multiple variants of Ensemble Neural Hawkes Process to diversify the models so that it can be adapted to varied applications for better uncertainty quantification.
- We propose new evaluation metrics for uncertainty estimaion which can be used for time-to-event modeling.
- We demonstrate the effectiveness of ENHP for uncertainty modelling and prediction of event times and locations on several simulated and real world data.

2 METHODOLOGY

2.1 Problem Definition

We consider the input sequence $S = \{t_i\}_{i=1}^N$ in the observation interval $[0, T]$ and inter-event time interval as $\tau_i = t_i - t_{i-1}$. Our goal is to predict the time of occurrence of next events along with uncertainty estimates over the predicted time.

2.2 Ensemble Neural Hawkes Process

[13] proved deep ensembles have improved performance along with uncertainty estimation over single models. A plausible reason for this could be attributed to the fact that deep ensembles present different samples from various modes of the loss setting [18]. Along this direction, a recent theoretical work [2] has suggested that multi-view structure of data can be captured by ensemble neural network in a better way. Leveraging these ideas, we propose Ensemble Neural Hawkes Process where each model of ensemble captures different features, hence boosting the predictive performance along with uncertainty quantification for events as well.

Each base learner of the ensemble will be neural Hawkes processes [17], which has modeled intensity function through the combination of recurrent neural network followed by feedforward neural network. Recurrent Neural Networks (RNNs) is used to represent history using hidden representation $\boldsymbol{h}_i$ at time t_i, can be denoted as $\boldsymbol{h}_i = RNN(\tau_i, \boldsymbol{h}_{i-1}; W_r)$ where W_r represents the parameters associated with RNN. This is used as input to a feedforward neural network to compute the hazard function which is used for computing the likelihood of event occurrences. In the proposed model, we consider input to the feed-forward neural network 1) hidden representation from RNN, 2) time of event occurrence and 3) elapsed time from the most recent event. To better capture the uncertainty in predicting points in future time, we considered the time at which the intensity function needs to be evaluated. We model the conditional intensity as a function of the elapsed time from the most recent event $\lambda(t|H_t) = \lambda(t - t_i|\boldsymbol{h}_i, t_i)$ where $\lambda(\cdot)$ is a non-negative function referred to as a hazard function. Therefore, we define cumulative hazard function in terms of inter-event interval $\tau = t - t_i$ as $\Phi(\tau|\boldsymbol{h}_i, t) = \int_0^\tau \lambda(s|\boldsymbol{h}_i, t)ds$. Cumulative hazard function is modeled using a feed-forward neural network (FNN) as $\Phi(\tau|\boldsymbol{h}_i, t) = FNN(\tau, h_i, t; W_t)$.

We use a combination of M such models to create a more powerful model (ENHP) which is capable of giving better predictions along with the uncertainty quantification. The recurrent neural network of each base model can be denoted as $\boldsymbol{h}_i^j = RNN(\tau_i, \boldsymbol{h}_{i-1}^j; W_r^j)$. Feed-forward neural network (FNN) models cumulative hazard function for each base model separately which can be expressed as $\Phi^j(\tau|\boldsymbol{h}_i^j, t) = FNN(\tau, h_i^j, t; W_t^j)$. Here, W_t^j represents the parameters associated with feed-forward neural network for j^{th} base model. Combining this, the log-likelihood can be modified as -

$$\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N}\Big(\log\big(\frac{\partial}{\partial\tau}\Phi^j(\tau_i|\boldsymbol{h}_{i-1}^j, t_i; W^j)\big) - \Phi^j(\tau_i|\boldsymbol{h}_{i-1}^j, t_i; W^j)\Big) \quad (1)$$

where $\tau_i = t_i - t_{i-1}$ and $W^j = \{W_r^j, W_t^j\}$ represents the combined weights associated with RNN and FNN for j^{th} ensemble model.

For promoting diversity in ensemble, one can go beyond relying on random initializations in order to avoid redundancy in model averaging. Several works [15, 26] have introduced methods to bring diversity in the ensembles. We will use the combination of following ways to consider variety into ensembles - **1) Random initialization:** The loss landscape of neural networks in non-linear due to incorporation of activation functions. Therefore, random initialization in neural network can lead to different training results. **2) Data shuffling:** In practical settings, neural networks are trained in a better way with the use of mini-batches. Hence, data shuffling can lead to variety in base learners. **3) Bootstrapping:** This allows to vary the distribution of the training set by sampling new sets of training samples for different base learners. **4) Different Network Architecture:** Upon using different network architectures, one can use a combination of different loss landscapes, hence bringing in the variety in ensembles.

The advantage of the approach is that these individual networks can be trained in parallel to get M values of cumulative intensity for each event. These M cumulative intensities are then used to predict future events. Further we discuss prediction methodology.

2.3 Prediction and Uncertainty Estimation

For prediction, ensemble models perform model combination where a weighted combination of all the base learners are used to predict time of occurrence of next event rather than a point estimate of the parameters. This allows them to make robust predictions and model uncertainty over the predictions. The probability of predicting the time of next event given the history of previous event times with last event at t_N is computed as

$$p(t_*|\mathcal{H}) = \frac{1}{M}\sum_{j=1}^{M} p(t_*|\mathcal{H}; W^j)$$

where W^j represents the weights from j^{th} model of ensemble. For each network, we use the bisection method [17] to predict the time of the next event. Bisection method provides the median t_* of the predictive distribution over next event time using the relation $\Phi(t_* - t_N|\boldsymbol{h}_N^j, t_*; W^j) = \log(2)$. We obtain S median event times, with each median event time t_{*_j} obtained using a network with weight W^j. The mean event time $t*$ is then found by averaging M times obtained from different sampled weights. The variance of predicted time is obtained as $Var(t_*) = {\sigma_*}^2 = \frac{1}{M}\sum_{j=1}^{M}(t_{*_j} - t_*)^2$. We also define lower and upper bound of predicted time as -

$$t_*^L = t_* - k\sigma_* \quad t_*^U = t_* + k\sigma_* \tag{2}$$

3 EXPERIMENTS

Dataset Details. We extensively perform experiments on six datasets including two synthetic and four real-world datasets. **1) Simulated Poisson (Sim-Poisson):** We simulate a homogeneous poisson process with conditional intensity $\lambda = 1$. **2) Simulated Hawkes (Sim-Hawkes):** We use the Hawkes process, in which the kernel function is given by the sum of multiple exponential functions. The conditional intensity function is given by $\lambda(t) = \mu + \sum_{t_i<t}\sum_{j=1}^{2}\alpha_j\beta_j\exp(-\beta_j(t-t_i))$. We have used $\mu = 0.05$, $\alpha_1 = 0.4$, $\alpha_2 = 0.4$, $\beta_1 = 1.0$, $\beta_1 = 20.0$. **3) Crime:** This dataset contains the records of the police department calls for service in San Francisco [1]. Each record contains the crime and timestamp associated with the call along with other information. **4) Music:** This dataset contains the history of music listening of users at *lastfm* [2].

Baselines. To the best of our knowledge, this is the first work in this direction. Therefore, we propose our own baselines which are as - **NHP:** Neural Hawkes process (NHP) [17] serves as a baseline for evaluating the performance of predicted time of occurrence of event. **SHP:** We consider the standard Hawkes Process with exponential kernel as another baseline. NHP and SHP model probability distribution over event times $p(\tau|h_i)$, though they do not model epistemic uncertainty. **Ensemble Hawkes (EH):** We propose a new baseline capable of modelling epistemic uncertainty - an ensemble of parametric HP using exponential kernels with different hyperparameters. To avoid convergence issues, we use least-square loss [25] for training EH. Due to this, MNLL of EH won't be comparable with other methods which use survival likelihood. Each ensemble consists of 5 models with decays ranging between 0.001 to 0.1.

Evaluation metrics. We consider these metrics for evaluation **1) Mean Negative Log Likelihood (MNLL):** Log-likelihood considers the probability of predicting the actual observations, and we expect a good model to have a lower MNLL score. **2) Mean Absolute Error for prediction (MAE):** The measure computes purely the absolute error in predicting the time without considering the probability. A model with low MAE will reflect a better model. **3) Prediction Ratio (PR):** We propose this metric to determine quality of intervals for event prediction task. It represents the ratio of number of times an actual event occurs within the estimated interval out of total number of events. Different values of k are used while finding interval in (2). Higher PR will represent a better model. **4) Prediction Bounds (PB):** It represents the length of difference between upper bound and lower bound of the predicted interval (Refer (2)). We can't always consider that higher PB will be better because too large intervals may not represent meaningful uncertainty. So, we use this metric for qualitative evaluation.

Implementation Details. Our ensemble includes 5 neural Hawkes process as base learner. We investigate techniques mentioned in Section 2.2 to diversify the samples. We report results on the variants of the models as discussed below - *ENHP-Model1:* Each base learner of ensemble uses different initializations. Also, training data is shuffled for modeling diverse distributions. The dropout rate and regularization constant is 0.2 and 1e-2 respectively for this. *ENHP-Model2:* Each model of ensemble uses different initializations, hyperparameters and shuffled training data. Training data is shuffled for modeling diverse distributions. The dropout rate and regularization constant for each model is randomly sampled from [0.1, 0.2, 0.4, 0.5, 0.8, 0.9] and [1e-1, 1e-2, 1e-3] respectively. *ENHP-Model3:* In addition to the settings of *ENHP-Model2*, this model is fed bootstrapped samples of input. The inputs are sampled with replacement with probability 0.2 to get input for each base-learner. *ENHP-Model4:* In addition to the settings of *ENHP-Model3*, this model consists of ensemble of feed-forward network which models cumulative hazard function but it contains common recurrent neural network. For aggregation, we perform uniform weights for each deep ensemble base learner. We perform single step lookahead prediction on test inputs.

4 RESULTS AND ANALYSIS

4.1 Results

We report our results in Table 1. We can observe that methods from the proposed methodologies are performing better in terms of predictive evaluation for all datasets. The variants of ENHP model achieve lower MNLL, implying better predictive performance. Also, MAE is lowest for the proposed methods for all the datasets. We can also observe that proposed methodologies are better in terms of uncertainty quantification. ENHP variants are showing better coverage as compared to baselines. Therefore, we have created wider intervals for time-to-event modeling which can capture more events within the predicted intervals. We can observe that among the variants of ENHP, there is no model which performs for all datasets. We can select the model whichever better suits our needs for a

[1] https://catalog.data.gov/dataset/police-calls-for-service
[2] https://www.dtic.upf.edu/~ocelma/MusicRecommendationDataset/lastfm-1K.html

Dataset	Method	MNLL	MAE	PR@1	PR@2	PR@5	PB
Sim-Poisson	SHP	1.21	0.81	0.0	0.0	0.0	0.0
	NHP	1.36	0.76	0.0	0.0	0.0	1e-6
	EH	*	0.172	0.019	0.031	0.061	0.0361
	ENHP-Model1	**-1.023**	**0.071**	**0.009**	**0.017**	**0.043**	0.0019
	ENHP-Model2	-1.015	**0.071**	0.006	0.011	0.028	0.0012
	ENHP-Model3	-1.014	**0.071**	**0.009**	**0.017**	**0.043**	0.0019
	ENHP-Model4	-1.013	**0.071**	0.005	0.0.011	0.026	0.0019
Sim-Hawkes	SHP	1.08	0.382	0.0	0.0	0.0	0.0
	NHP	1.26	0.83	0.0	0.0	0.0	1e-6
	EH	*	0.234	0.012	0.029	0.058	0.0431
	ENHP-Model1	-0.503	**0.114**	0.055	0.112	0.246	0.017
	ENHP-Model2	-0.507	0.109	0.235	0.433	**0.756**	0.071
	ENHP-Model3	**-0.523**	0.109	0.204	0.379	0.724	0.061
	ENHP-Model4	-0.470	0.109	**0.251**	**0.441**	0.701	0.074
Crime	SHP	1.809	0.903	0.0	0.0	0.0	0.0
	NHP	1.42	0.82	0.0	0.0	0.0	1e-6
	EH	*	0.242	0.002	0.005	0.015	0.0515
	ENHP-Model1	**-1.051**	0.172	0.174	0.531	**1.000**	0.218
	ENHP-Model2	-1.045	0.103	0.559	**0.945**	**1.000**	0.218
	ENHP-Model3	-1.048	0.101	**0.564**	**0.945**	**1.000**	0.215
	ENHP-Model4	-0.917	0.067	0.152	0.297	0.661	0.031
Music	SHP	1.239	0.997	0.0	0.0	0.0	0.0
	NHP	0.97	0.58	0.0	0.0	0.0	1e-6
	EH	*	0.104	0.265	0.312	0.458	0.045
	ENHP-Model1	-1.519	**0.074**	0.048	0.277	0.679	0.047
	ENHP-Model2	**-1.278**	0.061	0.464	**0.542**	**0.965**	0.093
	ENHP-Model3	-1.186	0.054	**0.471**	0.531	0.775	0.089
	ENHP-Model4	-1.148	0.046	0.177	0.338	0.491	0.018

Table 1: Comparison of the proposed approaches against the baselines (Please note that 1) EH uses least square loss and hence MNLL for EH is not comparable to our results, hence MNLL for EH is marked with '*'. 2) We can't always consider wider intervals are better because uncertainty might lose its importance in that case, so we have not boldfaced PB)

particular application. For example, if we consider *Music* dataset, if we need tighter bounds we can use ENHP-Model1 which exhibits lower PR@1 and PB. However, if we want wider intervals, we should use ENHP-Model2. So, the choice of model will depend on the requirement of the application and how we need to set trade-off between predictive performance and uncertainty quantification.

4.2 Qualitative Evaluation of Prediction Bounds

An important aspect of uncertainty quantification for time-to-event modeling is the evaluation of the predicted uncertainty bounds. Ideally, a more confident model will exhibit smaller uncertainty for correct predictions. So, prediction bound should be low for correct predictions and incorrect prediction bounds should have larger prediction bound. Also, correct predictions should have lower MAE and incorrect predictions will have higher MAE. So, an ideal model should lead to lower prediction bound for lower MAE and larger prediction bound for higher MAE. We acknowledge this behavior by finding correlation between MAE and PB. Positive correlation implies that when model is predicting correctly, the predictions bounds are of smaller length, hence implying that model is more confident about the predictions. Table 2 displays the correlation coefficient between MAE and PB. We can observe that there is a positive correlation for all dataset, which is congruous with the expected behavior.

Table 2: Correlation between MAE and PB

Dataset	Model1	Model2	Model3	Model4
Sim-Poisson	0.004	0.148	0.249	0.149
Sim-Hawkes	0.073	0.0004	0.004	0.0005
Crime	0.067	0.026	0.006	0.063
Music	0.134	0.236	0.709	0.407

Table 3: Effect of using weighted base learner in Ensemble

	Sim-Hawkes		Crime	
	Model2	Model4	Model1	Model3
MAE	0.109	0.109	0.171	0.098
PR@1	0.220	0.249	0.169	0.559
PB	0.069	0.070	0.216	0.211

4.3 Discussion

We have considered uniform weights to combine outputs of the ensemble in the experimental results. This is done because our goal is to improve predictive performance (MNLL and MAE) as well as uncertainty modeling ability (PR@k, PB). Using a weighted combination of ensemble could improve the predictive performance but can affect the uncertainty modeling. This is because, typically weights are learnt from the validation data based on predictive performance. Moreover, weighted ensemble will imply using weighted mean and weighted variance. The weighted variance would be lower if the weights are not uniform (high weights for one or a few models), since weighted average will be close to the values of a base learner with high weight and a base learner with high deviation will have low weights and vice-versa. Consequently, PR@k and PB metrics considering the prediction interval gets affected, worsening the uncertainty quantification ability of the model. This behavior is reflected in the results displayed in Table 3, where we considered the weights of 5 ensembles as [0.5, 0.1, 0.1, 0.1, 0.2] and choose the best weight combination from the validation data. We can observe that although weighted model gives better predictive performance, their uncertainty quantification ability is lower than their unweighted counterparts. Therefore, we use unweighted combination of ensembles to better capture uncertainty in predictions.

5 CONCLUSION

We propose Ensemble Neural Hawkes Process, a framework combines the advantages of neural Hawkes process and ensembles for time-to-event prediction along with uncertainty quantification over the predicted time. We also propose different variants of Ensemble Neural Hawkes Process which can help in diversifying each neural Hawkes process. We also propose evaluation metrics for uncertainty quantification suited towards time-to-event modeling. The proposed framework improves the predictive performance along with providing uncertainty estimates for the time of occurrence of event. Our experiments on several simulated and real world datasets showcase the efficacy of the proposed approach. For future work, we want to consider uncertainty modeling for neural marked temporal point process where we can consider uncertainty over marks along with time of occurrence of event, hence can be useful for more applications.

REFERENCES

[1] Ahmed Alaa and Mihaela Van Der Schaar. 2020. Discriminative jackknife: Quantifying uncertainty in deep learning via higher-order influence functions. In *International Conference on Machine Learning*. PMLR, 165–174.

[2] Zeyuan Allen-Zhu and Yuanzhi Li. 2020. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816* (2020).

[3] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. 2015. Hawkes processes in finance. *Market Microstructure and Liquidity* 1, 01 (2015), 1550005.

[4] Paidamoyo Chapfuwa, Chenyang Tao, Chunyuan Li, Irfan Khan, Karen J Chandross, Michael J Pencina, Lawrence Carin, and Ricardo Henao. 2020. Calibration and Uncertainty in Neural Time-to-Event Modeling. *IEEE Transactions on Neural Networks and Learning Systems* (2020).

[5] Wen-Hao Chiang, Xueying Liu, and George Mohler. 2021. Hawkes process modeling of COVID-19 with mobility leading indicators and spatial covariates. *International journal of forecasting* (2021).

[6] Francesco D'Angelo and Vincent Fortuin. 2021. Repulsive deep ensembles are bayesian. *Advances in Neural Information Processing Systems* 34 (2021), 3451–3465.

[7] Peter Diggle, Barry Rowlingson, and Ting-li Su. 2005. Point process methodology for on-line spatio-temporal disease surveillance. *Environmetrics: The official journal of the International Environmetrics Society* 16, 5 (2005), 423–434.

[8] Paul Embrechts, Thomas Liniger, and Lu Lin. 2011. Multivariate Hawkes processes: an application to financial data. *Journal of Applied Probability* 48, A (2011), 367–378.

[9] Yarin Gal. 2016. *Uncertainty in deep learning*. Ph. D. Dissertation. PhD thesis, University of Cambridge.

[10] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[11] Sebastian Hainzl, D Steacy, and S Marsan. 2010. Seismicity models based on Coulomb stress calculations. *Community Online Resource for Statistical Seismicity Analysis* (2010).

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

[13] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *Advances in Neural Information Processing Systems* 30 (2017).

[14] George O Mohler, Martin B Short, P Jeffrey Brantingham, Frederic Paik Schoenberg, and George E Tita. 2011. Self-exciting point process modeling of crime. *J. Amer. Statist. Assoc.* 106, 493 (2011), 100–108.

[15] Giung Nam, Jongmin Yoon, Yoonho Lee, and Juho Lee. 2021. Diversity matters when learning from ensembles. *Advances in Neural Information Processing Systems* 34 (2021), 8367–8377.

[16] Radford M Neal. 2012. *Bayesian learning for neural networks*. Vol. 118. Springer Science & Business Media.

[17] Takahiro Omi, Kazuyuki Aihara, et al. 2019. Fully Neural Network based Model for General Temporal Point Processes. *Advances in Neural Information Processing Systems* 32 (2019), 2122–2132.

[18] Tim Pearce, Alexandra Brintrup, Mohamed Zaki, and Andy Neely. 2018. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International Conference on Machine Learning*. PMLR, 4075–4084.

[19] Alexandre Ramé, Rémy Sun, and Matthieu Cord. 2021. Mixmo: Mixing multiple inputs for multiple outputs via deep subnetworks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 823–833.

[20] Esko Valkeila. 2008. An Introduction to the Theory of Point Processes, Volume II: General Theory and Structure, by Daryl J. Daley, David Vere-Jones.

[21] Haoyun Wang, Liyan Xie, Alex Cuozzo, Simon Mak, and Yao Xie. 2020. Uncertainty Quantification for Inferring Hawkes Networks. *Advances in Neural Information Processing Systems* 33 (2020).

[22] Florian Wenzel, Jasper Snoek, Dustin Tran, and Rodolphe Jenatton. 2020. Hyperparameter ensembles for robustness and uncertainty quantification. *Advances in Neural Information Processing Systems* 33 (2020), 6514–6527.

[23] Andrew G Wilson and Pavel Izmailov. 2020. Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems* 33 (2020), 4697–4708.

[24] Zhiliang Wu, Yinchong Yang, Peter A Fashing, and Volker Tresp. 2021. Uncertainty-Aware Time-to-Event Prediction using Deep Kernel Accelerated Failure Time Models. In *Machine Learning for Healthcare Conference*. PMLR, 54–79.

[25] Hongteng Xu, Dixin Luo, Xu Chen, and Lawrence Carin. 2018. Benefits from superposed hawkes processes. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 623–631.

[26] Shingo Yashima, Teppei Suzuki, Kohta Ishikawa, Ikuro Sato, and Rei Kawakami. 2022. Feature Space Particle Inference for Neural Network Ensembles. *arXiv preprint arXiv:2206.00944* (2022).

[27] Jize Zhang, Bhavya Kailkhura, and T Yong-Jin Han. 2020. Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning. In *International conference on machine learning*. PMLR, 11117–11128.

[28] R Zhang, C Walder, MA Rizoiu, and L Xie. 2019. Efficient non-parametric Bayesian hawkes processes. In *IJCAI International Joint Conference on Artificial Intelligence*.

[29] Rui Zhang, Christian Walder, and Marian-Andrei Rizoiu. 2020. Variational inference for sparse gaussian process modulated hawkes process. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 6803–6810.