Deep Learning Based Crime Prediction Models: Experiments and Analysis

Preprint · July 2024

DOI: 10.48550/arXiv.2407.19324

5 authors, including:

Tanzima Hashem
Bangladesh University of Engineering and Technology
71 PUBLICATIONS **876** CITATIONS

DEEP LEARNING BASED CRIME PREDICTION MODELS: EXPERIMENTS AND ANALYSIS

A PREPRINT

Rittik Basak Utsha[1], **Muhtasim Noor Alif**[1], **Yeasir Rayhan**[2], **Tanzima Hashem**[1], **and Mohammad Eunus Ali**[1]
[1]Department of Computer Science and Engineering, Bangladesh University of Engineering and Technology, ECE Building, Dhaka, 1000, Bangladesh
[2] Purdue University, West Lafayette, IN

July 30, 2024

ABSTRACT

Crime prediction is a widely studied research problem due to its importance in ensuring safety of city dwellers. Starting from statistical and classical machine learning based crime prediction methods, in recent years researchers have focused on exploiting deep learning based models for crime prediction. Deep learning based crime prediction models use complex architectures to capture the latent features in the crime data, and outperform the statistical and classical machine learning based crime prediction methods. However, there is a significant research gap in existing research on the applicability of different models in different real-life scenarios as no longitudinal study exists comparing all these approaches in a unified setting. In this paper, we conduct a comprehensive experimental evaluation of all major state-of-the-art deep learning based crime prediction models. Our evaluation provides several key insights on the pros and cons of these models, which enables us to select the most suitable models for different application scenarios. Based on the findings, we further recommend certain design practices that should be taken into account while building future deep learning based crime prediction models.

Keywords Crime prediction · Deep learning · Experimental analysis · Performance evaluation

1 Introduction

Crime prediction problem has been extensively studied in the literature. Early prediction of crime helps the security enforcing authorities to take preventive measures. The task of crime prediction involves analyzing a city, region, or space to forecast the likelihood of a specific type of crime occurring or the number of crimes that might happen, based on past crime data in that area. Modeling crime patterns in a target area is a complex task because various factors can influence them. For instance, locations like shopping malls, businesses, and universities may experience higher crime rates due to the large number of people frequenting these areas. This is referred as the spatial dependency of crimes. In our paper, we refer to these places as points of interest (POIs). Additionally, crime rates in a region can vary depending on the time of day, month, or year. Urban areas with poor lighting and security measures at night may attract criminals. This dependency of crime on time is referred as temporal dependency. Moreover, crime patterns in one region can be affected by neighboring regions; one type of crime may increase the likelihood of another. For instance, a robbery might lead to a hit-and-run incident. The correlations of one crime to another is referred as categorical dependency. These dependencies are illustrated in Figure 1. These factors significantly impact crime rates in a region, making it essential for state-of-the-art models to consider them when modeling crime patterns. Researchers have developed statistical and classical machine learning based crime prediction methods Chen et al. [2008], Breiman et al. [1984], Friedman [2001], Zhang et al. [2017], Bahdanau et al. [2015] in the last decade. To further improve the prediction accuracy, recent research have considered developing deep learning based crime prediction models: DeepCrime (DC) Huang et al. [2018], MIST Huang et al. [2019], CrimeForecaster (CF) Sun et al. [2020], HAGEN Wang et al. [2022], ST-SHN Xia et al. [2021] and ST-HSL Li et al. [2022], AIST Rayhan and Hashem [2023] for crime prediction. As these approaches use a wide variety of settings in their experiments, and no longitudinal study exists in an unified environment, it is quite

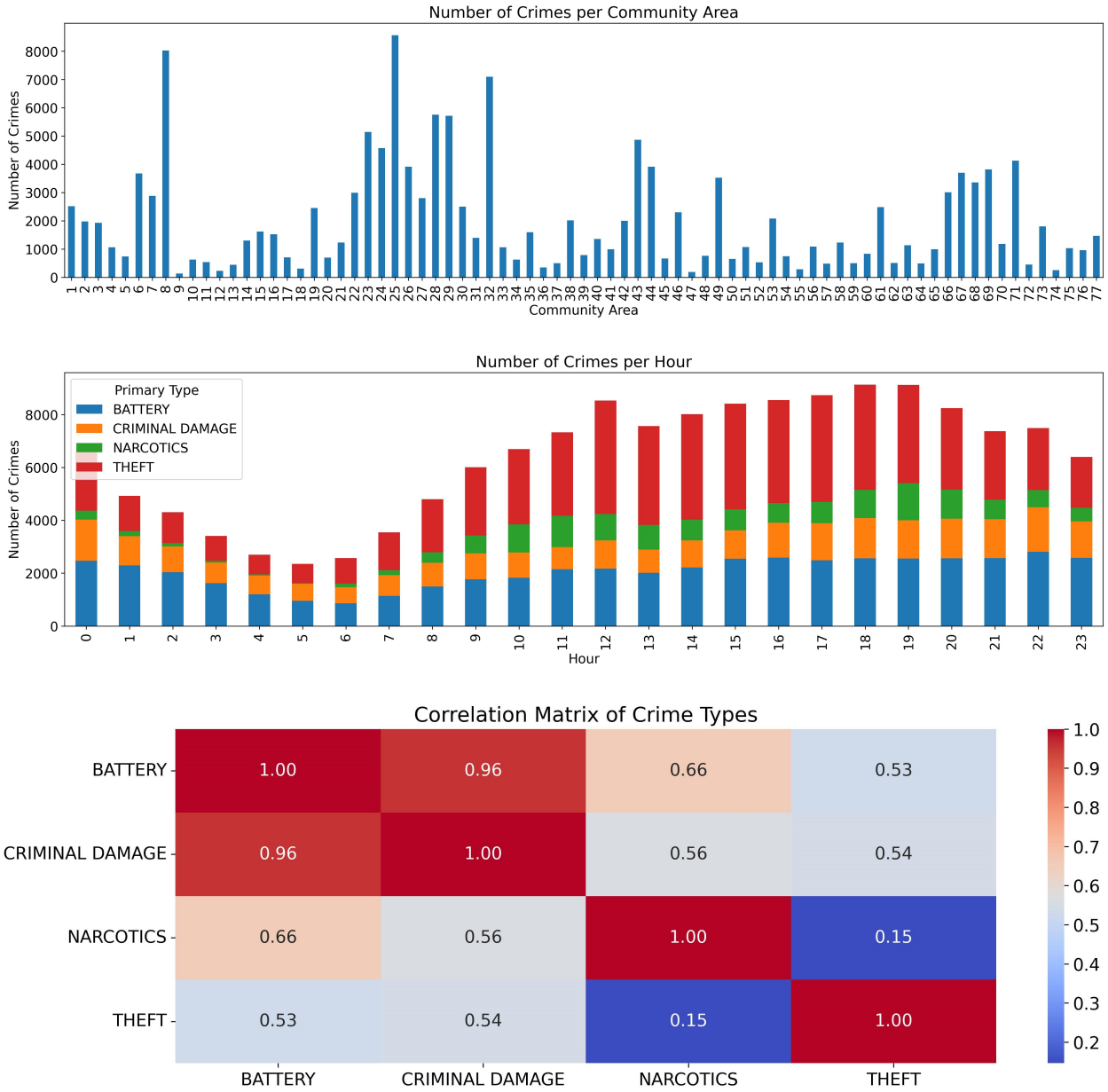


Figure 1: (a) Spatial Dependency: Crimes can exhibit different patterns for different regions. The communities of Chicago exhibit a wide range of number of crimes in 2019. (b) Temporal Dependency: Crime frequency can be different on different time of the day. The earlier hours of a day have less occurrences of crimes than later in the day in our dataset. (c) Categorical Dependency: One crime can be dependent on another. A heatmap depicting the correlations between the crime categories in our dataset is shown.

challenging to identify the most suitable model for a particular scenario, which limits the applicability of these models in a real-life environment. To overcome this limitation of existing research, we perform an experimental analysis of the proposed seven deep learning based models, and identify the key insights and the pros and cons of these state-of-the models, which enable us to find the most suitable model in different real-life settings.

The motivation behind our experimental study are as follows.

Missing competitors (M1). The deep learning based crime prediction models under consideration compare themselves with one or two other deep learning based models in experiments (Table 1). For example, AIST was compared with MIST and DeepCrime, HAGEN was compared with MIST and CrimeForecaster, and ST-SHN was compared with

DeepCrime. However, all of the models claim their superiority over statistical and classical machine learning based crime prediction methods. As a result, there is a need to compare all of the deep learning based models to know their actual performance.

Missing experiments (M2). The deep learning based crime prediction models only show their performance for different crime categories. They fail to answer questions like how does a model perform if crime data density or the region area or the time interval for which the crime occurrence is predicted vary. Some deep learning based crime prediction models Li et al. [2022], Xia et al. [2021], Rayhan and Hashem [2023] consider predicting the number of crime occurrences, whereas others Huang et al. [2018, 2019], Sun et al. [2020], Wang et al. [2022] predict whether a crime would occur or not for a region at a particular time interval.

Lack of uniform experiment settings (M3). The deep learning based crime prediction models in our study do not follow any uniform experiment settings. For example, AIST considers 4 hour time interval for crime prediction, whereas ST-SHN uses 24 hour time interval. Furthermore, the models also use different datasets in experiments.

Though there are a few studies (Wu et al. [2020], Zhang et al. [2020], Jenga et al. [2023], Mandalapu et al. [2023]) that compare crime prediction methods in the literature, none of them perform experimental analysis for the recent deep learning models. A study in Wu et al. [2020] presents a comparison among Bayesian networks, random trees, and neural networks for analyzing crime patterns. In another study Zhang et al. [2020], various classic machine learning based crime prediction algorithms, including K-nearest neighbors (KNN) Cover and Hart [1967], Support Vector Machines (SVM) Vapnik and Cortes [1995], Long Short-Term Memory (LSTM) Hochreiter and Schmidhuber [1997a], and Convolutional Neural Networks (CNN) LeCun et al. [1998] were compared. Both Jenga et al. [2023] and Mandalapu et al. [2023] present surveys of existing crime prediction methods and find their trends. They do not perform any experimental analysis.

Table 1: Considered baselines for deep learning based crime prediction models in experiments. (A blank cell denotes that the two models did not compare with each other, and a hyphen marked cell represents that there were no scope for comparisons as the model appeared earlier.)

	DC	MiST	CF	HAGEN	ST-SHN	ST-HSL	AIST
DC (Huang *et al.* 2018) Huang et al. [2018]	-	-	-	-	-	-	-
MiST (Huang *et al.* 2019) Huang et al. [2019]		-	-	-	-	-	-
CF (Sun *et al.* 2019) Sun et al. [2020]		✓	-	-	-	-	-
HAGEN (Wang *et al.* 2021) Wang et al. [2022]		✓	✓	-	-	-	-
ST-SHN (Xia *et al.* 2021) Xia et al. [2021]	✓				-	-	-
ST-HSL (Li *et al.* 2022) Li et al. [2022]	✓				✓	-	-
AIST (Rayhan & Hashem 2023) Rayhan and Hashem [2023]	✓	✓					-

In this paper, we perform a comprehensive experiment analysis of deep learning based crime prediction models under a unified experiment setting. Specifically, we evaluate the performance of the models by varying the size, crime data density of the target region. We further evaluate the models while varying the granularity of the temporal precision of the predictions. With these categorization, we aim to do an exhaustive search over all possible scenarios in terms of the geographical property, sparsity of crime data, temporal granularity and evaluate the models under different scenarios. This approach further helps us gain insights into the most suitable neural network architecture for different scenarios.

Our ***key findings*** include when crime data is very sparse, models (e.g., AIST) that attempt to capture the interaction between external features and crime data tend to perform better for regression task. However, as the sparsity of crime data tends to decrease, models (e.g., CrimeForecaster, HAGEN) benefit from utilizing the information from regions with similar crime profile. We observe the same phenomenon when we vary the area of the target regions due to the positive correlation between the area of a target region and its crime occurrences. Contrary to regression for classification task, we find that models that explicitly capture the different temporal trends of crime data, i.e., recent, daily and weekly tend to perform better for classification task across all possible scenarios presented in our experimental setting.

Through a detailed experimental study and critical comparative analysis of the results, we ***pinpoint eight critical questions*** that are of utmost importance for any crime prediction system and answer them one by one. Refer to Section 5 for the questionnaire list. These include figuring out (i) the best performing model across all scenarios for regression (**Q1**) and classification task (**Q2**), (ii) the best performing model under specific scenarios, i.e., precise temporal precision of prediction (**Q3**), very sparse crime data (**Q4**), (iii) evaluating the necessity (**Q5**) and utilization techniques (**Q6**) of external features, (iv) impact of capturing spatial dependencies (**Q7**), and (v) the desired model characteristics of a crime prediction model for regression and classification task (**Q8**).

Based on these findings, we further ***recommend*** certain design practices that should be taken into account while building deep learning based crime prediction models in terms of (i) the modeling of spatio-temporal correlation (**R1**) of crime data, (ii) prediction in the absence of external features (**R2**), (iii) utilization schemes of external features when present

(**R3**), (iv) desiderata of a crime prediction model for regression task (**R4**), (v) desiderata of a crime prediction model for classification task (**R5**).

In summary we have made the following contributions.

- We have done a critical analysis of different deep learning architectures and components used in these crime prediction approaches, where we have specifically identified commonness and differences of these architectures. We have also analyzed the models in terms of different characteristics such as used crime data features, considered spatial and non-spatial neighborhood, and prediction types (Section 2).
- We have performed a comprehensive experiment analysis of deep learning based crime prediction models in a unified experimental setting. We have evaluated the performance of the models by varying the crime data density, the area of the region and the time interval for which the crime occurrence is predicted to find the best model for each scenario (Section 3)
- We have summarized the findings with respect to answers of key questionnaire that can be vital in choosing crime prediction models (Section 5) and provide a set of recommendations (Section 6).

2 Deep Learning Models for Crime Prediction

Recently deep learning models have become popular in many domains like image processing, speech recognition and natural language processing. Deep learning models have also been recently used to capture the non-linear spatio-temporal dependencies of crime data for better crime prediction performance Huang et al. [2018], Wang et al. [2022], Rayhan and Hashem [2023]. In this section, we first present an overview of the architecture of seven major deep learning based crime prediction models: DeepCrime Huang et al. [2018], MIST Huang et al. [2019], CrimeForecaster Sun et al. [2020], HAGEN Wang et al. [2022], ST-SHN Xia et al. [2021] and ST-HSL Li et al. [2022], AIST Rayhan and Hashem [2023], that we have selected for our experimental evaluation (Section 2.1). We then critically analyze those models to find similarities and dissimilarities (Section 2.2).

2.1 Model Architectures

In this section, we discuss deep learning architectures used in different crime prediction models. The high-level diagram of these architectures is shown in Figure 2.

2.1.1 DeepCrime

DeepCrime Huang et al. [2018] is one of the first models to employ deep learning for crime prediction. The model claims to capture the dynamic patterns of crime and their correlation with other influencing data (POI, anomaly) across different time steps.

Figure 2a depicts the high level architecture of DeepCrime. In its architecture, the intra-region and intra-crime correlations are first embedded, and these embeddings are fed through a MLP to encode the region-category interactions in a weighted vector. The POI information is used while creating the region embedding. Then a hierarchical recurrent framework with three Gated Recurrent Units (GRU) Chung et al. [2014] is used to encode the temporal dependencies of crime sequence, anomaly sequence, and their inter-dependencies. Next an attention layer Vaswani et al. [2017] is introduced that models the influence of past crimes for the prediction of future crimes. Finally, a MLP network maps the learned vectors to output the crime probability. DeepCrime authors compare their model with traditional ML models such as SVR Chang and Lin [2011], ARIMA Chen et al. [2008], LR Hosmer Jr et al. [2013], and GRU Chung et al. [2014] with NYC crime data nyc [2017] of 2014.

2.1.2 MiST: Multi-View Deep Spatial-Temporal Network

MiST Huang et al. [2019] is a model designed to predict the occurrence of spatial-temporal abnormal events, such as crimes, urban anomalies, etc. Figure 2b shows the overview of the MiST architecture. MiST first divides the target city with a grid-based map segmentation. With the data mapped into the cells of this grid, MiST first employs a Long Short-Term Memory (LSTM) Hochreiter and Schmidhuber [1997b] based network, named "Context Aware Recurrent Framework", to encode the time-dependent nature of the data. Then the inter-region and cross-category correlations are captured through the use of an attention mechanism, named "Multi-Modal Pattern Fusion Module." After that, the complex interactions between the spatial-categorical fusion module and temporal recurrent module are integrated through another recurrent module, named "Conclusive Recurrent Network". Finally, the output of the recurrent module are fed through a MLP network to generate the occurrence probabilities.

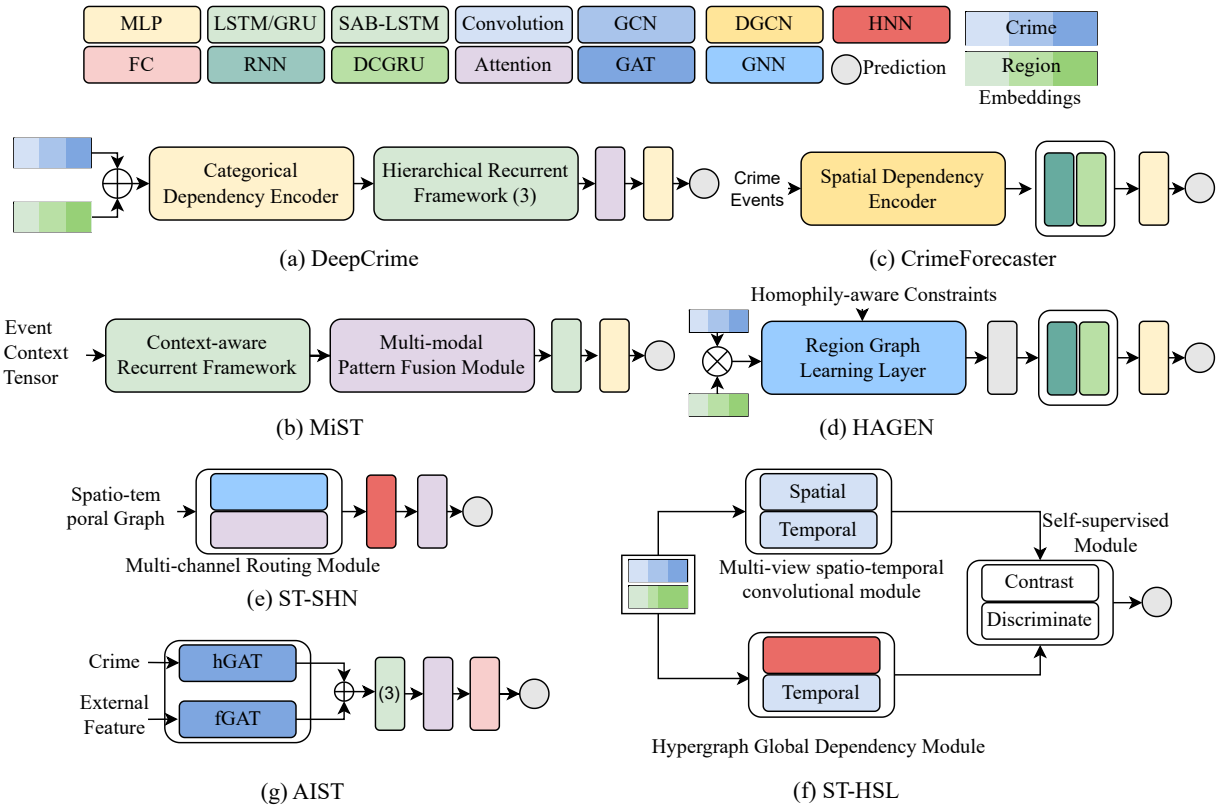


Figure 2: The Architectural Synopses of Deep Learning Based Crime Prediction Models. (MLP: Multi-Layer Perceptron, FC: Fully Connected NN, LSTM: Long Short-Term Memory, GRU: Gated Recurrent Unit, RNN: Recurrent Neural Network, SAB-LSTM: Sparse Attention-Based LSTM, DCGRU: Diffusion Convolution GRU, GCN: Graph Convolutional Network, GAT: Graph Attention Network, DGCN: Diffusion GCN, GNN: Graph Neural Network, HNN: Hypergraph Neural Network)

The authors experimented their model with NYC Crime data nyc [2017], NYC Urban Anomaly Data and Chicago Crime Data dat [2019].

MiST is compared against various traditional ML models such as SVR Chang and Lin [2011], ARIMA Chen et al. [2008] and some early deep learning models such as ST-RNN Liu et al. [2016], RDN Hu et al. [2017], and ARM Feng et al. [2018].

2.1.3 CrimeForecaster

Crimeforecaster Sun et al. [2020] authors argue that the spatial nature of crime depends on the temporal nature, i.e. a community can have different crime patterns depending on the time of the year, month, week or day. They claim that traditional methods process the spatial and temporal information separately, hence neglect the spatio-temporal dependency. Crimeforecaster models this spatio-temporal dependency using a Diffusion Convolution module. Figure 2c shows a high level diagram of CrimeForecaster. CrimeForecaster represents the neighborhoods and learns their correlations through a graph, and then uses diffusion convolution on that graph to predict crimes. CrimeForecaster uses a DCGRU (Diffusion Convolutional Gated Recurrent Units) Li et al. [2018] encoder to learn the complex intra and inter-region correlations across the previous time slots. The diffusion convolution operation models the spreading nature of crimes to nearby and similar regions with time.

2.1.4 HAGEN: Homophily-Aware Graph Convolutional Recurrent Network

HAGEN Wang et al. [2022] authors claim that two areas may exhibit similar crime patterns if they share some common traits such as proximity in geographical distance or similar points of interest. Instead of creating individualized graphs for every instance, HAGEN suggests identifying similar regions in the graph using adaptive graph learning.

Table 2: Components used by the models to capture various aspects of crime data

Model	Categorical Dependency	Spatial Dependency	Temporal Dependency
DC	MLP		GRU and Attention
MiST	Attention		LSTM
CF	DCGRU		
HAGEN	MLP, Graph Learning		DCGRU
ST-SHN	Message Passing, Attention	Hypergraph	Message Passing, Attention
ST-HSL	Convolution (local)		
	Hypergraph (global)		
AIST	hGAT	hGAT, fGAT	SAB-LSTM

Figure 2d shows the high level diagram of HAGEN. It first applies a region-crime dependency encoder, which under the hood learns the graph adaptively, using a homophily-aware constraint. Crime and region embeddings are utilized to jointly capture the interactions between regions and crimes. Region embeddings are created based on geographical distances and points of interests. After this layer, the temporal dependency of the crimes are caputred using a diffusion convolutional recurrent module, namely DCGRU, just like Crimeforecaster. The final predictions is obtained by an MLP-based decoder.

The authors used almost similar experimental setup as CrimeForecaster Sun et al. [2020]. They claim that HAGEN outperforms CrimeForecaster and other competitive models.

2.1.5 ST-SHN: Spatial-Temporal Sequential Hypergraph Network for Crime Prediction with Dynamic Multiplex Relation Learning

Xia et al. Xia et al. [2021] suggest using a graph-based method for message passing among different regions for different crime categories, incorporating hypergraph learning to address spatial and temporal changes within a broad context.

ST-SHN first generates a region graph representing the regions and their geographical adjacencies. This graph is utilized in the "Spatial Dependency Encoder" module, which captures the complex spatial dependencies among regions regarding various crime types. It incorporates a message propagation approach called the "Multi-Channel Routing Mechanism" to model how different types of crimes influence each region. It employs a multiplex mutual attention network under the hood. To further improve cross-region learning, a hypergraph neural network named "Cross-Region Hypergraph Relation Learning" is introduced to understand the broader relationship between crimes. Temporal dependencies of crimes are handled similarly to spatial dependencies, with a network resembling the "Spatial Dependency Encoder" that spreads temporal messages within and across regions and different crime categories. The overall architecture of ST-SHN is shown in Figure 2e.

ST-SHN has been evaluated and compared to other baseline models like ARIMA Chen et al. [2008], DeepCrime Huang et al. [2018], DCRNN Li et al. [2018], GMAN Zheng et al. [2020] using both the NYC and Chicago crime data, and shown to have superior performance over these models.

2.1.6 ST-HSL: Spatial-Temporal Hypergraph Self-Supervised Learning for Crime Prediction

Li et al. Li et al. [2022] claim that the problem in the state-of-the-art crime prediction models are that they all follow supervised learning methods, which require labeled data. But in real life scenario, the labels are very scarce compared to the vastness of a city. This scarcity of labeled data creates a challenge in effectively training these models effectively for real-world datasets. This is why the authors of ST-HSL proposes a dual-stage self-supervised learning paradigm.

ST-HSL encodes the spatial and temporal interactions of crimes between geographically neighboring regions through a convolutional module, named "Multi-View Spatial-Temporal Convolution", creating a local crime embedding. At the same time, it devises a hypergraph structure and employs a hypergraph-guided message passing learning framework to create a global crime embedding. A temporal convolution network is fused with the global module to inject temporal context to it. With the global and local embeddings, a "Dual-Stage Self-Supervision Learning Paradigm" is designed to tackle the challenge of data sparsity by self-supervised learning. One stage of this module generates a corrupt crime embedding and tries to discriminate between the original and corrupt graph. The other stage of the module enhances the training process by finding the contrast between the local and global embedding of crimes. The architecture of ST-HSL is illustrated in Figure 2f.

Extensive experiments show that ST-HSL performs better than other state-of-the-art models such as ARIMA Chen et al. [2008], DCRNN Li et al. [2018], GMAN Zheng et al. [2020] and deep learning models like ST-SHN Xia et al. [2021], DeepCrime Huang et al. [2018] for crime prediction.

2.1.7 AIST

Rayhan and Hashem Rayhan and Hashem [2023] claim that the current deep learning models from crime predictions are not interpretable; they also fail to address the long term temporal correlation, and do not effectively incorporate external features. Hence they propose an attention based deep learning model, namely AIST that uses external features like taxi flow and point of interest along with past crime data to predict crime occurrence. The high level diagram of AIST is shown in Figure 2g.

AIST uses neighbourhood graph and hierarchical structure of the regions to capture the spatial dependency of the crime data. A variant of Graph Attention Network Velickovic et al. [2018], hGAT to learn region crime embedding by incorporating the hierarchical information of diferent regions. Another variant, fGAT is used to embed the external features into the model. This crime embedding and feature embedding are concatenated to get the spatial embedding. The spatial representation generated at different time-steps are then fed into three Sparse Attention Based LSTM (SAB-LSTM) to capture the recent, daily and weekly trends of the crime data, thereby capturing the temporal dependency into a final weight vector that is then used to predict the crime occurrence at the next time-step.

The authors compared AIST with various other models like DeepCrime Huang et al. [2018], MiST Huang et al. [2019], STGCN Yu et al. [2018] and found its superior performance them in terms of prediction accuracy. Also, attention weights associated with different parts of the model can be exploited to interpret its prediction.

2.2 Comparative Analysis

In this section, we present a comprehensive comparative analysis of our crime prediction models on several key aspects: different architectural components, variants of data types employed by each model, different data features incorporated into their architectures, and the types of predictions made by these models.

2.2.1 Sub-components of Model Architectures

The crime prediction problem spans three dimensions: space, time, and category. Therefore, deep learning models focus on effectively modeling the interactions among these three aspects to produce accurate predictions. Table 2 summarizes the architectural synopsis of the above seven models in the context of categorical, spatial and temporal dependencies. Previous models like DeepCrime and MiST attempt to capture spatial and categorical dependencies using MLP and attention-based layers, respectively. However, these models do not account for the graph-like relationships between regions, potentially missing important geographical semantics. CrimeForecaster, HAGEN, and AIST address this by using graph neural networks to model spatial interactions. ST-SHN and ST-HSL further utilize hypergraph neural networks to encode spatial correlations while considering the global dependency of crime among different regions. Temporal dependencies are typically captured by recurrent network variants such as GRU and LSTM, with attention mechanisms commonly used to identify important focal points in space or time. Figure 2 depicts the block diagrams of different categories of components using different color codes.

2.2.2 Utilization of Crime and External Datasets

Crime prediction models typically focus on a city, using data on crimes committed within that city. The city may be divided into various regions or communities, with the number of crimes committed in each region during specific time intervals serving as the primary data for training the models. The division of the target city can vary; for instance, MiST divides the city into a grid, while CrimeForecaster considers the city as a graph with communities represented as nodes. Crimes are usually categorized into multiple types. For example, Chicago is divided into 77 different communities, and the count of various crime categories (e.g., murder, burglary, robbery, hijacking) in these communities at 4-hour intervals can form a crime dataset.

Also, crimes in a region can be influenced by external events. Solely depending on the crime data, the models cannot capture these influences properly. So, existing deep learning models incorporate external datasets in addition to crime data of the region to enhance their prediction accuracy. These datasets provide important contextual information that help the models capture the complex factors that influence the crime pattern. The models we are using in this experimental study also utilize various external datasets to predict crime. Table 3 summarizes the datasets used by the models we are considering for this study:

Table 3: Data used by various models

Model	Data Used
DC	Crime data, POI data, Urban anomaly data
MiST	Crime data
CF	Crime data
HAGEN	Crime data, POI data
ST-SHN	Crime data
ST-HSL	Crime data
AIST	Crime data, Taxi trip data, POI data

2.2.3 Different Types of Prediction

Different crime prediction models adopt distinct approaches to tackle the problem. Based on the type of prediction they make, these models can be classified into two groups. The first group, DeepCrime, MiST, CF, and HAGEN, focus on predicting solely the occurrence of a crime, which is categorized as classification. The second group of models, ST-SHN, ST-HSL, and AIST, aim to predict the quantity of crimes that may occur during a specific time step, which is referred to as regression. Only ST-SHN consider both types of predictions.

3 Experimental Settings

This section provides an overview of the experimental setup for our experiments. We outline the datasets employed, evaluation criteria and parameter settings of the models for our experiments in the following subsections.

3.1 Dataset

We conduct all the experiments on 2019 Chicago crime data for the following four crime categories: theft, criminal damage, battery and narcotics. In addition to the crime data, we include external feature datasets, e.g., POI information, taxi flow and urban anomaly data as required by the respective models, the details of which are presented in Table 4. We divide the dataset such that first 8 months (January to August) are used for training, next 1 month (September) is used for validating and the data for last 3 months (October to December) are used for testing.

We choose Chicago as our target city as most models have used Chicago crime dataset themselves and due to the availability of all the external features required to train the respective prediction models. On top of that, we claim that our categorization of Chicago communities into different groups based on the crime density can capture the properties exhibited by other target cities, i.e., Los Angeles, New York City.

3.2 Parameter Settings

Refer to Table 5 for the parameter settings of the seven models under different scenarios.

3.3 Evaluation Metrics

We use mean average error (MAE) and rooted mean square error (MSE) to evaluate the prediction models on the regression task.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

We use Macro-F1 and Micro-F1 to evaluate the prediction models on the classification task.

$$\text{Micro-F1} = \frac{1}{J}\sum_{j=1}^{J}\frac{2TP_j}{2TP_j + FN_j + FP_j}$$

$$\text{Macro-F1} = \frac{2\sum_{j=1}^{J} TP_j}{2\sum_{j=1}^{J} TP_j + \sum_{j=1}^{J} FN_j + \sum_{i=j}^{J} FP_j}$$

Table 4: Details of the Chicago datasets

Datasets	Category	#Records
Crime Dataset (2019) dat [2019]	Theft	62484
	Criminal Damage	26681
	Battery	49513
	Narcotics	15069
311 Public Service Complaint Data (2019) 311 [2019]	Blocked Driveway	51939
	Noise	43542
	Building/Use	32235
	Safety	1415
POI data (2019) POI [2019]	Food	12532
	Residence	4236
	Travel	13863
	Arts & entertainment	4780
	Outdoors	5022
	Recreation	3782
	Education	5646
	Nightlife	21014
	Professional	18402
	Shops and event	47
Taxi Trip Data (2019) of Chicago Portal [2019]	Taxi inflow	14557888
	Taxi outflow	14552209

Table 5: Parameter settings for the models(B=Batch size, A=Attention Dimension size, L=#MLP layers, E=Embedding size, lr=Learning rate, Hsd=Dimension of hidden state, Ep=#Epochs, Rl=#RNN layers, Ds=Diffusion step, Drl=Dimension of RNN layers, Dr=Decay rate, Ss=Subgraph size, Sr=Saturation rate, H=#Hyperedges, Ks=Kernel size, Cl=#CNN layers, Rt=#Recent Timesteps, Hsl=Hidden state in SAB-LSTM)

Model	Parameter Name	Experiment Name	
		Area/Density	Temporal Granularity
DC	(B, A, L, E, lr)	(10, 64, 3, 128, 5e-4)	(10, 64, 3, 128, 1e-4)
MiST	(Hsd, E, A, B, lr, Ep)	(32, 32, 32, 64, 1e-3, 150)	(32, 32, 32, 32, 1e-3, 200)
CF	(B, Rl, Ds, Drl, lr, Dr)	(64, 3, 2, 64, 1e-2, 0.1)	(64, 6, 4, 64, 1e-2, 0.1)
HAGEN	(Rl, Drl, Ss, Sr, lr, Dr)	(3, 64, 50, 3, 1e-2, 0.1)	(3, 64, 50, 3, 1e-2, 0.1)
ST-SHN	(B, H, lr, Dr)	(16, 128, 1e-3, 0.96)	(32, 128, 1e-3, 0.96)
ST-HSL	(B, Cl, Ks, H, lr)	(16, 4, 3, 128, 1e-3)	(32, 4, 3, 128, 1e-3)
AIST	(B, Rt, Hsl, Ep, lr)	(42, 20, 40, 180, 1e-3)	(42, 20, 40, 120, 1e-3)

Here, Here, n represents the number of predictions, y_i represents the ground truth and $\hat{y}_i$ represents the predicted result., whereas $J = 2$ represents the total number of classes and $TP_j, FP_j \, and \, FN_j$ denote the number of true positive, false positive, false negative values in each class, respectively.

3.4 Evaluation Criteria

Our experiments are categorized into three criteria, which help us assess the performance of the models under under specific conditions and determining their performance. These criteria not only help identify models that are more likely to perform better but also provide insights into the most suitable architectures for different scenarios. The three criteria are as follows.

Evaluation based on area. Chicago has 77 communities with different sizes from very large (O'Hare - 34.55 sq km) to very small (Oakland - 1.5 sq km). Our aim is to evaluate if geographical properties, i.e, area has an impact on the prediction performance of these deep learning models

Evaluation based on crime density. While it is generally true that larger areas tend to have higher crime rates, there maybe exceptions to this observation as well. In some cases, smaller areas with significant points of interest can exhibit a high density of crimes. Therefore, we introduce a new criterion, crime density, to assess the performance of our models in capturing these variations. To compute the crime density of a community, we divide the total number of

Table 6: Grouping criteria of the Chicago communities based on area.

Group	Area Range (Sq. km)	#Communities	#Crimes
Very Small	< 4	13	18070
Small	$4 - 6$	17	47813
Medium	$6 - 8$	15	52979
Large	$8 - 10$	18	78933
Very Large	> 10	14	63409

crimes by its area. Formally,

$$\text{Crime density of a community} = \frac{\text{\#Crimes of the community}}{\text{Area of the community}}$$

Evaluation on prediction interval. The models are evaluated on different granularity of prediction time intervals for all Chicago communities. We try to evaluate how accurate the models' prediction performance is under under different required temporal precision.

4 Experiment Results

4.1 Evaluation on the Area of the Target Region

In this experiment, we aim to assess the impact of the target region's size while predicting crimes. We further want to compare the effectiveness of these models when the size of the target regions vary and identify the models that excel across all different sized target regions. A thorough investigation of the 77 communities in Chicago was carried out to identify areas where the number of communities in each division closely aligns. Five groups were chosen to maintain consistent area ranges while ensuring close community counts within each division. The five groups based on area are as follows: (a) very small, (b) small, (c) medium, (d) large, and (e) very large. Refer to Table 6 for the grouping criteria.

4.1.1 Performance Comparison for Regression Task

- It is evident from table 7 that the models tend to perform worse as the community size increases. With the increasing community size, the number of crimes increase too (Refer to Table 6). With the increasing number of crimes, it becomes difficult for the models to predict the exact number of crimes. This can be the reason the models' gradually worse performance for larger areas.
- Table 7 shows that AIST and HAGEN are the best performing models across all 5 groups. AIST performs better than all the other models including HAGEN with regards to the MAE metric. This suggests that the graph attention layer used to model the spatial dependencies and capture the crime-external feature interaction captures the crime distribution quite well. However, when it comes to capturing the sudden spikes of crime distribution, HAGEN's Homophily-aware Graph Diffusion Convolution architecture performs comparatively better. It can be attributed to the fact HAGEN not only captures the spatial dependencies of the neighboring regions, it further utilizes the spatial embedding of the regions with same crime profile. On the contrary, AIST only considers the spatial dependencies of the important neighboring regions ignoring the distant regions.

  AIST outperforms HAGEN in the MAE metric for all area categories. However, as the area size increases, HAGEN outperforms AIST in the RMSE metric. This suggests that AIST is less sensitive to outliers, with its errors being more uniformly distributed and not having extreme values that significantly skew the error metric. On the other hand, HAGEN's better performance in RMSE indicates that it handles large errors better. This model likely has fewer large outliers or predicts more accurately most of the time but occasionally makes larger mistakes. HAGEN's better performance can be attributed to the fact HAGEN not only captures the spatial dependencies of the neighboring regions, it further utilizes the spatial embedding of the regions with similar profile. On the contrary, AIST only considers the spatial dependencies of the important neighboring regions ignoring the distant regions.
- CrimeForecaster stands out as the next best model across these different categories. CrimeForecaster uses the same graph diffusion convolution network architecture as HAGEN to capture the spatial dependencies, however it does not consider the homophily learning approach thus ignoring the distant regions. This explains the downgrade in performance of CrimeForecaster compared to HAGEN for the RMSE metric.
- ST-SHN's use of hypergraphs to address spatial dependencies does not yield better performance for small sized communities. But as the area gets larger, i.e., the number of crimes increase, it performs comparatively better.

- ST-HSL applies a self-supervised learning mechanism to backup its performance when the data is sparse. This makes the model invariant to area or density.
- For MiST, the MAE is relatively stable across different area sizes, indicating consistent performance. The RMSE is relatively stable but increases slightly in medium and very large areas.
- DeepCrime models spatial dependency using an MLP network, which fails to capture the spatial correlations among regions as effectively as the more advanced spatial modules in other models. Consequently, as the area of the target region increases, its performance gradually deteriorates.

Table 7: Regression metrics for groups based on area.

Model	Criteria	Very Small	Small	Medium	Large	Very Large
DC	MAE	0.80	1.08	1.32	1.36	1.49
	RMSE	1.03	2.05	2.60	2.10	2.60
MiST	MAE	0.85	0.73	0.76	0.74	0.73
	RMSE	0.92	0.85	0.87	0.86	0.86
CF	MAE	0.58	0.54	0.54	0.53	0.54
	RMSE	0.66	0.61	0.62	0.60	0.61
HAGEN	MAE	0.46	0.50	0.54	0.52	0.51
	RMSE	0.52	**0.55**	**0.54**	**0.55**	**0.55**
ST-SHN	MAE	0.92	0.65	0.77	0.68	0.63
	RMSE	1.34	0.88	1.53	0.99	0.80
ST-HSL	MAE	1.00	1.01	1.01	1.01	1.02
	RMSE	1.02	1.05	1.03	1.04	1.04
AIST	MAE	**0.10**	**0.35**	**0.38**	**0.39**	**0.42**
	RMSE	**0.35**	0.77	0.82	0.83	0.90

4.1.2 Performance Comparison for Classification Task

- Contrary to our findings for regression task Table 8 shows that models tend to perform better as the community size increases. As evident from Table 6 the number of crimes occurrences are comparatively larger in big communities. In larger areas, crime patterns might be more aggregated, making it easier for models to identify and classify crime hotspots. The larger number of crimes provides more data points, leading to better training and higher performance in classification. On the other hand, regression tasks require predicting exact numbers, which is more sensitive to outliers and noise in the data. In larger areas, the complexity and variability increase, making it difficult for models to capture the exact relationships.
- AIST performs the best across all the groups for both metrics, Macro-F1 and Micro-F1. This is due to the usage of graph attention layers to capture the spatial dependencies and interaction between crime and external features as well as separate temporal modules to capture different trends, i.e., recent, daily and weekly.
- CrimeForecaster and HAGEN exhibit competitive performance, with CrimeForecaster generally performing slightly better in predicting crime occurrences indicating the ineffectiveness of homophily-aware learning paradigm introduced in HAGEN.
- DeepCrime, originally designed for the classification task surprisingly exhibits better performance with its primitive architectures implying the significance of capturing the temporal dynamics and external feature interactions for the classification task.

Findings. For regression task, AIST is the best performing model across all the groups in terms of the MAE score. In terms of RMSE score, it achieves the best performance for smaller areas with very limited crime occurrences. In contrast, HAGEN, performs the best in terms of RMSE score for larger areas with larger crime occurrences. For classification task, AIST is the winner across all the groups for both Macro-F1 and Micro-F1 metrics.

4.2 Evaluation on the Crime Density of the Target Region

Depending on the socio-economic factors associated with individual communities, even a smaller region can face a large number of crimes, whereas a large but sparsely populated community can hardly face any crimes. By dividing the communities based on crime density, we aim to evaluate how limited as well as ample number of crime records impact the prediction performance of these models. We conducted an analysis of the Chicago communities and divided them

Table 8: Classification metrics for groups based on area.

Model	Criteria	Very Small	Small	Medium	Large	Very Large
DC	Macro-F1	0.24	0.30	0.33	0.38	0.42
	Micro-F1	0.36	0.41	0.55	0.56	0.59
MiST	Macro-F1	0.18	0.22	0.28	0.31	0.35
	Micro-F1	0.21	0.34	0.39	0.42	0.45
CF	Macro-F1	0.20	0.39	0.38	0.47	0.43
	Micro-F1	0.23	0.45	0.45	0.53	0.51
HAGEN	Macro-F1	0.25	0.36	0.37	0.42	0.41
	Micro-F1	0.27	0.39	0.41	0.45	0.44
ST-HSL	Macro-F1	0.39	0.37	0.29	0.32	0.38
	Micro-F1	0.44	0.48	0.43	0.47	0.60
AIST	Macro-F1	**0.46**	**0.50**	**0.48**	**0.52**	**0.61**
	Micro-F1	**0.54**	**0.60**	**0.68**	**0.77**	**0.73**

Table 9: Grouping criteria of the Chicago communities based on crime density.

Group	Density Range (crime per sq. km)	#Communities
Very low	< 150	12
Low	$150 - 300$	16
Medium	$300 - 450$	19
High	$450 - 600$	9
Very high	> 600	21

into five groups ((a) very low, (b) low, (c) medium, (d) high and (e) large) with equally distributed crime densities, each containing a moderate number of communities. Refer to Table 9 for the categorization criteria.

Table 10: Regression metrics for groups based on crime density.

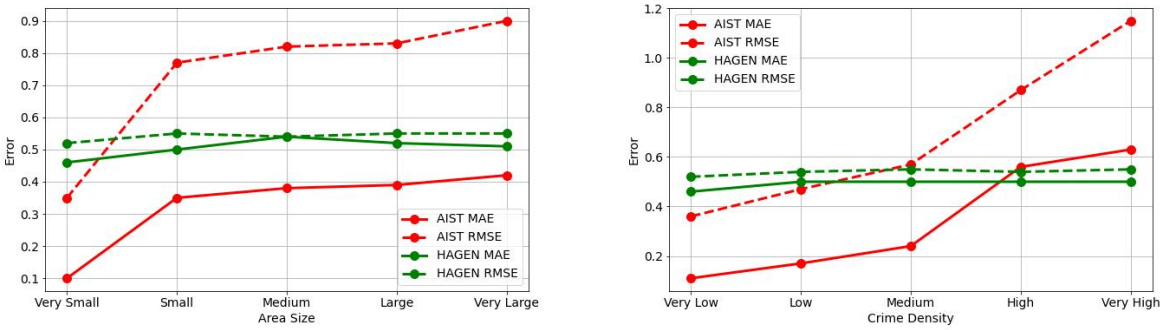
Model	Criteria	Very low	Low	Medium	High	Very high
DC	MAE	0.75	0.76	0.86	1.20	1.68
	RMSE	0.91	0.97	1.15	1.84	2.65
MiST	MAE	0.91	0.87	0.70	0.69	0.67
	RMSE	0.95	0.93	0.84	0.83	0.82
CF	MAE	0.57	0.56	0.54	0.53	0.52
	RMSE	0.65	0.65	0.62	0.60	0.58
HAGEN	MAE	0.46	0.50	0.50	**0.50**	**0.50**
	RMSE	0.52	0.54	**0.55**	**0.54**	**0.55**
ST-SHN	MAE	1.00	0.83	0.71	0.63	0.66
	RMSE	1.18	1.54	1.08	0.93	0.91
ST-HSL	MAE	1.01	1.01	1.01	1.01	1.01
	RMSE	1.02	1.03	1.03	1.05	1.06
AIST	MAE	**0.11**	**0.17**	**0.24**	0.56	0.63
	RMSE	**0.36**	**0.47**	0.57	0.87	1.15

4.2.1 Performance Comparison for Regression Task

- With increasing crime density, models exhibit a somewhat worse performance. As the crime density increases, it becomes difficult for the models to predict the exact number of crimes that is likely to occur. Hence, the models show a greater error when trying to predict for higher crime density communities.
- Among the crime prediction models evaluated, AIST and HAGEN are the best performers. Both models use external features and complex architectures that are capable of capturing cross-region and cross-temporal relationships. Particularly, HAGEN exhibits consistent performance across all density groups and shows better performance in high-dense regions. HAGEN utilizes a graph with regions of similar features connected together to effectively capture crime embeddings. So, high density regions can better influence each other that results in improved predictions.
- CrimeForecaster ranks as the next best performing model. It captures cross-region dependencies and temporal correlations but does not use external features, resulting in relatively poorer performance compared to the other models.

- ST-SHN is the only model that performs better in high density regions compared to low density regions. This model uses a hypergraph-based approach to capture the relationship between one region and all other regions. It also incorporates cross-category dependencies. As a result, it obtains better embeddings for higher density regions which lead to improved performance.
- Although MiST attempts to capture spatial, temporal, and categorical co-dependencies, it relies on simple attention and LSTM architectures. Additionally, MiST does not consider any external features in its predictions, which results in it lagging behind other models in terms of performance.
- ST-HSL exhibits consistency across all density groups. Its dual-stage self-supervised learning algorithm addresses the sparsity issue inherent in crime data. Its hypergraph global dependency model also handles the skewed distribution of crime occurrence at geographical regions. These architectures allow the model to maintain performance regardless of regional density.
- DeepCrime, with its basic MLP network for modeling spatial-categorical dependencies, performs the poorest among all the models.
- From Figure 3, it is evident that HAGEN's RMSE and MAE values are very consistent across all area and density classes, with little variation. This indicates that HAGEN performs reliably regardless of the area size and crime density. Both RMSE and MAE values for HAGEN are relatively low and constant, suggesting that HAGEN is an accurate model for predicting across different area sizes and densities. AIST shows an increasing trend in both RMSE and MAE as the area size and crime density increases. This indicates that AIST's performance degrades slightly as the area size or crime density becomes larger. AIST's RMSE values are higher than its MAE values, especially noticeable in larger area sizes or densities. This suggests that AIST is more sensitive to outliers, and has some larger prediction errors that significantly affect the RMSE more than the MAE.

Figure 3: Comparison of regression performance between HAGEN and AIST



(a) Errors vs Area Size

(b) Errors vs Crime Density

Table 11: Classification metrics for groups based on crime density.

Model	Criteria	Very low	Low	Medium	High	Very high
DC	Micro-F1	0.22	0.25	0.28	0.31	0.35
	Macro-F1	0.26	0.29	0.38	0.44	0.49
MiST	Micro-F1	0.15	0.15	0.30	0.3	0.45
	Macro-F1	0.21	0.23	0.26	0.31	0.38
CF	Micro-F1	0.20	0.25	0.35	0.44	0.60
	Macro-F1	0.24	0.30	0.43	0.5	0.65
HAGEN	Micro-F1	0.26	0.28	0.37	0.37	0.50
	Macro-F1	0.29	0.32	0.39	0.40	0.52
ST-HSL	Micro-F1	0.37	0.37	0.36	0.33	0.32
	Macro-F1	0.57	**0.58**	**0.57**	0.52	0.47
AIST	Micro-F1	**0.45**	**0.48**	**0.46**	**0.51**	**0.63**
	Macro-F1	**0.65**	0.57	0.56	**0.73**	**0.86**

4.2.2 Performance Comparison for Classification Task

- Contrary to our observation for regression task, for classification the models' performance get better as crime density increases. With higher crime density, there is more data for the models to incorporate the spatio-temporal and external factors to predict the occurrence of a crime. For this reason, although predicting the exact number of crime becomes more difficult, the prediction of the mere occurrence of a crime becomes easier.
- AIST emerges as the best performing model for predicting the occurrence of crimes across all scenarios. AIST incorporates multiple external features to effectively capture spatial and temporal dependencies, resulting in accurate predictions. Thus, it performs well for both regression and classification approaches.
- In communities characterized by low and medium crime density, ST-HSL performs comparably well to AIST. Its self-supervised learning mechanism for sparse data allows it to provide reliable predictions of crime occurrences when crime data is limited, hence it does comparatively better when crime data is less dense. However, as crime density increases, other models tend to outperform ST-HSL.
- For high and very high density communities, CrimeForecaster stands out as the second-best model for predicting crime occurrences. With its original classification-based approach and the utilization of DCGRU architecture, CrimeForecaster effectively captures spatial and temporal dependencies, leading to accurate predictions of crime occurrences.
- HAGEN, an improvement over CrimeForecaster, shows changing performance depending on the density of the region. It performs worse in low-density regions but surpasses CrimeForecaster for high-density regions. This is because of HAGEN's adaptive learning of region dependencies using similar regions as connected nodes in a graph. In high-density regions, HAGEN benefits from more available data, allowing it to learn features more effectively.
- DeepCrime and MiST perform quite poorly on predicting crime occurrences compared to others. Although these models' performances tend to get better as the crime density increases, other models do even better than them. DeepCrime models the spatial correlation with an MLP networks, and MiST neglects external features. As crime density gets higher, these factors become more crucial in predicting occurrence of crimes, hence the models perform poorly.

Findings. For regression task, when the crime data is sparse, i.e., for regions with very low to medium crime density, AIST performs best in terms of both MAE and RMSE score. On the contrary, when the number of available crime records is high, i.e., for regions with high to very high density HAGEN performs best. However, for classification task, AIST is the sole best performer.

4.3 Evaluation on the Temporal Granularity of Prediction

Crime forecasting models can be trained to predict crimes at different temporal granularity, i.e., hourly, daily, weekly and so on. We evaluate the models based on the following four temporal precision: (a) 4 hours, (b) 6 hours, (c) 12 hours, and (d) 24 hours. Our aim is to evaluate how accurate the models are in predicting crimes at different temporal precision and whether the temporal precision adversely affects a model's prediction performance or not.

Table 12: Regression metrics for different temporal granularity.

Model	Criteria	4h	6h	12h	24h
DC	MAE	0.88	0.93	0.99	1.22
	RMSE	0.87	0.91	0.89	1.68
MiST	MAE	0.37	0.49	0.57	0.76
	RMSE	0.54	0.57	0.67	0.87
CF	MAE	0.59	0.53	0.55	0.54
	RMSE	0.71	0.65	0.64	0.62
HAGEN	MAE	0.50	0.49	0.50	0.49
	RMSE	**0.53**	**0.53**	**0.54**	**0.54**
ST-HSL	MAE	1.02	1.00	1.02	1.01
	RMSE	1.03	1.02	1.03	1.04
AIST	MAE	**0.34**	**0.35**	**0.41**	**0.44**
	RMSE	0.57	0.61	0.65	0.67

4.3.1 Performance Comparison for Regression Task

- A general trend here is that as the temporal granularity becomes coarser the models' performance deteriorate (Refer to Table 12). This is down to the fact that at finer granularity, the number of crimes in adjacent time steps remain more consistent compared to when the temporal granularity is coarser. Thus, it becomes hard for the models to predict crime accurately.
- Similar to previous observations, AIST performs the best across all temporal granularity for the MAE metric. In all time intervals AIST does well in predicting number of crimes. But it comes second best for the RMSE metric. The best performing model across all the temporal granularity for RMSE metric is HAGEN. This is due to the homophily-aware architecture that HAGEN provides which makes it easier for HAGEN to capture sudden spikes.
- CrimeForecaster, HAGEN, ST-HSL show consistent performance across different temporal granularity. Despite having to predict larger crime numbers with coarser temporal granularity, these models benefit from the fact that the actual number of zero crimes is small. DeepCrime performs well at finer resolution compared to when the temporal resolution is coarser.

4.3.2 Performance Comparison for Classification Task

- As the temporal resolution becomes coarser, the prediction performance of all the models improve which suggests that for classification task it is harder for models to predict at a finer temporal resolution (Refer to Table 13). This can be attributed to the fact that crimes exhibit strong daily, weekly, monthly and seasonal correlation. With coarser temporal resolution these trends become more obvious for the models to capture.
- AIST outperforms all the other models in terms of both the Macro-F1 and Micro-F1 score at different temporal granularity. The performance gap of AIST is more prominent while predicting at finer temporal granularity. This is due to the fact that contrary to the other models AIST explicitly captures the recent, daily and weekly trends in crime data by using three different LSTM modules.
- In contrast to its performance in other experiments, HAGEN performs worse, more so at finer temporal granularity. The adaptive graph architecture utilized by HAGEN struggles to effectively learn and adapt when the available feature information becomes limited.
- Contrary to the regression task, for classification DeepCrime shows a strong performance when the temporal granularity of the prediction is coarser. Classification, being a comparatively simpler task than regression, allows DeepCrime to leverage its simple design and effectively predict the crime occurrences at coarser temporal precision.

Findings. For regression task, it is harder for the models to predict crimes at coarser temporal granularity. AIST performs best in terms of the MAE score, while HAGEN performs best in terms of the RMSE score. Contrary to regression, for classification task it is harder for the models to predict crimes at finer temporal granularity. AIST is the best performing model for both Macro-F1 and Micro-F1 metrics for classification.

Table 13: Classification metrics for different temporal granularity.

Model	Criteria	4h	6h	12h	24h
DC	Macro-F1	0.24	0.32	0.34	0.42
	Micro-F1	0.24	0.38	0.40	0.55
MiST	Macro-F1	0.29	0.35	0.38	0.32
	Micro-F1	0.30	0.38	0.41	0.37
CF	Macro-F1	0.16	0.21	0.30	0.38
	Micro-F1	0.20	0.25	0.35	0.44
HAGEN	Macro-F1	0.09	0.13	0.23	0.37
	Micro-F1	0.10	0.15	0.25	0.40
ST-HSL	Macro-F1	0.13	0.23	0.34	0.34
	Micro-F1	0.16	0.32	0.51	0.54
AIST	Macro-F1	**0.45**	**0.48**	**0.50**	**0.51**
	Micro-F1	**0.78**	**0.59**	**0.64**	**0.68**

Table 14: Impact of various spatial components on performance

Model	Spatial Correlation On			Regression Metric		Classification Metric	
	Crime	External Data	Extent	MAE	RMSE	Micro F1	Macro F1
HAGEN	Y		Y	0.287	0.226	0.707	0.663
	Y		N	0.283	0.226	0.706	0.663
	N		Y	0.486	0.280	0.526	0.472
	N		N	0.491	0.277	0.492	0.473
AIST	Y	Y		0.440	0.670	0.680	0.510
	Y	N		1.330	2.058	0.674	0.496
	N	Y		1.332	2.052	0.684	0.503
	N	N		1.332	3.011	0.674	0.496

4.4 Exploring the Impact of Area Size with Fixed Crime Density

The aim of these experiments is to explore the impact of different community areas on performance when crime density remains constant. To accomplish this, we select a specific density group and divide the communities in that group into five categories based on their area size: very small, small, medium, large, and very large. Next, we calculate the mean of the metrics for communities for each of the five area categories. We then compute the variance of the five means for that density group, which gives us the variance of performance among the different area groups, while holding density constant. Our findings indicate that the majority of the models demonstrate minimal variances, typically ranging from 1e-2 to 1e-6. This indicates that for a given density category, there is minimal variation in the performance of communities with different sizes.

4.5 Exploring the Impact of Spatial Components on Crime Prediction Performance

Previous experiments indicate that HAGEN and AIST consistently outperform other models across various scenarios. In our subsequent experimentation, we focus only on these two models, deactivating different spatial components to explore their relative importance in these models' success.

HAGEN and AIST utilize complex architectures designed to capture various spatial correlations in crime data. We can identify three types of spatial correlations captured in these models: spatial correlation on crime, spatial correlation on external data, and spatial correlation on extent.

HAGEN's Graph Learning Layer leverages POI data to construct a regional graph, capturing spatial correlations of crimes across different regions. The Homophily constraint ensures that neighboring nodes in the regional graph display similar crime patterns, thereby capturing spatial correlation on extent.

AIST incorporates two variants of the Graph Attention Network: hGAT and fGAT. The hGAT component learns region crime embeddings by integrating hierarchical information, capturing the spatial correlation on crime. Meanwhile, the fGAT component embeds external features into the model, capturing the spatial correlation on external data.

We deactivate the components responsible for capturing these three types of spatial correlation to determine their importance in the superior performance of the models. Table 14 presents the results of our study.

Findings. When its components for capturing spatial correlation are active, HAGEN performs better in all scenarios. In contrast, when these components are active, AIST performs noticeably better on the regression task and marginally better on the classification task. The findings show that techniques for capturing spatial correlations in crime data improve the models' ability to predict crimes.

4.6 Exploring the Impact of External Features on Crime Prediction Performance

To find out the effect of external data on the performance of the models we conduct ablation study on the models that use external datasets: DeepCrime, HAGEN, AIST. Table 15 shows the performance of the models with and without using external feature data. DeepCrime takes POI and 311 public service complaints data as external features along with crime data. For both regression and classification metrics, DeepCrime performs slightly better when external data is used. HAGEN uses POI data to build the region graph. Without the POI external feature, HAGEN performs better for regression and negligibly worse for classification. AIST uses POI data and taxi flow data for predicting crimes. With these external features turned off AIST perform worse for both the prediction types. However, for classification the degradation is small.

Findings. Compared to other models, the introduction of external features have more visible impact on the prediction performance of AIST due to its attempt to explicitly capture the interaction between crime and external data. In contrast, the gain in performance is marginal for models (e.g., DeepCrime, HAGEN) which treat external datasets only as an additional feature ignoring its interaction with crime data.

Table 15: Performance of models in ablation study. (R: Regression, C: Classification)

Model	Task	Metric	w External	w/o External
DC	R	MAE	**0.912**	0.916
		RMSE	**0.937**	0.939
	C	Macro-F1	**0.426**	0.420
		Micro-F1	**0.549**	0.546
HAGEN	R	MAE	**0.493**	0.498
		RMSE	**0.541**	0.547
	C	Macro-F1	**0.369**	0.367
		Micro-F1	**0.401**	0.396
AIST	R	MAE	**0.323**	0.331
		RMSE	**0.634**	0.653
	C	Macro-F1	**0.453**	0.451
		Micro-F1	**0.777**	0.775

5 Key Findings

Q1. Which model/models exhibit superior performance in crime regression task? Table 12 (See 24h column) reveals that AIST and HAGEN demonstrate the best performance among the competing models. AIST excels in terms of the MAE metric, indicating its superior performance in predicting the average number of crimes. On the other hand, HAGEN performs best in terms of the RMSE metric, showcasing its ability to capture sudden spikes in crime occurrences. These two models perform at a similar level for overall crime prediction. Their utilization of complex architectures and consideration of multiple external features contribute to their superior performance.

Q2. Which model/models exhibit superior performance in crime classification task? AIST demonstrates superior performance compared to all other models in predicting the occurrence of crimes, as evident from Table 13 (See 24h column). This aligns with its overall better performance in regression tasks. AIST's proficiency in predicting the number of crimes naturally translates to better performance in predicting their occurrence.

Q3. Which model/models exhibit superior performance when the temporal granularity of the prediction is very fine, i.e., 4h? AIST, the best performing model in terms of MAE score outperforms its closest competitor MiST by 8.1%, whereas HAGEN the best performing model for RMSE score outperforms its closest competitor MiST by 1.85%. The performance gap in predicting crimes at a finer granularity is more evident for crime classification task. AIST comprehensively outperforms the next best performing model with an improvement of 55.17% and 160% in terms of Macro-F1 and Micro-F1 metrics, respectively.

Q4. Which model/models exhibit superior performance when the crime dataset is very sparse? AIST comes across as the best performing model across both the regression and the classification task for all the evaluation metrics due to the model's design architecture which separately captures different temporal trends as well as the interaction between the external features and the crime data.

Q5. Is the crime data itself enough for crime prediction task in absence of external features? Table 15 suggests that crime data itself is enough for a good crime prediction model be it for regression or classification task. Even though incorporating external features improve the prediction performance for the respective models, the improvement is marginal. In our experiments we find that incorporating external features improve the prediction performance of DeepCrime, HAGEN and AIST $0.43\%, 1\%, 2.41\%$ in terms of MAE, $0.21\%, 1.09\%, 0.29\%$ in terms of RMSE, $1.43\%, 0.54\%, 0.44\%$ in terms of Macro-F1 and $0.54\%, 1.26\%, 0.26\%$ in terms of Micro-F1, respectively.

Q6. Does introducing external features always improve the prediction performance of the crime prediction models? Analyzing the models such as DeepCrime, HAGEN and AIST that use external features for predicting crimes, we found that introducing external features always improve the prediction performance of the models, be it marginal. Out of these three models only AIST attempts to capture the interaction between the crime data and external features by introducing a graph attention layer. The other two models DeepCrime and HAGEN incorporate anomaly and POI data with the temporal and spatial view by directly concatenating with the temporal and spatial view, respectively ignoring the interaction between the crime and external features. As a result, the improvement in performance of AIST

with external features is comparatively better than the other two, i.e., incorporating external features improve AIST's performance 2.41% compared to $0.41\%, 1\%$ for DeepCrime and HAGEN for MAE, respectively.

Q7. Is capturing the spatial correlation among different regions actually necessary for crime prediction task? Out of all the competing models, only DeepCrime ignores the spatial correlation between regions and performs the worst for regression task (Refer to Table 12), suggesting that incorporating spatial correlation of the regions is essential for the regression tasks. However, its superior performance over all the competing models (excluding AIST) for classification task suggests that considering spatial correlation is not a must for the classification task as long as the temporal correlation and external features are integrated into a model's design architecture. Refer to Table 13.

Q8. Do regression and classification task require designing model architecture with different properties? Models that leverage complex neural network architecture to capture the spatial dependencies tend to perform better for regression task specially when there is ample crime data, e.g., On the other hand, models that explicitly captures different temporal trends tend to perform better for the classification task, e.g., AIST.

6 Recommendations

Capturing spatio-temporal correlation (R1). Design crime prediction models such that it can explicitly capture both the spatial correlation among the regions and the temporal correlation of the crime data. For spatial correlation, it is essential to not only capture the dependencies of the neighboring regions, but also the similar regions with same crime profile that may be far away from the target region. As for temporal correlation, it helps to explicitly capture daily, weekly temporal dynamics of the crime data.

In absence of external features (R2). In case of the unavailability of the external features, building a model that can utilize the spatial and temporal is good enough for a crime prediction model. A combination of GNN (i.e., GCN, GAT) and RNN (i.e., LSTM, GRUs) for capturing the spatial and temporal correlation is a good starting point while designing the crime prediction models. Avoid using LSTMs for capturing the spatial correlation or convolution networks for capturing the temporal correlation.

Utilizing external features (R3). If external features are available, it is always beneficial to utilize it. However, incorporate explicit modules in your crime prediction model such that it can capture the interaction between the crime data and the external features to benefit fully from utilizing external features.

Regression-only task (R4). While designing crime prediction models only for regression purpose, the spatial module requires special attention since it is the driving force that will decide the prediction performance. The spatial module should be able to capture not only the spatial dependencies of the neighboring regions but also it should be able to learn from the non-neighboring regions that exhibit similar crime patterns to the target region.

Classification-only task (R5). While designing crime prediction models only for classification purpose, utmost importance should be given to capture the different trends of crime. Models should include separate modules to capture recent, daily, weekly, monthly crime trends. Only after the temporal dynamics is fully captured, spatial and external feature modules can be introduced that complement the temporal module.

7 Conclusion

In this paper, we have conducted a detailed experimental study of all major modern deep learning based crime prediction models and compared them in a unified environment. As a number of different deep learning based models have been proposed for crime prediction in recent years, and there has been a lack of direct comparisons among these models, researchers and practitioners face hurdles in analyzing and adapting these models. Therefore, we have systematically compared the models' performance across different scenarios, including variations in community area size, crime density, and prediction intervals. This experimental study has enabled us to identify the models that performed best in specific scenarios and gain insights into the suitability of different architectures under various conditions.

References

P. Chen, H. Yuan, and X. Shu. Forecasting crime using the arima model. In *FSKD*, pages 627–630, 2008.

Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5): 1189–1232, 2001.

Junbo Zhang, Yu Zheng, and Dekang Qi. Deep spatio-temporal residual networks for citywide crowd flows prediction. In *AAAI*, pages 1655–1661. AAAI Press, 2017.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

Chao Huang, Junbo Zhang, Yu Zheng, and Nitesh V. Chawla. Deepcrime: Attentive hierarchical recurrent networks for crime prediction. In *CIKM*, pages 1423–1432, 2018.

Chao Huang, Chuxu Zhang, Jiashu Zhao, Xian Wu, Dawei Yin, and Nitesh Chawla. Mist: A multiview and multimodal spatial-temporal learning framework for citywide abnormal event forecasting. In *WWW*, page 717–728, 2019.

Jiao Sun, Mingxuan Yue, Zongyu Lin, Xiaochen Yang, Luciano Nocera, Gabriel Kahn, and Cyrus Shahabi. Crime-forecaster: Crime prediction by exploiting the geographical neighborhoods' spatiotemporal dependencies. In *ECML/PKDD*, pages 52–67, 2020.

Chenyu Wang, Zongyu Lin, Xiaochen Yang, Jiao Sun, Mingxuan Yue, and Cyrus Shahabi. HAGEN: homophily-aware graph convolutional recurrent network for crime forecasting. In *AAAI*, pages 4193–4200, 2022.

Lianghao Xia, Chao Huang, Yong Xu, Peng Dai, Liefeng Bo, Xiyue Zhang, and Tianyi Chen. Spatial-temporal sequential hypergraph network for crime prediction with dynamic multiplex relation learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, aug 2021. doi:10.24963/ijcai.2021/225. URL https://doi.org/10.249632Fijcai.20212F225.

Zhonghang Li, Chao Huang, Lianghao Xia, Yong Xu, and Jian Pei. Spatial-temporal hypergraph self-supervised learning for crime prediction. In *ICDE*. IEEE, may 2022. doi:10.1109/icde53745.2022.00269. URL https://doi.org/10.11092Ficde53745.2022.00269.

Yeasir Rayhan and Tanzima Hashem. Aist: An interpretable attention-based deep learning model for crime prediction. *ACM Trans. Spatial Algorithms Syst.*, 9(2), 2023. ISSN 2374-0353. doi:10.1145/3582274. URL https://doi.org/10.1145/3582274.

Shaobing Wu, Changmei Wang, Haoshun Cao, and Xueming Jia. *Crime Prediction Using Data Mining and Machine Learning*, pages 360–375. 01 2020. ISBN 978-3-030-15823-1. doi:10.1007/978-3-030-14680-1_40.

Xu Zhang, Lin Liu, Luzi Xiao, and Jiakai Ji. Comparison of machine learning algorithms for predicting crime hotspots. *IEEE Access*, 8:181302–181310, 01 2020. doi:10.1109/ACCESS.2020.3028420.

Karabo Jenga, Cagatay Catal, and Gorkem Kar. Machine learning in crime prediction. *Journal of Ambient Intelligence and Humanized Computing*, 14:1–27, 02 2023. doi:10.1007/s12652-023-04530-y.

Varun Mandalapu, Lavanya Elluri, Piyush Vyas, and Nirmalya Roy. Crime prediction using machine learning and deep learning: A systematic review and future directions. *arXiv preprint*, 2023. URL https://arxiv.org/abs/2303.16310.

T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi:10.1109/TIT.1967.1053964.

Vladimir N Vapnik and Corinna Cortes. A training algorithm for optimal margin classifiers. *Machine learning*, 20(3): 273–297, 1995.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997a.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Junyoung Chung, Çaglar Gülçehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, abs/1412.3555, 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, 2:27:1–27:27, 2011.

David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied Logistic Regression*, volume 398. John Wiley & Sons, 2013.

New york city crime data, 2017. URL https://www.nyc.gov/site/nypd/stats/crime-statistics/crime-statistics-landing.page.

Sepp Hochreiter and J"urgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997b.

Chicago crime data, 2019. URL https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2.

Q. Liu, Shu Wu, Liang Wang, and Tieniu Tan. Predicting the next location: A recurrent model with spatial and temporal contexts. In *AAAI*, 2016.

Linmei Hu, Juan-Zi Li, Liqiang Nie, Xiaoli Li, and Chao Shao. What happens next? future subevent prediction using contextual hierarchical lstm. In *AAAI*, 2017.

Jie Feng, Yong Li, Chao Zhang, Funing Sun, Fanchao Meng, Ang Guo, and Depeng Jin. Deepmove: Predicting human mobility with attentional recurrent networks. In *WWW*, pages 1459–1468, 2018.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018.

Chuanpan Zheng, Xiaoliang Fan, Cheng Wang, and Jianzhong Qi. GMAN: A graph multi-attention network for traffic prediction. In *AAAI*, pages 1234–1241, 2020.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.

Bing Yu, Haoteng Yin, and Zhanxing Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In Jérôme Lang, editor, *IJCAI*, pages 3634–3640, 2018.

Chicago 311 public service complaint data, 2019. URL https://data.cityofchicago.org/Service-Requests/311-Service-Requests/v6vf-nfxy.

Chicago poi data, 2019. URL https://mygeodata.cloud/data/download/osm/points-of-interest/united-states-of-america--illinois/cook-county/chicago.

City of Chicago Portal. Chicago taxi trips data, 2019. URL https://data.cityofchicago.org/Transportation/Taxi-Trips-2019/h4cq-z3dy.