PREDICTION OF CRIME OCCURRENCES: A DATA-DRIVEN APPROACH FOR SINGLE DOMAIN AND CROSS-DOMAIN LEARNING

by

Fateha Khanam Bappee

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
December 2020

Dedicated to my beloved parents and brother,
for their unconditional love, support, and sacrifices.

Table of Contents

List of Tables

List of Figures

Abstract

Nowadays, urban data such as demographics, infrastructure, and crime records are becoming more accessible to researchers. This has led to improvements in quantitative crime research, such as identifying factors that contribute to criminal activities. However, data from smaller cities are not as available or comprehensive. Applying the same research techniques to both urban regions and smaller domains is difficult due to nonlinear connections and data dependencies. To address this challenge, we examine an extensive set of features link to different domains from various perspectives and provide explanations for each link.

Our study aims to build data-driven models for predicting future crime occurrences. We first examine the geographic aspect of crime by focusing on a single domain, the city of Halifax, Nova Scotia. We apply reverse geocoding technique to retrieve spatial information using Open Street Map, and propose a density based spatial clustering algorithm to generate crime hotspots. A spatial distance feature is then computed based on the location of different hotpoints extracted from hotspots considering different types of crime. Next, we unite the Internet of Things (IOT) and social media data, as well as explore the smart city context likely to provide a large volume of heterogeneous, city-relevant data in near future. We propose employing streetlight infrastructure and Foursquare data along with demographic characteristics for improving crime prediction. Finally, we address the same task from a cross-domain perspective to tackle the data insufficiency problem in a small city. We create a uniform outline for all geographic regions in Halifax by adapting and learning knowledge from two different domains (Toronto, Vancouver) which belong to different but related distributions with Halifax. For transferring knowledge among source and target domains, we propose applying instance-based transfer learning settings. Each setting is directed to learn knowledge based on a seasonal perspective with cross-domain data fusion. We choose ensemble learning methods for model building as it has generalization capabilities over new data. We evaluate the classification performance for both single and multi-domain representations and compare the results with baseline models. Our findings demonstrate the effectiveness of integrating diverse sources of data to gain satisfactory classification performance.

List of Abbreviations Used

AUC Area Under the Curve.

BNEs Break and Entries.

DA Dissemination Area.

DNN Deep Neural Network.

FPR False Positive Rate.

GB Gradient Boosting.

GIS Geographic Information Systems.

HDBSCAN Hierarchical Density-Based Spatial Clustering of Applications with Noise.

HRM Halifax Regional Municipality.

KDE Kernel Density Estimation.

LBSN Location-Based Social Networks.

LKDE Localized Kernel Density Estimation.

LR Logistic Regression.

M.O. Modus Operandi.

MCI Major Crime Indicator.

NLP Natural Language Processing.

OSM Open Street Map.

POI Points-Of-Interest.

RBF Radial Basis Function.

RF Random Forest.

ROC Receiver Operating Characteristics.

SLV Streetlight Vision.

SVC Support Vector Classification.

SVM Support Vector Machine.

TNR True Negative Rate.

TPR True Positive Rate.

UCR Uniform Crime Reporting Survey.

Acknowledgements

First and foremost, I would like to express my deepest and most sincere gratitude to my respected Supervisor, Professor Stan Matwin, for his excellent guidance and continuous support throughout my PhD. He is a prominent leader with such an amiable and generous personality who provided me the opportunity to explore my research ideas and interests as well as directed me to choose the right path. Thank you Dr. Matwin for always being very optimistic and supportive. This work would not have been possible without your assistance, patience, and consistent encouragement.

I would like to express my deepest appreciation to my thesis committee: Dr. Luis Torgo, Dr. Vlado Keselj and Dr. Xu (Sunny) Wang for their insightful comments, effective advice, and encouragement during my doctoral program. I would also like to thank Dr. Luiza Antonie for generously accepting to be my external thesis examiner and giving her valuable time to review my thesis.

I am also grateful to Halifax Regional Police, Halifax Regional Municipality, Nova Scotia Health Authority and Injury Free Nova Scotia for supporting me with data and ideas.

I would like to express my sincere appreciation to all of my colleagues at the Institute for Big Data Analytics. Special thanks to my colleagues Dr. Amilcar Soares, Lucas May Petry, and Xiang Jiang for helping me and giving me their valuable feedback on my work and papers. Lucas, my friend, you are the one who inspired me to come to the lab on my bad days when I was feeling demotivated and down, and cheered me up with your fun activities. Xiang, you are awesome, I am lucky to have a friend like you. Moreover, I would like to mention some names: Eman, Yamani, Habibeh, Salil, Farshid, Ahmad, Dijana, and Mohammad who became my friends during this journey and made me think of this lab as my family.

I would like to express my heartfelt gratitude to my friends: Fahim, Sazzad, Sarah and Zaahirah for listening me and supporting me through the good and bad times.

Last but not the least, I would like to thank my family for their unconditional love, encouragement, and patience during this long journey.

Chapter 1

Introduction

Crime is a well-known social problem that affects the quality of life and slows the economic growth of a country. In recent years, with the availability of a high volume of crime data, scientists have been motivated to pursue research in the field of crime and criminal investigations. For police and law enforcement agencies, it is very challenging to analyze the increasing volume of crime data without the intervention of advanced analytics. Understanding the factors related to different categories of crimes and their consequences is particularly essential.

Traditionally, criminology researchers study and analyze historical crime data by focusing on sociological and psychological theories to obtain crime and criminal behavioral patterns. However, such strategies may introduce bias from the theory-ladenness of observation [15]. Previous research found that crime in the real-world highly correlates with time, place, and population, which make the researcher's task more complicated [21]. In addition, criminal activities have been correlated to socioeconomic factors such as educational facilities, ethnicity, income level and unemployment, and human behavioral factors [51, 47, 17].

Leveraging data mining and machine learning techniques with crime research offers the analysts the possibility of better analysis and crime prediction, as well as crime pattern detection. Such research would help police and law enforcement make more efficient decisions for public safety. Crime rate or crime occurrence prediction has received considerable attention in many studies, including [119, 101, 103]. Several studies highlighted the importance of spatio-temporal patterns in crime analysis and prediction [62]. For example, Feng et al. [41] and Yu et al. [115] established a relation of committing property crime and residential burglary with geographic space and time periods respectively to their study. Mapping spatio-temporal crime hotspots may identify the reasons for relative crime proximity between time and space for specific crime occurrences. Nowadays, advanced techniques are applied to detect

different crime patterns such as spatio-temporal, demographic, meteorological, and human behavioral patterns for crime prediction. Nevertheless, it is challenging to make accurate estimations from diverse data sources due to nonlinear relationships and data dependencies. Existing data-driven crime research mostly addresses big cities with dense and diverse characteristics [15, 58, 34]. However, the demography, urbanization, and societal factors differ by region and city size. Hence, the analysis and resultant outcomes based on mega-cities might be different than smaller towns or cities. In our definition, a small city refers to a city with low or medium population density, i.e., under 1 million [5]. On this premise, we based our research on the small city of Halifax (population: 403,131 in 2016) [1], Nova Scotia, Canada, to gain an inherent understanding of its physical and human impact characteristics. We explore various nontraditional data derived from location-based social network alongside some conventional datasets. As far as we could possibly know, utilizing socioeconomic information with human behavioral factors for crime incidents prediction is the first endeavor for a small city. We extract five different categories of features: raw features (based on spatio-temporal and historical information), demographics, streetlights, Points-Of-Interest (POI), and human mobility dynamic features from diverse sources of data. We conduct our experiments with well-known ensemble learning methods, Random Forest (RF) and Gradient Boosting (GB), based on the combinations of these five features. The results demonstrate the effectiveness of integrating population-centric features, streetlight, and Foursquare POI features with raw features for crime prediction.

With the increasing accessibility of crowd-sourced and open data in big cities, there has been an interest in applying domain adaptation and transfer learning techniques across cities, and to transfer knowledge from big city to small city [121]. In recent times, the thought of knowledge transfer among different domains has been applied effectively in numerous real-world applications [79, 108]. A number of studies have explored transfer learning on a renowned machine learning field known as natural language processing (NLP) [122, 123, 99]. Liu et al. [121] introduced a domain adaptation network to identify parking hotspots with shared bikes for Beijing city by utilizing the knowledge learned from Shanghai city. Another study investigated

transfer learning for building predictive models of C. difficile infection by using information from multiple hospitals [109]. Nevertheless, only limited research has been done to explore transfer learning and domain adaptation in crime prediction. Considering the challenges of preparing a satisfactory amount of labeled training data for crime prediction build on Halifax city, we formulate our research from transfer learning and domain adaptation point of view. We propose multi-source domain adaptation techniques by adapting different domains such as Toronto and Vancouver cities with Halifax city (different but related distribution). For domain adaptation, we apply local and global min-max normalization techniques. For transferring knowledge among different domains, we consider instance-transfer to learn informative instances by seasonal subset selection. We represent different transfer learning scenarios based on the seasonal perspective with cross-domain data fusion. We tested all setups with GB classifiers and compared the results with the RF method and some well-known ensemble-based transfer learning methods. The results show the satisfactory performance with GB for crime prediction by incorporating Toronto and Vancouver domains with Halifax.

1.1 Problem and Motivation

Our study aims to build machine learning models to predict the relationships between criminal activity and geographical regions, as well as other environmental and socioeconomic factors. We implement a data-driven approach for single and cross-domain learning by integrating different data sources and fusing knowledge from them. We address the problem of predicting future crime incidents for small geographic areas (i.e., dissemination areas as defined by Statistics Canada) in Halifax. We set up the complete work from three different levels of observations through three distinct phases.

In the first phase, we observe and learn the spatial relationships of crime and criminal activities considering a single domain, Halifax. Examining spatial patterns reveals the spatial distribution and aggregation of crime. Some existing research also examined the geographic influence on future crime prediction and crime rate estimation [101]. We analyze four different categories of crime at this stage: (i) alcohol-related; (ii) assault; (iii) property crime; and (iv) motor vehicle. Law and

criminology departments have already uncovered a hypothetical relationship between alcohol consumption and violent behavior. For example, the experimental findings of Exum [37] indicate that alcohol has a causal influence on violent behavior. Another experiment has been conducted by Yu et al. [117] to check the association between a sudden loss of alcohol access and a reduction of assault violence. In particular, a study from Statistics Canada presents that the national rate of heavy alcohol consumption in Canada is 17.4% where Nova Scotian's surpasses the average by 4.9% [6]. In addition to alcohol-related crime, Halifax's crime statistics from 2012 to 2016 show that the violent crime and the property crime rates per 100,000 residents are 5,680.19 and 16,524.19 respectively [7]. These studies inspired us to continue research on crime prediction connected to alcohol-related and other violent crimes. In this phase, we primarily focus on creating spatial features to predict crime by using geocoding and crime *hotspots* identifications. Geocoding allows researchers to find various kinds of location information immediately by computing boundaries and distances. At the same time, crime *hotspots* may indicate areas where a crime type is more likely to occur. We show how geocoding can be used to create features using Open Street Map (OSM) data. Crime *hotspots* are created using a density-based clustering algorithm (HDBSCAN — Hierarchical Density-Based Spatial Clustering of Applications with Noise), then *hotpoints* are extracted from the *hotspots*. Next, we use the *hotpoints* as features for classifiers. We show using a real-world scenario that these two new features increase the performance of different classifiers for predicting the four different types of crime.

The second phase consolidates streetlight infrastructure and demographic characteristics with geographic profiling for crime analysis. Few works to date have reported the impact of streetlight distributions on criminal behavioral patterns and crime predictions. However, exploiting the hidden patterns of streetlight data as well as human demographics with crime statistics might be beneficial for crime research. Motivated by the findings of the research [111] on Detroit city, we observe the effect and significance of streetlight distribution on crime prediction for Halifax city. In addition to the streetlight and demographic patterns, we explore human behavioral patterns based on Foursquare POI and check-in behavior. Recently, location-based social networks

(LBSN) such as Foursquare are widely used in various machine learning research, particularly in urban computing, to create smart cities. Foursquare allows users to share their real-time location information and activities with others, along with the visible circumstances which can reflect the dynamic picture of the region. Inspired by author Kadar et al.'s work on understanding criminal patterns in New York City [57], we utilize the Foursquare venue and check-in information for our crime research in Halifax. We propose a data-driven approach by integrating all extracted feature categories on a single domain for future crime incidents prediction.

The third phase introduces the thought of domain adaptation and transfer learning with crime research. In general, urban profiling links to comprehensive, dynamic, and diverse patterns of each neighborhood. These patterns must then be efficiently solved computationally to gain the highest benefit. With significant advances in machine learning techniques, it is possible to promote crime research with prominent urban features. However, due to data insufficiency, privacy concerns, as well as geographically asymmetric crime data distribution, it is challenging to develop a uniform outline for all regions in a small city like Halifax. The existence of population movements and commuting facilities between cities, as well as cross-city interoperability features, motivate us to leverage transfer learning techniques with the crime prediction problem. Considering Halifax as our target city, we import knowledge from two other big cities, Toronto and Vancouver. Instead of using training and testing data with the same probability distributions, transfer learning adapts data from the different distributions. This strategy will save developers time by not relearning the model when testing a new group of data from different areas or domains. In general, the adequacy of knowledge transfer is influenced by the connection among different sources and target domains. Including multiple sources of data helps discover firmly connected sources to the target and to promote transferring positive knowledge. We propose a cross-domain learning approach based on instance knowledge transfer. We obtain source instances closely connected to the target and some labeled target instances for model training.

1.2 Contributions

The contributions of the thesis are presented in three different phases:

Phase I: In this phase, our objective is to predict the relationships between criminal activity and geographical regions based on a single domain. The contributions of this phase are:

- We propose two spatial features for crime prediction by using geocoding technique and crime hotspots. For hotspots creation, we propose to use a density based clustering algorithm, HDBSCAN and extract hotpoint for each created hotspot.
- We examine four different types of crime individually using ensemble learning methods and evaluate the prediction performance of our engineered spatial features.

Phase II: In this phase, we address the crime prediction problem using diverse feature combinations based on small geographic areas of a single domain. The contributions of this phase are:

- We propose the use of streetlight infrastructure data and Foursquare data with demographic characteristics for improving future crime prediction. Its effectiveness is demonstrated in our experimental evaluation results;
- We propose data-driven models to predict future crime occurrences in smaller cities. This implies that fewer data points are applicable for training the models;
- We experimentally show the effect of each feature group proposed in previous works and this paper on crime prediction, evaluating the classification performance of different feature combinations.

Phase III: In this phase, we focus on cross-domain learning for crime occurrence prediction. The contributions of this phase are:

- We propose to apply supervised domain adaptation and transfer learning approaches on urban crime data. To the best of our knowledge, this work is the

first to adjoin crime research with transfer learning research. We analyze multiple sources of domains to find out the related source domain with target and to increase positive knowledge transfer for target domain.

- We study instance-based transfer learning methods and propose a seasonality based subset selection method for transferring knowledge of instances.

- We present that ensemble machine learning techniques can adopt generalization for different but related domains. We evaluate that ensemble learning method such as Gradient Boosting outperforms the baselines on crime prediction in three cities.

1.3 Published Papers

The main publications supporting the content of this thesis are the following:

- Bappee F. K. Identification and classification of alcohol-related violence in Nova Scotia using machine learning paradigms. In *Advances in Articial Intelligence - 30th Canadian Conference on Articial Intelligence, Canadian AI 2017, Edmonton, AB, Canada, May 16-19, 2017, Proceedings,* pages 421-425, 2017

- Bappee F.K., Soares Júnior A., Matwin S. (2018) Predicting Crime Using Spatial Features. In: Bagheri E., Cheung J. (eds) *Advances in Artificial Intelligence. Canadian AI 2018.* Lecture Notes in Computer Science, vol 10832. Springer.

- Bappee F.K., Petry L. M., Soares A., Matwin S. (2020) Analyzing the Impact of Foursquare and Streetlight Data with Human Demographics on Future Crime Prediction. 16th Int. Conference on Data Science (ICDATA'20), Springer Nature - Book Series: Transactions on Computational Science & Computational Intelligence.

1.4 Outline

The rest of the thesis is organized as follows:

Chapter 2 provides a review of the existing works on urban crime research and performs a comprehensive study on crime factors contributing to different criminal activities and their consequences. It also summarizes existing research scenarios based on crime inference and prediction, and presents the state-of-the-art algorithms explored for these purposes.

Chapter 3 explores spatial feature engineering techniques for crime hotspots detection and prediction. It introduces the data source, its retrieval, and usage in our research, and also explains data preparation activities which include data cleaning, labeling, as well as data visualization to see the graphical representation of data. It illustrates and evaluates the experimental results obtained by the proposed classifiers trained on all raw features and the engineered spatial features. It also reports the data and experimental error, as well as some ideas for future extension.

Chapter 4 proposes a data-driven approach for Halifax city by investigating an extensive set of features from different aspects. We mainly focus on human behavioral aspects, streetlight features, and the traditional demographic features for future crime prediction. It evaluates the theory and its implementation, and compares the performance of different models built from consecutive feature groups. It also points out some potential extensions for the next phase.

Chapter 5 addresses the scope of domain adaptation and transfer learning on crime research. It focuses on instance-based transfer learning for predicting future crime incidents. It reviews some existing research on instance transfer. This Chapter explores different subset selection methods and domain adaptation methods to transfer knowledge from multiple sources. Finally, it evaluates the experimental results obtained from ensemble learning technique and compares the results with some baseline transfer learning methods.

Chapter 6 summarizes the overall findings and concluding remarks of the study. This Chapter also presents some future research ideas and suggestions for urban crime study.

Chapter 2

Literature Review

The relationship between crime and various factors has been studied in much scientific and criminology research. Nowadays, researchers can use spatial information from the real world using Geographic Information Systems (GIS). Likewise, demographic information is easily accessible from different statistical sources. The use of historical facts and temporal dynamics between neighborhoods and crimes have also been broadly noted in criminology. After analyzing the factors related to different categories of crimes and their consequences, researchers have emphasized the feasible computation solutions for urban crime. The research also reviews different sources of data connected to urban crime data including crowd-sourced and open data. The existing work on crime pattern and prediction analysis can be grouped into two different sections: crime pattern analysis (Section 2.1) and crime prediction analysis (Section 2.2) based on the patterns discovered from the crime data and factors related to the crime, as well as the proposed computational tasks for crime prediction. We review the relevant literature from both single domain and cross-domain learning perspective. In our study, we define domain as an unrelated surrounding, group, or context which has some changes in data distribution.

2.1 Crime Pattern Analysis

Identifying crime patterns and trends are of great importance for crime analysts and law enforcement agencies. Crime patterns tell us the story about environment, demography, temporality, and how criminals interact with those factors. According to the crime pattern theory of criminology, three main ideas such as nodes, paths, and edges have to be considered for crime research [40]. The node indicates human activity and movement. Path refers to the route that people use in their everyday activities. Edge, the third idea of crime pattern theory, indicates the boundaries of neighborhood people live in or has social interaction. However, there are some other

factors related to time, environment and virtual community that might contribute to instigating crime.

2.1.1 Temporal Pattern

Temporal patterns of crime are learned from sequential crime data by analyzing the structure (various intervals) of temporal resources. Crime rates can be examined for hours of the day, different days of the week, months, seasons, years and others.

Works that consider the temporal aspect of crime prediction are detailed below. Bromley and Nelson [19] reveal temporal patterns of crime to predict alcohol-related crime in Worcester city. They also provide valuable insight into the spatial characteristics of the alcohol-related crime. Ratcliffe [85] focuses on temporal dynamics of crime pattern detection. He proposes three types of temporal *hotspots*: diffused, focused, and acute. The author also explains by which way the spatial and temporal factors integrate inside the *hotspot* matrix. However, the author did not apply any machine learning strategy to predict crime. Carroll and Brower [20] analyze four categories of crime data which include liquor law violations, assaults and batteries, vandalism, and noise complaints. Different categories of crime show different temporal patterns. Their studies show that serious crimes happen at the bar closing time between 2 am and 3 am and less severe crimes happen between 11 pm and midnight. Similarly, Cusimano et al. [30] found the relationship between ambulance dispatch and bar closing time to be from 12 am to 4 am in their study. The month and year of crime occurrence are included in [17, 15] while predicting crime, though the authors mainly focus on demographic features and mobile network activity. The study conceded that by aggregating weekly, daily and hourly trace of crime events would give finer outcomes [17]. Wang et al. [97, 98] presented a periodic temporal pattern with hourly crime intensity and holiday information for crime forecasting. In another study [13], we added 'incident start time' as a temporal feature which shows significant improvement in the classification task. Many researchers have studied how to identify temporal patterns among criminal incidents [58, 119, 35, 8, 101, 81]. The drunk driving incidents and other criminal incidents occur during Saturday nights and bar game nights close to the bar, as well as sports season close to the stadium [115, 114]. In several literature [42, 11, 16], the authors arranged the crime data as

six 4-hour time intervals. According to [42], the residential break and entry crimes decreased from 1.00 to 6.00. However, at 8.00, 12.00 and 18.00 crimes increased dramatically. For commercial break and entries (BNEs), crimes increased between 3.00 to 5.00 as well as 17.00 to 18.00. Another study [67] produced a time series for each region based on the incidents happen by the day, week and month. Later, the time series signal was converted to a binary signal with time. Crime increases during hot summer months, and holidays such as Thanksgiving, Christmas have a visible effect on crime as well [95]. Temporal trends have been analyzed in several studies [33, 72, 71] for crime research.

Several works also focus on historical information along with temporal knowledge to predict future crime incidents [116]. Nevertheless, previous analysis implies that the temporal influence of crime may change over geographic regions.

2.1.2 Spatial Pattern

Environmental criminology reveals that crimes are correlated with environment contexts. The geographic area of crime analysis may vary from one place to another. Crimes are not randomly distributed throughout the space. The aim of spatial pattern analysis is to discover the spatial distribution and aggregation of crime.

Here, we list the work focuses on spatial features of crime incidents. Chainey et al. [26] identify crime *hotspots* using Kernel Density Estimation (KDE) to predict spatial crime patterns. Similarly, in another study, Nakaya and Yano [73] create crime *hotspots* with the help of KDE. However, they combine temporal features with crime *hotspots* analysis. Brower and Carroll [20] clarify crime movement through the city of Madison using GIS mapping. The authors investigate the relationships among high-density alcohol outlets and different neighborhoods. The paper [85] proposes three categories of spatial *hotspot*: dispersed, clustered, and *hotpoint*. Nath [74] employs a semi-supervised clustering technique for detecting crime patterns. Geospatial Discriminative Patterns (GDPatterns) was introduced by Wang et al. [100] to capture the spatial properties of crime. Spatial autocorrelation is considered in [116] where the average number of neighbors is calculated for each grid. Several studies analyzed spatial patterns in conjunction with some other patterns while predicting crime occurrence [58, 16, 46]. Spatial analysis aims to predict the real-time crime in several

studies [97, 98]. Fitterer et al. [42] exhibit the residential and commercial BNEs hotspots to predict property crime in Vancouver. In addition, the authors in [11, 10] found spatial patterns (hotspots) for crime prediction using the Apriori algorithm and LKDE respectively. Different hotspot techniques are introduced and compared in [33] to retrieve spatial information of crime. The study believe that spatial information along with other geo-coded events are associated with crime tendency [72]. The geographical profiling of school, subway, parks, etc. can help to get the hints about crime scene [63]. For geographic profiling, the study apply discrete distance decay function. Recently, the authors [15] infer that spatial information can dig up the better insight about crime distribution in urban area and are highly significant for crime prediction. In our study [13], we engineered spatial features by using geocoding technique and by generating hotspots.

2.1.3 Spatio-temporal Pattern

The purpose of Spatio-temporal pattern analysis is to obtain understanding from geo- and time-related crime data. As the distributions of crimes vary in time and space, identifying patterns from the dynamic interaction among time, space and crime is very challenging. In 2017, Zhao and Tang [119] explored Spatio-temporal correlation for their study. Another study [114] proposed a global spatio-temporal pattern for crime prediction with 800-meter by 800-meter grid range for a specific location. The local crime distributions are learned at the beginning for different crime periods. Then the spatio-temporal patterns are induced from each distribution using the Cluster-Confidence-Rate-Boosting (CCRBoost) algorithm. The algorithm iteratively picks some local patterns with minimum classification error. These are called an ensemble spatio-temporal pattern. At the end of the process, a global spatio-temporal pattern is learned from these ensemble patterns which is then utilized for crime forecasting. Spatio-temporal dynamics for break and entries (BNEs) crime are investigated by Fitterer et al. [42]. The study applied near-repeat modeling from Ratcliffe [86] to calculate the spatio-temporal distance between each crime. They structured crime data for six 4-hour time intervals using a 200-meter by 200-meter grid. Later, the residential and commercial BNEs within 500, 850 and 1000 m from the starting event and from 1 to 30 days of incident creation were assessed as observed pattern. The

cascading spatiotemporal pattern (CSTP) was discovered by Mohan et al. [70] for crime analysis. In another literature [67], the authors established the applicability of spatiotemporal STL prediction method. One of the contributions of the study is Dynamic Covariance Kernel Density Estimation method (DCKDE). Spatiotemporal patterns inherent into the crime incidents may affect the crime risks [35]. A study [95] examined the spatiotemporal correlation of criminal offenses based on eigenvalue spectrum analysis and Random Matrix Theory. Mohler et al. [71] modeled the spatio-temporal cluster in order to extract the crime patterns using self-exciting point process. Wang et al. [106] presented the spatio-temporal generalized additive model (STGAM) to incorporate the information for specific time and space. Newton et al. and Leong et al. [75, 62] separately reviewed theoretical analysis of different spatio-temporal patterns into crime events. Leong illustrated the spatio-temporal topological relationship (STTR) on crime analysis. The study also presented the flow pattern based on spatio-temporal sequence. Yu et al. [115] discovered the Ensemble Spatio-Temporal Pattern (ESTP) to represent the global spatiotemporal characteristics of various regularities. Feng et al. [41] presented spatio-temporal characteristics of property crime in Beijing. Spatio-temporal regularity was revealed for crime forecasting in [97, 98]. Another study [118] investigated intra-region temporal and inter-region spatial patterns for crime prediction based on cross-domain learning. Nevertheless, considering the geographic influence may add a little help for crime prediction as the neighboring community shares similar demographics.

2.1.4 Demographic Pattern

Traditional demographic features have been extensively used in many research for crime prediction [17, 58, 15]. This field of research focuses on crime pattern detection using more demographic information and criminal profiling. For example, Buczak and Gifford [21] apply a fuzzy association rule mining technique to detect community crime pattern. For mining association rules, the authors mainly consider demographic features such as population density, mean people per household, people in the urban area, people under the poverty level and people in dense housing with some other features. Demographic and socioeconomic features have widely been used

by researchers for crime prediction [17]. Another study [72] discovered the association of construction permits, foreclosures etc. with crime tendency. Researchers have applied demographic data, such as population, number of vacant houses, owner-occupied houses, number of people who are married or separated [106], population density, poverty, residential stability [101, 11, 100], type of premises [105], education, ages, income levels [15, 58, 115, 16], property values [42]. However, using only demographic feature is insufficient to understand the implicit characteristics of crime and criminals.

Recently, Fatehkia et al. [38] proposed leveraging Facebook 'interests' data from the Facebook Advertising API with demographic data for crime rate prediction. Interests are analyzed based on four different groups such as movie, game, music and relationship-related interests for specific age groups and gender. The study found that integrating Facebook interests data with demographic census data improves the models prediction power. Few works reported the impact of streetlight distributions on the criminal behavioral pattern and crime prediction. An inverse relationship between streetlight density and crime rates based on the census block groups in Detroit has been found by the researchers in 2018 [111]. In our study, we also consider extracting streetlight features, but for crime incidents prediction. However, due to human mobility, the demographic characteristics of a region may change for a short or long period of time.

2.1.5 Meteorological Pattern

Having knowledge from criminology, it has been found that meteorology and crime are correlated [84]. Motivated by this, Zhao and Tang [119] collected meteorological data which includes weather, temperature, wind strength, snowfall from NYC meteorological station while predicting crime. Environmental factors like daily weather records have statistically significant impacts on crime rate [95]. The study found that crime rate increases with the rise of temperature, and precipitation leads to decrease the crime rate. Moreover, the study discovered interesting differences by comparing the coefficients of the effect of weather for different types of crime. In another study [17], the authors include weather data from the Open Data Institute. Similarly, in 2017, researchers [97, 98] incorporated temperature, wind speed, and special events,

which include fog, rain, and thunderstorms, for weather feature. Inspired by the theory of temperature/aggression and routine activity which state that high temperature or warm weather leads to rising crime occurrence, the authors [58] collected weather data for their research. Based on the same theory, D.V.S. Pereira et al. [81] tested the relationship between homicides and weather pattern using Pearson correlation coefficient. Xinyu Chen et al. [28] discovered that temperature has a significant influence that leads a person to aggressive behaviors.

2.1.6 Human Behavioral Pattern

Human behavioral pattern aims to obtain understanding from human behavior, mobility, and networks.

Criminal behavior analysis considers the following signs: modus operandi (M.O.), Crime Scene Signature, Depersonalization, Staging, Undoing Behavior, Ritual Behavior. In the context of criminal investigations, modus operandi (M.O.) detects someone's behavior or working habit while committing the crime. Wang et al. [105] identify the M.O. of the particular offender. The study captures several important aspects of patterns. First, each M.O. is different. According to this aspect, for the housebreaks prediction problem, some offenders choose weekdays for their operation while the residents are at work; some choose night time while the residents are sleeping. They also capture the attributes of whether the offenders favor large apartment buildings or single-family houses. Another aspect of general commonalities in M.O. do exist. Sometimes, similarities in time and space are often found even though the pattern is different. Third, patterns can be dynamic. For example, the M.O. changes with experience between novice and experienced offenders. Occasionally, the offender shows a fantasy-driven, repetitive crime scene behavior while committing a criminal act. These are called ritual behavior and crime scene signature. This signature aspect is not dynamic and might be the same. For example, offenders sometimes show some unnecessary acts while killing the victim like fill the victim's mouth with dirt, pull their hairpins out and press their hands together [59].

Human mobility and network analysis are also important for crime analysis. Bogomolov et al. [17] investigated the predictive power of aggregated and anonymized

human behavioral data derived from a multimodal combination of mobile network activity and demographic information. These data are specified as the Smartsteps data considering the real source of data which is Telefonica Digital's Smartsteps product. Specifically, footfall or the estimated number of people within each cell is derived from the mobile network by aggregating every hour the total number of unique phone calls in each cell tower and mapping the cell tower coverage areas to the Smartsteps cells. The study also estimated how many people are in the cell per hour and the percentage of residents, workers, visitors among those people. Zhao and Tang [119] extracted check-in information, pick-up and drop-off points from POI data and taxi trajectories dataset respectively for these purposes. The taxi flow data of [101] reflect how people commute in the city. The authors speculated taxi flows as "hyperlinks" in the city to connect the locations. For each taxi trip, they recorded pickup/dropoff time, pickup/dropoff location, operation time, and the total amount paid. The authors hypothesized that the social interaction among two communities propagates crime. The taxi flow feature indicates how much crime in the target area is contributed by its neighboring areas through social interaction. Fitterer et al. [42] used LandScan ambient population data to represent human activity pattern over 24h for their study. Similarly, Andrey Bogomolov et al. [16] derived an estimation of footfall from mobile network activity using the unique phone calls in each cell tower. Traunmueller et al. [96] analyze footfall count from telecommunication data and find a correlation between crime and metrics derived from population diversity. A data-driven approach is presented by Belesiotis et al. [15] for crime rate prediction that also considers road network, transportation nodes, and human mobility. Recently, crime event prediction for Brisbane and New York are studied in [57, 90] using dynamic features extracted from foursquare data. The authors measure the region popularity by determining the total number of observed check-ins in that region for a specific time interval. Also, the total number of unique users that checked in to a specific venue and the number of tips users have ever written about that venue are counted to measure the popularity and heterogeneity, as well as the quality of the region.

2.1.7 Other Patterns

Images can provide valuable insights into the characteristics of a location and may relate to indicators of criminal incidents. Belesiotis et al. [15] analyzed Flickr photos from Yahoo which locate the area of greater London. The paper extracted two types of features such as photo timestamps and photo tags. In another study, the authors [58] collected image data from Google street view images to relate the crime prediction with environmental context information. Besides image feature, the authors [28] extracted polarity score for each tweet within the Chicago city boundary and showed a 3-day trend of polarity score. They used these extracted features with other crime features to construct their crime prediction model. Some other patterns, for instance, racial and ethnic diversity are explored by [101, 16, 21, 58] in predicting crime. Andrey et al. [16] considered the statistics of political control, for instance, the proportion of seats won by Labour, Liberal Democrats, and Conservatives, and election turnout. Street light density and graffiti rate per 1000 persons are used by Fitterer et al. [42] to represent the urban environment and social instability. Matthew S. Gerber [46] extracted topics from tweets tagged within the city of Chicago and combined these features with standard Kernel Density Estimation. For topic modeling, Latent Dirichlet allocation (LDA) was used. Similarly, another study [107] extracted event based topics from tweets of a news agency covered the area of Charlottesville to predict future crime. Wang et al. [102] explored features from Twitter content along with Foursquare content for next place prediction.

Table 2.1 represents the articles that analyze urban crime based on spatio-temporal, demographic and others perspective. The majority of the studies investigating crime research, predominantly deal with a single domain (city). In this thesis, we investigate an extensive set of features connected to urban crime and propose a data-driven approach based on single domain (Chapter 3 and 4) and cross-domain (Chapter 5) learning.

2.2 Crime Prediction Analysis

Nowadays, data mining and machine learning techniques offer better crime analysis and prediction for police and law enforcement agencies. The purpose of applying

Table 2.1: Crime Pattern Analysis

Pattern Type	Approach/Example	Article
Spatio-temporal Pattern	CCRBoost	[114]
	Near-repeat modeling	[42, 86]
	ST-ResNet	[97, 98]
	CSTP	[70]
	DCKDE	[67]
	ESTP	[115]
	STTR, Flow Pattern	[62]
	Spatio-temporal hotspots	[41, 85]
	STGAM	[106]
	Correlation Matrices	[95]
Demographic Pattern	Population density in different level	[21, 101, 11, 100]
	Education, age, income levels, property values	[15, 58, 115, 16, 42]
	No. of vacant houses, owner-occupied houses	[106]
	Others	[105, 72]
Meteorological Pattern	Weather	[119, 95, 17, 81, 58]
	temperature, wind strength, snowfall	[119, 28, 97, 98]
	special events, fog, rain, thunderstorm	[97, 98]
Human Behavioral Pattern	Footfall	[17, 16, 96]
	check-ins, pick-up, drop-off	[119, 101]
	Foursquare POI and dynamic	[15, 57, 90, 118]
	taxi trajectories	[119]
	road network/transportation nodes	[15]
	modus operandi (M.O.)	[105]
	human mobility/activity	[15, 42]
	ritual behavior/signature	[59]
Other Pattern	Images	[15, 58]
	Event/topic	[46, 107]
	Racial and Ethnic diversity	[101, 16, 21, 58]
	Polarity Score	[28]
	Street light	[42, 111]
	textual content	[102, 12]
	political control	[16]

Table 2.2: Quantitative research for crime prediction

Target	Data Type and Source	Approach	Article
Crime Rate Prediction	Residential Burglary, LA Police Dept.	Self exciting point process model	[71]
	NYC Crime data	TCP	[119]
	Chicago Community area data	Linear Regression, Negative Binomial Regression	[101]
	NYC crime data	STCN in Deep CNN	[35]
	31 types of crime, Chicago, Illinois crime data	DNN	[58]
Crime Hotspot Prediction	14 types of crime, Greater London, UK police	Ridge Regression, RF, SVM	[15]
	Residential burglary, Northeast, USA	1NN, SVM, J48, NN, Bayes	[116]
	Residential Burglary, Northeast city, US	Hotspot Optimization Tool (HOT)	[100]
	11 types of Real crime data, London	LR, SVM, NN, DT, Ensemble, RF (best)	[17]
	Shoplifting, burglary, assault, Camden, South Chicago, San Francisco	PSTSS technique (optimal extent)	[8]
	Shoplifting, burglary, assault, Camden, South Chicago, San Francisco	PSTSS technique	[9]
	17 types of crime, Chicago, Illinois, USA	LKDE optimized by genetic algorithm	[10]
	11 types of crime, London	SVM, RF (best)	[16]
Crime Type Prediction	Northeastern city, residential burglary, motor vehicle	Empirical Discriminative Tensor Analysis	[72]
	Northeastern city, burglary, vehicle larceny	CCRBoost	[115]
	Hit-and-run incidents, Charlottes ville, Virginia	Linear modeling (GLM)	[107]
	Crime data from police dept., twitter, Chicago,Illinois	Binary Logistic Regression with KDE & LDE	[46]
	BNE & Property crime, Vancouver, Canada	Generalized Linear Logistic Regression	[42]
	14 types of crime, Denver, Los Angeles	Naive Bayes, Decision tree	[11]
	Theft incidents, Chicago city, Police dept.	LR with standard hot-spot (KDE)	[28]
	Alcohol related crime, NS court & Newspaper data, NS, Canada	Ensemble of SVM and RF	[12]
Future-location Prediction	Serial crime cases, Gansu, China	Bayesian Theory	[63]
	Serial Crime Cases	Rossmo's formula	[82]
	British Columbia (BC) Crime data, RCMP	CrimeTracer model (Random walk based approach)	[94]
	Twitter data, Chicago, Illinois	Text-enriched Model	[102]

predictive policing is mainly stopping or reducing crime before it happens [89]. This section reviews and presents the computational tasks and modeling based on urban crime under four different subsections. Subsection 2.2.1 summarises the prediction tasks established on crime data. Subsection 2.2.2 and 2.2.3 present the computational models of crime prediction build on regular machine learning and deep learning methods respectively. Subsection 2.2.4 summarises existing crime research based on cross-domain transfer learning.

2.2.1 Prediction Tasks

Crime Rate Prediction

Crime rate prediction helps to predict the crime rate of a given region which may occur in the future. In 2011, Mohler et al. [71] proposed a self-exciting point process model to predict the average daily percentage of residential burglary crimes. The study applied a nonparametric evaluation method to understand the spatio-temporal triggering function and temporal aptitude in the background burglary rate. Zhao et al. [119] designed a novel framework, TCP for crime number prediction which captures intra-region temporal and the inter-region spatial correlation. In another study [101], the authors presented crime rate inference problem for Chicago community areas. The study refers to crime rate as the crime count normalized by the population in a region. By accommodating 311 data with other crime data researcher found increased performance for crime risk prediction task [35]. The model can predict crime number of target region in the urban area during the time window. On the other hand, a deep-learning-based multimodal data fusion can accurately predict crime occurrences by considering environmental context information [58].

Crime Hotspots Prediction

Crime Hotspots refer to the places that have high crime intensity i.e., that attract a big number of potential offenders and victims at the same time. Predicting or detecting crime hotspots is of high importance for police, law enforcement agencies, and citizens. Belesiotis et al. [15] presented a purely data-driven methodology for predicting crime hotspots. According to the authors, a cell is classified as a crime hotspot if its crime rate is above the overall median (threshold). The study identified crime hotspots using the Getis-Ord spatial autocorrelation statistic [76] and trained the hotspot prediction model as a regression analysis problem on the level of individual areas. Yu et al. [116] analyzed a variety of classification algorithms to predict crime "hotspots" by leveraging the spatial knowledge inherent in the crime data. In another literature [100], Dawei Wang et al. model the relationship between target crime Hotspots and their underlying related variables. They proposed a new model — Hotspot Optimization Tool (HOT) to emphasize the identification of crime hotspots

where the hotspot boundaries are optimized by Geospatial Discriminative Patterns (GDPatterns). GDPatterns uncover the hidden information from crime's related variables. It indicates the closed frequent patterns where a user-defined threshold is less than the growth ratio. A study [17] tried to predict whether a specific area would be a crime hotspot or not. The research mainly provided valuable insights into the feature engineering while predicting 'high crime' or 'low crime' class. Prospective space-time scan statistic (PSTSS) has been employed to the grid-based predictive hotspots [9]. They developed a toolkit of evaluation metrics for various aspects of spatio-temporal point processes (STPP) based hotspot prediction. Later, Adepeju and Cheng [8] maximize the accuracy of crime hotspots prediction by determining the optimal value of spatial scan extent of PSTSS. Localized Kernel Density Estimation (LKDE) optimized by a genetic algorithm performed significant improvement to predict crime hotspots [10]. The research applied the concept of convolution filtering for this purposes where each grid cell of the study region is considered as an image pixel. The LKDE method can adapt itself by enlarging or reducing the kernel size based on the sparse or dense region. Genetic algorithm helps to learn kernel size and convolution values dynamically. Bogomolov et al. [16] used Random Forest (RF) ensemble classifier and Support Vector Machine (SVM) method for their study to accurately classify crime hotspots.

Crime Types Prediction

Crime Type Prediction aims to predict the types of crime (e.g., assault, burglary, alcohol-related crime) in a specific region and time based on the pattern of each crime. The Empirical Discriminative Tensor Analysis (EDTA) was proposed by Mu et al. [72] to predict residential burglaries. The study designed a fourth-order tensor data structure to obtain discriminative information regarding spatial and temporal aspects, and other relevant events of each residential burglary. Chung-Hsien Yu et al. [115] also predicted residential burglary using hierarchical spatio-temporal pattern. The study conducted by Wang et al. [107] presented how to predict future hit-and-run crimes based on Twitter based posts of criminal incidents. Yu et al. [114] built a crime prediction system to predict residential burglary in a northeastern city of the US. The study chose residential burglary as target crime due to the near repeat

hypothesis concept where a burgled residence neighborhood increases the likelihood of victimization of other residence. Matthew S. Gerber studied 25 types of crime and investigated the prediction performance by incorporating the Twitter extracted topics [46]. Break and entries (BNEs) and property crime prediction were conducted by Fitterer et al. [42]. BNEs are one the most patterned and foreseeable crime types for residential and commercial locations. In another study [11], the authors analyzed 14 different types of crime including aggravated assault, sexual assault, burglary, etc for finding crime patterns and predicting future crime. Future theft incidents were predicted by Chen et al. [28] based on Twitter sentiment and weather data. In 2017 [12], we model the relationships between alcohol consumption and violence based on information spread among different text media. Later, we tried to predict the relationships between criminal activities and geographical regions [13]. Our study focuses on four different categories of crime such as assault, property crime, motor vehicle crime and alcohol-related crime.

Future-Location Prediction

Future-Location Prediction predicts the upcoming crime location, an offender is going to commit a crime as reported by their historical trajectories and other related sign. Liao et al. [63] presented how to locate the neighborhood of next-crime scene. In order to predict the future crime location, the authors [82] applied Rossmo's formula with a traffic-network based geographic profiling. Tayebi et al. [94] designed a personalized random walk model for next crime location prediction. Researchers have investigated the approaches of incorporating textual content for next-location prediction [102]. The study built the model based on Twitter posts of user's spatial trajectories. Moreover, they also examined the correlation between future-location concentrations and the actual future-crime occurrences.

Table 2.2 shows the quantitative research for crime analysis and prediction reached from this subsection.

2.2.2 Regular Machine Learning Methods

Crime research with machine learning techniques exhibits improved crime analysis and prediction. Following the underlying foundations of crime research utilizing machine

learning is not simple. We mainly focus on the research explored over the last 10 years throughout our analysis. Section 2.2.1 covers most of the studies of predictive crime mapping where the researchers employed regular machine learning techniques.

Several studies have conducted their experiments for predicting particular crime based on Logistic Regression, Generalized Linear Model (GLM), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT) techniques [28, 107, 42, 11]. For instance, Wang et al. [107] used GLM model for future hit-and-run prediction. Cluster-Confidence-Rate-Boosting (CCRBoost) algorithm was proposed by Yu et al. [114] for predicting residential burglary. The study compared the prediction results with SVM, C4.5, Naive Bayes classier, and LADTree [52]. Gerber et al. [46] applied logistic regression model for crime type prediction by implicating GPS-tagged tweets. In another study [42], the authors designed two different models for break and entries (BNEs) and property crime prediction. In the first model, generalized linear logistic regression method was used which integrates human and urban environmental features with observed crime data. However, in the second model, the regression method was applied only on observed crime data. Decision Tree and Naive Bayesian classifiers were applied in [11] to predict potential crime types in a specific location within a particular time. To predict theft crime incidents, a logistic regression model was derived by Chen et al. [28].

On the other hand, future-location prediction is analyzed by Liao et al. [63] based on Bayesian learning theory. Researchers have investigated the Text-Enriched Model to classify user's nearest next place type using linear support vector machines [102]. In another study, the authors [101] focused on Linear Regression and Negative Binomial Regression model to build the crime inference model.

2.2.3 Deep Learning Methods

Deep learning gives a state-of-the-art performance on many predictive analytics for automatic feature identification [60]. Recently, it has been applied for crime modeling and prediction, and to study spatio-temporal data. Considering the challenges of real-time spatio-temporal crime prediction, Wang et al. [97] adapted deep learning architectures for their study. The authors tested both convolutional and non-convolutional structures to predict crime distribution. Crime dynamics can be captured through

convolutional architectures. For non-convolutional model, an ensemble of ResNet is applied to learn the time series on each grid. The model does not consider the transition of crimes between different grids. Though the prediction results in both space and time are accurate, the computational cost increases dramatically due to the super resolution regularization in space. To reduce the model size and speed up the prediction process, the authors explored the Ternarization of ST-ResNet in [98]. Kang et al. [58] proposed a feature-level data fusion method with environmental context based on a Deep neural network (DNN). The DNN structure configured with four different layers: spatial, temporal, environmental context and joint feature representation layers, and softmax classifier. According to the study, the DNN model can accurately predict crime occurrences than other models. Lian Duan et al. [35] proposed a Spatiotemporal Crime Network (STCN) based on deep Convolutional Neural Network (CNN) for automatically crime-referenced feature extraction. The authors claimed that the model proposed for the end-to-end crime prediction is flexible and minimizes feature engineering bias. The study developed the inception block and fractal block to extract complicated features from various spatiotemporal patterns.

However, it is not always possible to have enough labeled training data for deep learning models, particularly for smaller problem space. Consequently, the model tries to memorize the training data and fit the model closely based on the limited samples instead of generalizing the model for the unseen future instances; and hence, introduces overfitting problem.

2.2.4 Transfer Learning Methods

Many machine learning algorithms work well for single domain learning where the training and testing data are drawn from the same distribution and feature space. However, collecting adequate training data for many real-world applications is very difficult and expensive. In such cases, where there is a limited amount of training data, domain adaptation and knowledge transfer or transfer learning would be beneficial. In transfer learning, the model stores knowledge obtained from one problem or domain and applies that knowledge to a different but related problem or domain. It allows the domains and distributions used in the training and testing data to be different. The relationship between transfer learning and other machine learning techniques are

discussed in several studies [79, 108].

In our previous research [12], while solving the classification problem of Nova Scotia's alcohol-related crime, we tried to apply some knowledge obtained from three other provinces: Alberta, British Columbia, and Saskatchewan. The work compared the performance of transfer learning by transferring the knowledge of instances with the Recursive Partitioning (RP), Support Vector Machine (SVM), and Random Forest (RF) methods. An ensemble with the predictions of RF and SVM methods was created to improve the performance in transfer learning. Experimental results demonstrated that an ensemble of RF & SVM method gains more knowledge from data and achieves substantial classification performance on Nova Scotia data.

In another research, X. Zhao et al. [118] proposed a novel transfer learning framework to integrate crime-related features extracted from cross-domain datasets and model spatio-temporal patterns for crime prediction. The study focuses on intra-region temporal patterns and inter-region spatial patterns to understand how crime originates over time for a region in a city and the geographical influence among regions in the city respectively.

Many existing research have explored transfer learning focusing on real-world applications, though a very limited research has been dedicated to crime. For example, Liu et al. [121] detected parking hotspots by transferring knowledge of dockless shared bikes among different cities. Another study [109] applied transfer learning approaches to medical data. Transfer learning also helps to evaluate the similarities among spatial networks [53]. Many research have been done on NLP by utilizing domain adaptation and transfer learning [122, 123, 99]. Recently, Raghu et al. [83] investigated transfer learning on medical imaging. An in-depth study on transfer learning methods is presented in Phase 3 (Chapter 5).

2.3 Existing Research Gaps

This section points out some insights of existing crime research that lead us to explore various scopes of future research.

In order to improve crime research, various kinds of crime pattern and prediction analysis has been conducted to date. However, the vast majority of the current

research directed to large urban communities that exhibit dense and diverse characteristics. Whereas, urban planning paradigms, and societal variables contrast by locale and size of the community. Thus, the representing study and outcomes maybe not quite the same among large and small urban communities or cities. On this account, a data-driven crime research centering small cities is important.

Though a couple of studies [118, 12] mentioned domain adaptation and transfer learning for crime prediction, they did not actualize their ideas on cross-domain crime datasets. Aiming to learn a uniform model for all cities given different data distributions domain adaptation should get particular attention in crime research. Similarly, exploring transfer learning, when sufficient data is inaccessible and willing to gain knowledge from a different domain, is an interesting avenue of future crime research.

Out of the existing studies summarized in literature review, only two [42, 111] presented the impact of streetlight distributions on urban crime. However, this can be further analyzed with respect to crime incidents prediction. The effect and graveness of the decreased number of streetlights could be observed. Moreover, the efficient number of streetlights that should exist in a neighborhood might be determined to keep the neighborhood safe.

Regarding specific type of crime research, alcohol or drug related crime analysis and prediction are very limited in comparison to other crimes. Thus more attention should be made to the closeness of alcohol establishments and crime.

Moreover, handling non-linear relationships and data dependencies among different domains should be considered for accurate crime forecasting. Advanced techniques are desired to automatically incorporate nonlinear patterns from multiple sources. Researchers [58] tried to address this problem using a deep neural network (DNN) based feature-level data fusion. However, the authors were unable to reduce computational cost with this approach.

As the majority of the existing crime research deal with geographic, demographic and socioeconomic data, measuring data bias is important to avoid biased crime prediction. Addressing discriminatory decision patterns from historical data has been largely overlooked in most of the studies summarized.

2.4 Conclusions

In this chapter, we provide a comprehensive analysis on existing urban crime research. We identify some research gaps from the perspective of crime pattern detection and prediction. We direct our research in three different phases by following the identified limitations on current research. Chapter 3 defines the first phase of our research. In this phase, we study urban crime and criminal patterns based on a single domain, Halifax. It incorporates data collection, data preparation processes, as well as data visualization to understand the patterns of data. This chapter mainly extracts temporal and spatial knowledge for crime hotspots detection and crime type prediction.

Chapter 3

Phase I: Spatial Knowledge Extraction from Single Domain

In this chapter, we present the work from the first phase of our thesis. Our goal is to predict the relationships between geographical space and crime occurrences by identifying the hidden characteristics of different types of crime. We adopt crime data from a single domain, Halifax for this purpose. We propose two methods for spatial feature engineering by transforming the information from geographic location and detecting crime hotspots.

3.1 Introduction

In general, it is known that crime is not arbitrarily or consistently distributed inside an area. This uneven crime distribution is highly influenced by the uneven spatio-temporal distributions of objects in that place. Therefore, it is essential to grasp the knowledge from spatial and temporal attributes of an urban area to understand criminal activities. As indicated by the routine activity theory, three factors, such as a probable offender, a favorable target, and a favorable circumstance (or non-attendant circumstance) must converge in time and space to direct a crime incident [40]. Moreover, each crime has its distinctive influential factors and consequences. Distinguishing the implicit characteristics of various types of crime and their predictive learning are the main concerns in our research. In particular, we investigate four types of crime for this phase: assault, property damage, motor vehicle and alcohol-related.

We divide the whole approach into five different sections. Section 3.2 introduces the data sources and data collection process. It also illustrates the data analysis and visualization techniques. Section 3.3 defines the strategy of engineering spatial features. Section 3.4 deals with the evaluated classifiers to model the relationships between criminal activity and geographical regions. Moreover, Sections 3.5 and 3.6 present the detailed information regarding performance assessment criteria employed in this phase and the evaluated results obtained from those criteria respectively.

3.2 Data Analysis and Visualization

Our analysis involves crime data from Halifax, Nova Scotia (NS), Canada. Section 3.2.1 gives a detailed description of the data that we use in our study. Section 3.2.2 shows the graphical representation of data to better understand the hidden patterns of data.

3.2.1 Data Source, Collection and Labeling

Crime records from Halifax regional police (HRP) department are used in this work, and it covers most of the districts in Halifax Regional Municipality in Nova Scotia. Our dataset was extracted from the Uniform Crime Reporting Survey (UCR). The UCR was designed to measure the incidence of crime and its characteristics in Canadian society.

In total, 756,913 crimes were reported by UCR from 2006 to 2016 across the regional municipality of Halifax. After deducting data with zero, null and invalid geographic coordinates, we have total 257,017 data. For this phase, we only consider crime incidents that have the alcohol flag recorded [1]. We explore all of the offenses of 2016 which include 3726 data samples for our experiments. The crime attributes extracted from the source data include geographic location, incident_start_time, month, weekday, ucr_descriptions or incident type, and whether the incident happened because of alcohol.

We group our data using four different classes, named alcohol-related, assault, property damage, and motor vehicle using the ucr_descriptions and alcohol incident fields. Inspired by observations stated in Section 1.1, we select these four categories of crime. It has been observed that alcohol is involved in various criminal activities including assault and motor-vehicle crimes [37, 12]. Besides, these are some of the most patterned and well reported crime types [42, 4]. For the alcohol-related crimes, we considered all the cases where alcohol presence was reported in the UCR using the alcohol incident field. For all the remaining classes, the ucr_description field was used. The assault group covers all levels of assault including sexual assault,

[1] As the crime incidents with alcohol flag represents a very small proportion of the total actual crime incidents for this time period, missing data could substantially alter the current patterns generated from our model.

aggravated assault, bodily harm, threat, murder, etc. Property damage group covers break, theft, robbery, fraud, etc. Motor vehicle group covers all types of a motor vehicle accident, act violation and impair driving.

The total data distributions with different categories of crimes and their ratio are summarized in Table 3.1. For the alcohol-related group, 1742 crimes were reported as alcohol incident and 1984 crimes that have no relation to alcohol. The Assault group contains 1291 crimes related to assault and 2435 crimes with no relation to assault. Next, in the Property Damage group, 431 crimes were reported as damage crimes and rest 3295 crimes were considered as not property damage crimes. Finally, 686 crimes belong to motor vehicle group, and 3040 were grouped in the no motor vehicle-related group.

Later, to create the *hotspots* and *hotpoints* (geographic center of the hotspot), we used UCR form data from the year of 2015. We created *hotspots* for each positive class in Table 3.2 and the shortest distance from each 2016 incident to a *hotpoint* was used in the experiment. For example, when the positive class was alcohol-related, we used a single shortest distance to a *hotpoint* that was extracted from the examples of alcohol-related crime from 2015.

Table 3.1: Dataset description for 2016

Crime type	**Negative**	**Positive**	**Total**
Alcohol-related	1984(53%)	1742(47%)	3726(100%)
Assault	2435(65%)	1291(35%)	3726(100%)
Property damage	3295(88%)	431(12%)	3726(100%)
Motor vehicle	3040(81%)	686(19%)	3726(100%)

Table 3.2: Dataset description for 2015

Crime type	**Negative**	**Positive**	**Total**
Alcohol-related	1952(53%)	1931(47%)	3883(100%)
Assault	2465(65%)	1418(35%)	3883(100%)
Property damage	3473(88%)	410(12%)	3883(100%)
Motor vehicle	3187(81%)	696(19%)	3883(100%)

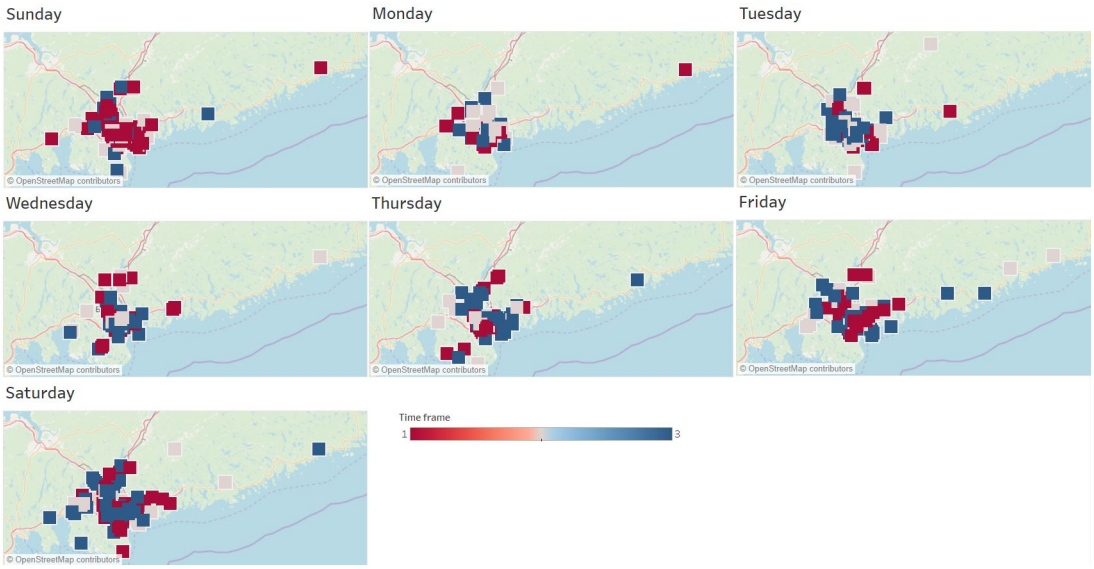


Figure 3.1: Crime density for alcohol-related crime based on three 8, 10 and 6-hour time frames of the weekdays

3.2.2 Data Visualization

Visualization helps decision makers to see the graphical representation of data and to capture new patterns from data. Before crime prediction, this chapter presents the visual picture of crime data 2016 to better understand the structure of data and to see if there is any hidden patterns in this data. The visualization shows the crime spatial distribution with respect to 7 days of the week in 2016 for alcohol-related crime [2] These graphs discover some dissimilar characteristics among the weekdays crime distribution. In order to better understand the spatial and temporal patterns, crime data are formatted with three time frames in a grid density map (Figure 3.1). Time frame 1, 2 and 3 indicate 8 hours [0-8), 10 hours [8-18), and 6 hours [18-24) duration per day with red, brown and blue colors respectively. From the figure, it is clear that crime intensity changes with time and space.

Figure 3.2 represents the time series plot for all four categories of crime from January 2016 to December 2016. It reveals features on different scales. For alcohol-related crime, most of the incidents happen on Saturday in September and October

[2]As this visualization may potentially create privacy issues, according to our agreement with HRM Police we have not included it in the final version of the thesis.

(Figure 3.2(a)). The number of incidents remain constant from January to June. A peak for assault related crime is noticeable on Friday in May (Figure 3.2(b)). On the other hand, for property related crime, Thursday in March and Friday in May are observable. Similarly, the majority of motor vehicle crimes occur in June on Friday.

Figure 3.3 compares the temporal patterns of crime based on days of a week for all four categories of crime. From Figure 3.3(a), we have following observations: time frame 1 [0-8) shows highest crime rate for Saturday and Sunday; time frame 2 [8-18) shares almost similar temporal distribution for each day. On the other hand, time frame 3 [18-24) has a bit high crime rate than time frame 2 which gradually increases on Fridays and Saturdays. Therefore, Saturday mornings (or Friday midnight) and Sunday mornings (or Saturday midnight) are the most unsafe time for alcohol-related crime. Assault crime shows different scenarios where most of the crimes happen in time frame 2 (Figure 3.3(b)). Moreover, it is difficult to blame more on any specific day for time frame 2. However, for time frame 3, maximum incidents occur on Fridays. For property damage crime, time frame 1 is the safest time frame among all and time frame 2 on Fridays is the most unsafe time period (Figure 3.3(c)). Figure 3.3(d) shows the highest crime rate for time frame 2.

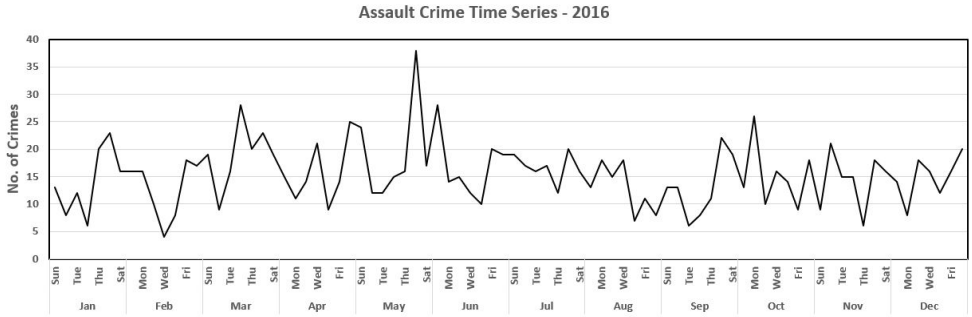
3.3 Engineering Spatial Features

This section discusses the details of spatial features created for crime prediction. Section 3.3.1 describes how the geocoding process is used to extract geographic information to create spatial features and Section 3.3.2 outlines the techniques for crime *hotspots* detection.

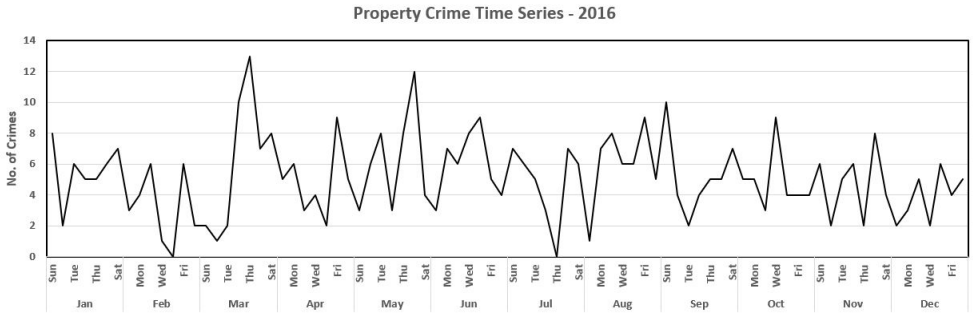
3.3.1 Geocoding

Geocoding is the process of spatial representation of a location by transforming descriptive information such as coordinates, postal address, and place name. When the geographic coordinates are converted to get a location description, it is defined as reverse geocoding. The geocoding process relies on GIS and record linkage of address points, street network and boundaries of administrative unit or region. For this work, we use a reverse geocoding technique to extract the spatial information from the

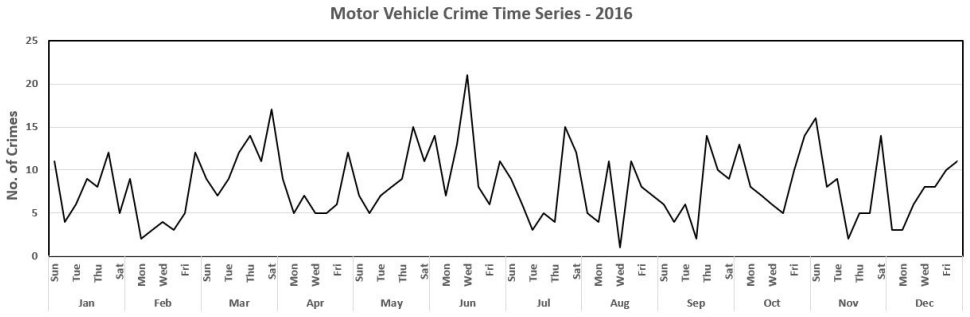
(a) Alcohol related crimes.



(b) Assault related crimes.

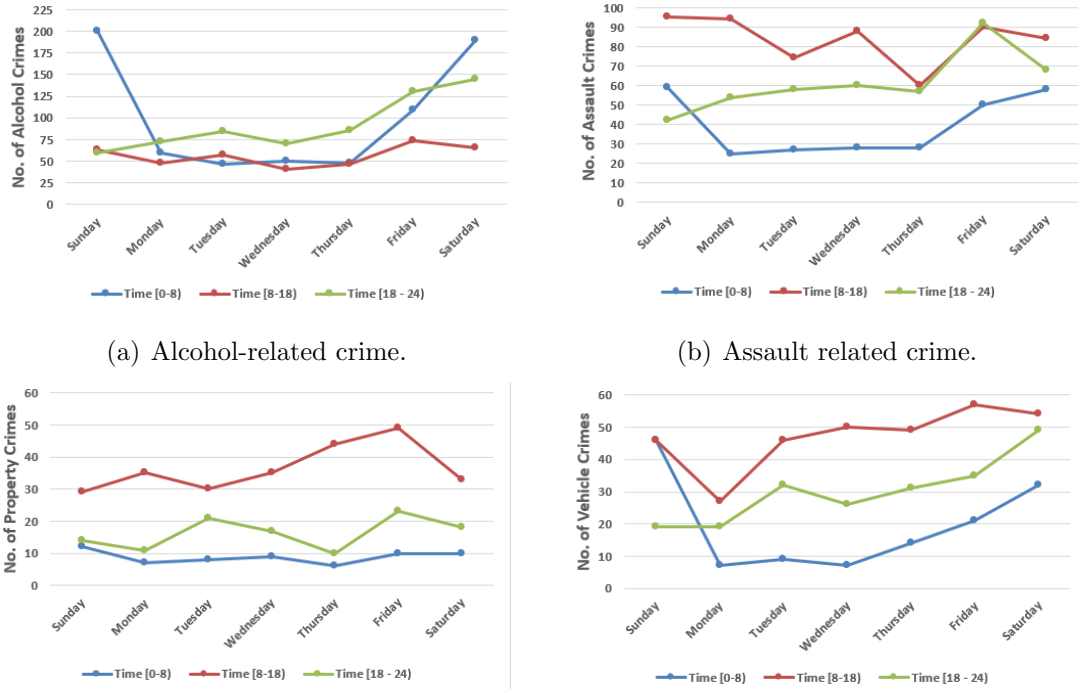


(c) Property related crimes.



(d) Motor vehicle related crimes.

Figure 3.2: A time series plot on days of a week in 2016. (a) alcohol crime time series, (b) assault crime time series, (c) property crime time series, (d) motor vehicle crime time series.



(a) Alcohol-related crime.

(b) Assault related crime.

(c) Property damage related crime.

(d) Motor vehicle related crime

Figure 3.3: Comparison of the number of crime incidents by days of a week and three 8-hours time frames. (a) alcohol-related crime, (b) assault related crime, (c) property damage related crime, and (d) motor vehicle related crime.

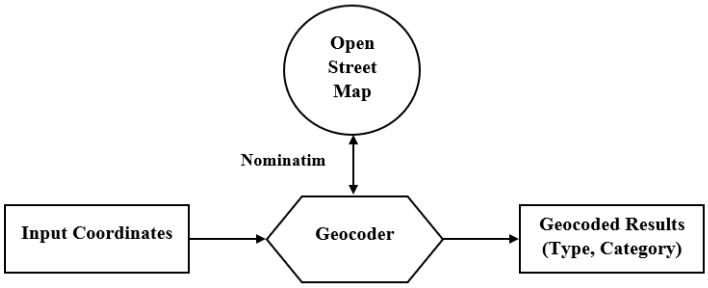


Figure 3.4: Geocoding Framework

crime data. The geocoder library [45], written in Python, was used for geocoding services with the Open Street Map (OSM) provider. Figure 3.4 presents the framework of geocoding process. Every crime in our dataset contains geographical coordinates (latitude and longitude) that are given to the Nominatim tool. Then, the tool queries the OSM dataset and outputs some information from geographic points.

The output of the Geocoder package can be 108 types of location including bar, pubs, bus stops, or hospitals from Nova Scotia. According to OSM documentation, all of these location types are grouped into 12 categories including amenity, shop, tourism, office, etc. We use both types of location and category as features to predict crime. Figure 3.5 (a) and (b) show, respectively, alcohol-related crimes where the output of the geocoding process returned as categories highway and amenity, and as type bus stop and pub [3].

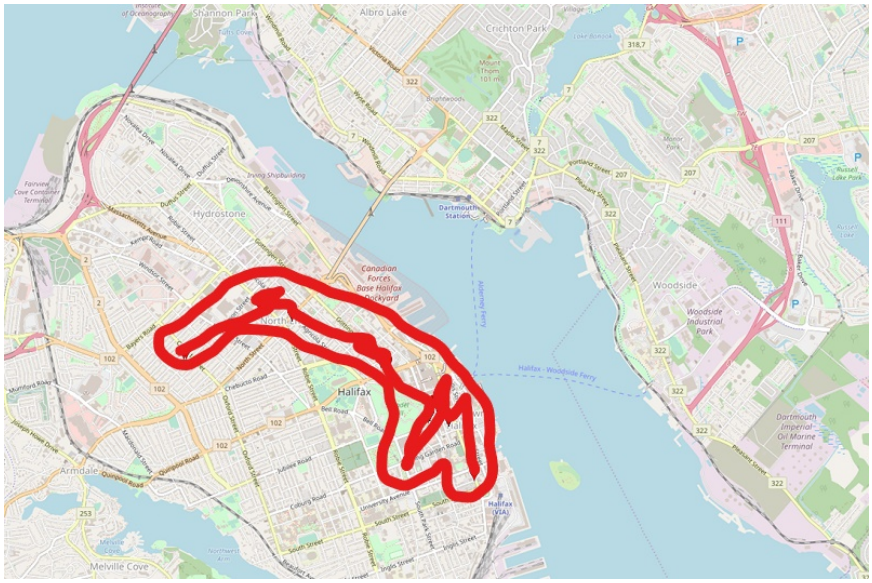
3.3.2 Clustering for Hotspot Creation

Hotspot analysis can emphasize the patterns of data regarding time and location of a geographic area. In the study of crime, one of the critical issues involves the analysis of where crimes occur in general. Therefore, for a crime analyst, the creation of *hotspots* became very popular to identify high concentrated crime area.

In this work, *hotspots* are created and transformed into a feature to predict different crime types. The idea is to cluster crime data into regions with a high rate of occurrence of the same crime type. Because we want to group crime data by density,

[3]We made the crime points unclear due to the privacy issues.

(a) Crimes around Bus stops.



(b) Crimes around Pubs.

Figure 3.5: Alcohol-related crimes (red pins) on and around bus stops (a) and pubs (b) in Halifax city.

a reasonable choice is to select the DB-Scan [36] algorithm. However, since this algorithm has a complexity of O(n^2), we decided to use the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [23]. The HDBSCAN algorithm we used works as follows (Step1 to Step5):

Step 1: Calculate core distance for all data points regarding m_{pts} (m_{pts} = minimum cluster size).

Step 2: Build the minimum spanning tree (MST) of the mutual reachability graph ($G_{m_{pts}}$).

Step 3: Construct a cluster hierarchy of connected components by extending MST.

Step 4: Shorten the cluster hierarchy by removing edges in decreasing order of weights based on minimum cluster size.

Step 5: Extract the stable clusters or the HDBSCAN hierarchy as a dendrogram from the tree found in Step 4.

We selected the HDBSCAN for several reasons: (i) can handle data with variable density; (ii) eliminates the ϵ (*eps*) parameter of DBSCAN which determines the distance threshold to cluster data; and (iii) has a complexity of $\mathrm{O}(n \log n)$.

The clusters can identify *hotspot* areas where different types of crime occur. Instead of considering the whole dense area as a feature, we extract one geographical point named *hotpoint* for every *hotspot* found by the HDSCAN. This *hotpoint* is determined by extracting the *hotspot* center, averaging the geographical positions inside the area. Finally, we extract the location of new crime reports and compute the distances to all *hotpoints* found in the data and select the shortest distance to a *hotpoint*. This shortest distance to a *hotpoint* is then used as a feature for crime prediction.

In this work, we used the Haversine distance in both HDBSCAN and shortest distance to *hotpoint*. The Haversine formula (given in Equation 3.1) determines the shortest distance between two points on Earth located by their latitudes and longitudes. In this equation, r is the average Earth radius (6371 km), l_1, l_2 define latitudes of two points and λ_1, λ_2 define longitudes of two points.

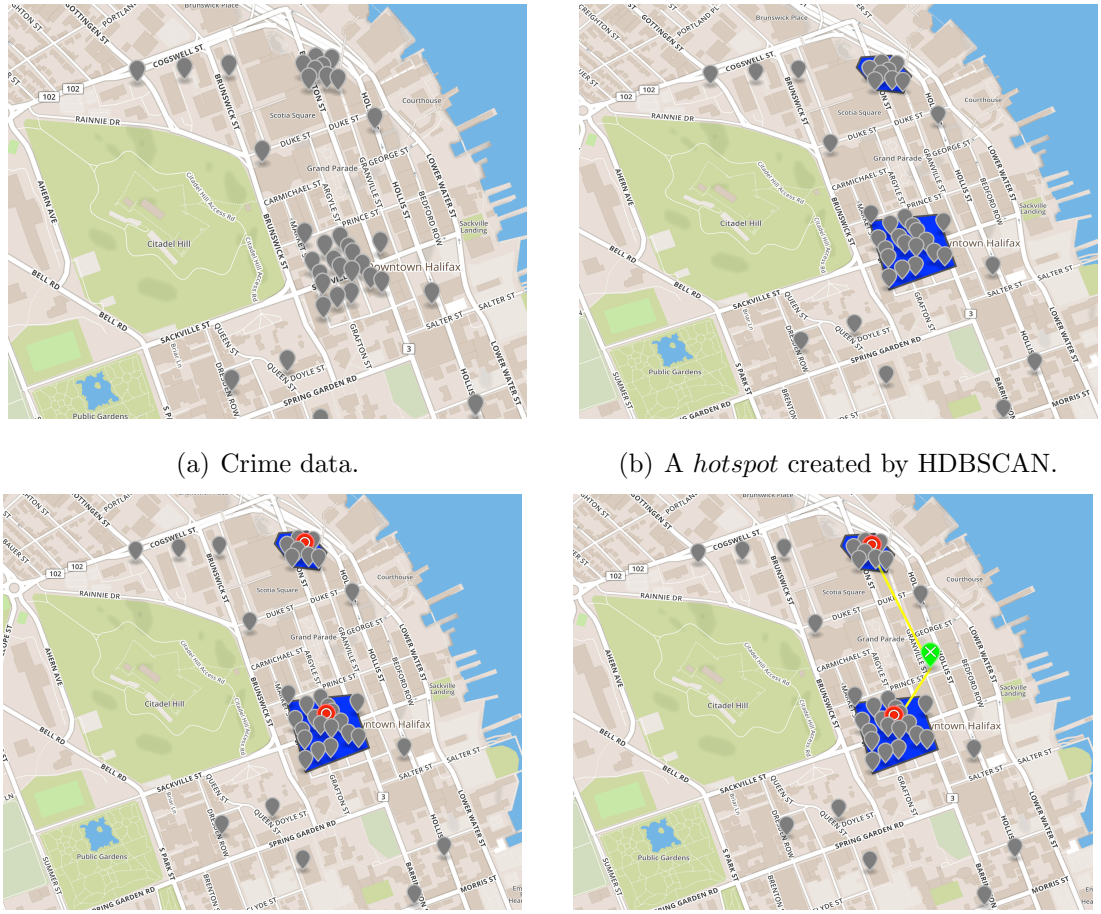
$$distance = 2 * r * arcsin(\sqrt{d}) \tag{3.1}$$

$$d = \sin^2(\frac{l_2 - l_1}{2}) + \cos(l_1) * \cos(l_2) * \sin^2(\frac{\lambda_2 - \lambda_1}{2}) \tag{3.2}$$

Figure 3.6 summarizes the overall process to produce the shortest distance for *hotpoint* feature [4]. Figure 3.6 (a) shows crime examples (gray pins) in downtown

[4]Due to privacy concerns dummy examples are used to draw these pictures instead of true crime examples.

Halifax area. Then, a *hotspot* (blue area) found by HDBSCAN is shown in Figure 3.6 (b). Figure 3.6 (c) shows a *hotpoint* (red pin) extracted from a *hotspot*. Finally, a new crime example (green pin) is evaluated, and the distances to *hotpoints* (yellow line) are calculated. We use the shortest distance to a *hotpoint* as a feature to classify a crime type for crime prediction problem.



(a) Crime data.

(b) A *hotspot* created by HDBSCAN.

(c) A centroid computed from the *hotspot*.

(d) Distance from centroid for new crime data around the *hotspot*.

Figure 3.6: An overview of the crime *hotspots*, *hotpoints* and distance to *hotpoint* feature.

Table 3.3 provides the details of the total selected features for raw and engineered feature categories.

Table 3.3: Details of the observed features

Feature category	Total features	Selected feature names
Raw Feature	6	Month, weekday, incident start time, term, time frame, crime type
Engineered Feature	3	Location type, location category, shortest distance to hotpoint

3.4 Evaluated Classifiers

In this study, we mainly focus on Logistic Regression (LR) [68], Support Vector Machine (SVM) [29], and Random Forest (RF) [18] methods to model the relationships among different crime features and various types of crime. SVM and RF are optimized by choosing different kernel functions and the optimal number of decision trees respectively. For LR, 'liblinear', 'newton-cg', 'sag', and 'lbfgs' solvers can be used in the optimization problem. LR, RF and SVM has gained considerable attention in crime prediction, because of their optimal classification performance.

3.4.1 Logistic Regression (LR)

Logistic Regression models the probabilities of the different possible outcomes of a response variable, given a set of input variables. The model is appropriate for two-class classification or prediction problem. Using logistic regression, the two-class classification problem can be modeled as,

$$\begin{aligned} L &= \log(\frac{pr(\mathbf{x})}{1-pr(\mathbf{x})}) \\ &= \beta_0 + \beta_1 x_1 + \cdots + \beta_m x_m \end{aligned} \tag{3.3}$$

where L and $pr(x)$ represent the log-odds and the probability of belonging to positive class respectively. β_0, β_1,$\cdots \beta_m$ are the parameters of the LR model.

For the experiment we use 'liblinear' solver that supports both L1 and L2 regularization. We set the random_state parameter and maximum number of iterations to 1415 and 100, respectively.

3.4.2 Support Vector Machine (SVM)

Support Vector Machines (SVMs) [29] can effectively work on both linear and non-linear classifications. SVM develops a maximum margin hyperplane or a decision boundary to divide the crime points into two classes for linear data. For non-linear

data, it uses kernel functions to transform non-linear data to linear data [91]. For our experiment we use C-support vector classification (SVC) with radial basis function (RBF) kernel. The implementation for SVC is based on 'libsvm'. We choose default parameter settings for gamma and maximum number of iteration. We set the value of '1415' as random state parameter similar to LR model.

3.4.3 Random Forest (RF)

Random Forest (RF) [18] is a powerful ensemble learning classifier. It trains many decision trees on the basis of the entire dataset. We use 200 trees in our experiment. A final output of a particular target variable is made based on the results of these 200 trees. In this structure, each individual tree predicts a decision for each crime record in the test data. The class with the majority of votes is used by the model for each record among all resultant class or decisions. A random subset of training data and features are considered for every split. Therefore, Random Forest can easily manage hundreds to thousands of potential input features [18]. Random Forests correct for decision trees' overfitting habit to their training set. For the experiment, we use the same random state parameter as LR and SVM.

3.4.4 An ensemble of LR, RF and SVM

Ensemble learning combines multiple learning algorithms to obtain better predictive performance. Model diversity is the key for creating a powerful ensemble. Combining different models with diverse focus promotes the prediction performance compared to the models with identical focus [32, 69]. An ensemble with the predictions of LR, SVM and RF methods are created to improve the performance of crime prediction. The ensemble learning formulations of our problem are given below.

$$\text{predictions} \leftarrow (\text{lr\_predictions} + \text{rf\_predictions} + \text{svm\_predictions})/3$$

$$\text{predictions} \leftarrow (\text{lr\_predictions} + \text{rf\_predictions}*2 + \text{svm\_predictions}*2)/5$$

In this work, we used Scikit-Learn library [80] versions of LR, SVM, and RF to build models from crime data. We used voting classifier from Scikit-Learn ensemble module to combine three different classifiers. The method collects the predicted

probabilities from each classifier and multiplies this by the assigned weight for the specified classifier. Later, the weighted average probability is calculated to predict the final class. We chose soft voting and weights: $LR = 1, RF = 2, SVM = 2$ for our experiment. The baseline used in this work to verify if the newly engineered features help a classifier to improve the crime prediction power was the raw data contained in the UCR form.

Section 3.5 reports two popular performance assessment criteria, such as accuracy and area under the curve (AUC), which are employed in the study for evaluating a classifier's performance. Section 3.6 shows the comparison of the results for four different categories of crime based on our proposed engineered features.

3.5 Performance Assessment Criteria

We consider a two-class (positive and negative) classification problem in our study. For performance assessment, we consider using accuracy, and AUC score to evaluate the prediction performance. We decided to use both metrics because the accuracy and the AUC complement each other. Both of the metrics depend on an estimate of the true probability of being positive for new crime data. The new data or test data is referred as positive for specified crime if the estimated probability is greater than 0.5.

Table 3.4 presents a general classification table or confusion matrix for binary classification problem. N_{pp} refers to the number of positive crime data which our model classified as positive (true positive); N_{nn} indicates the number of negative crime data which are classified as negative (true negative); N_{pn} is the number of positive crime which are classified as negative (false negative) and N_{np} refers to the number of negative crime which are classified as positive (false positive).

Accuracy gives the proportion of correct results the model gets among the total number of crime data. It evaluates the degree of closeness between a measured quantity value and a true quantity value of a measurand [56]. This metric performs better when the dataset is balanced.

$$\text{Accuracy} = \frac{N_{pp} + N_{nn}}{N_{pp} + N_{pn} + N_{np} + N_{nn}}$$

Table 3.4: Classification table

	Predicted class	
Actual class	*positive crime*	*negative crime*
positive crime	N_{pp}	N_{pn}
negative crime	N_{np}	N_{nn}

The AUC is the area under the receiver operating characteristic (ROC) curve, where the ROC is plotted by true positive rate (TPR or sensitivity) against false positive rate (FPR or 1 - specificity). AUC is computed from prediction scores. In AUC, target scores can be probability estimates of the positive class. It is also possible to compute the AUC scores by using an average of a number of trapezoidal approximations [49].

$$\text{TPR} = \frac{N_{pp}}{N_{pp} + N_{pn}}$$

$$\text{FPR} = \frac{N_{np}}{N_{nn} + N_{np}}$$

After obtaining the method's evaluation results, paired t-tests are applied to test the statistical difference significance of raw and engineered features. The significance level of the paired t-test is 0.05. As the resultant p-value from the experiment is very small ($\sim 6.17e-09$ to .005) in the majority of cases, using paired t-test instead of multiple comparison tests does not affect the statistical significance analysis.

3.6 Comparison of the Methods

This section gives a detailed description of the experimental results to understand the impact of the features proposed in this work. A 10-fold cross-validation is used in all phases to estimate model prediction performance correctly. We want our model to be trained in a way that the model has experience on all possible values of a feature, so as to generalize the different seasons or special events. As an example, if the special event 'Christmas Eve' is not present in the training data, the resulting model will not be able to make meaningful prediction.

Table 3.5 shows the classification accuracy for LR, SVM, RF and an ensemble of these methods for all four categories of crime. For each method, the first column

displays the accuracy of raw features and the second column for raw features with engineered spatial features. The asterisk (*) in Table 3.5 symbol indicates that the method fails for the statistical hypothesis testing, i.e., the p-value is higher than 0.05.

For the Alcohol-related group, the results show that new spatial features achieve better accuracy in comparison with raw features for all four methods with statistical evidence support, and the Ensemble method performs better than others (75.52% of accuracy) with almost 17% accuracy improvement.

The accuracy values of the engineered features for the Assault and Property damage groups show that all methods, except LR, benefit from their inclusion. For example, adding engineered features with raw features improves nearly 11% (Assault group) and 5% (Property damage group) of accuracy for RF method. Finally, for the Motor vehicle group, all the classifiers showed improvement, except for the Ensemble classifier. Figure 3.7 shows the graphical representation of accuracy for all four categories of crime where raw and eng. indicate accuracy applying raw only features and raw with engineered features respectively.

The reason to get improved accuracy with engineered features is that, our raw features only explore temporal pattern of crime data. Besides, the distribution of crime does not always follow the same temporal trend for all geographic regions. For instance, most of the alcohol-related crime may occur at time frame 1 (Section 3.2.2) in a place which is close to the bar or pub, not in any residential area. Therefore, incorporating local geographic information of crime with temporal pattern helps to improve classification accuracy.

On the other side, for a couple of observations, we notice that the methods fail to show significant accuracy improvement with engineered features. Our explanation behind this review, especially for Property damage and Motor vehicle crimes, is class imbalance problem.

Table 3.6 shows the AUC scores for LR, SVM, RF and an ensemble of LR, SVM & RF methods. For Alcohol-related and Motor vehicle crimes, the results discovered that spatial features give better AUC scores than raw features for all four methods. For instance, the Ensemble method gives 82.5% and 69.4% AUC score for Alcohol-related and Motor vehicle crimes respectively based on engineered features. Similarly, for Assault and Property damage crime, LR, RF and Ensemble methods perform

Table 3.5: Results for accuracy. The scores with the asterisk (*) symbol specify statistically insignificant results

	LR		RF		SVM		Ensemble	
Crime type	**raw**	**raw+eng.**	**raw**	**raw+eng.**	**raw**	**raw+eng.**	**raw**	**raw+eng.**
Alcohol-related	59.36	65.27	57.73	73.51	59.28	71.31	58.61	75.52
Assault	65.35	65.03*	47.94	58.89	63.53	65.27	55.96	64.41
Property damage	88.43	88.41*	84.03	88.57	88.19	88.43	88.43	88.44*
Motor vehicle	81.59	82.31	71.82	81.45	81.11	81.45	81.56	81.80*

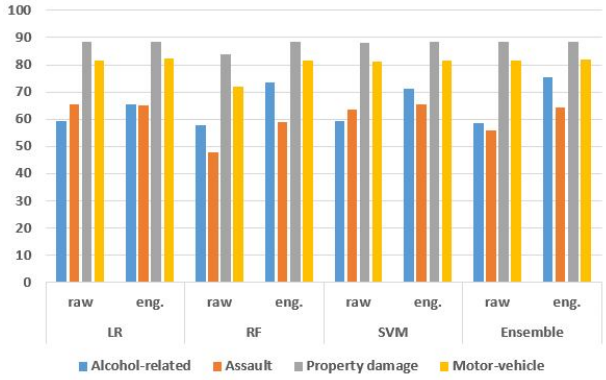


Figure 3.7: Accuracy for four different categories of crime (based on Table 3.5)

significantly better with engineered features. Adding engineered features with raw features gives 56.7% and 65.7% AUC score for Assault and Property damage crime respectively with the Ensemble method. Therefore, using spatial features, the Ensemble method performs at least 10% improvement in AUC score for all four categories of crime. However, for SVM method, there is no significant evidence of improvement. The interpretation of non-significant findings for a few observations might be the reasons of different variances between two populations, as well as non-normal distribution of variables.

Additionally, as we see from Tables 3.5 and 3.6, the results of accuracy are higher than the AUC scores for almost all cases except alcohol-related crime. This is because the class distributions for assault, property, and motor-vehicle crimes are slightly imbalanced. AUC metric tries to balance the class sizes and avoid overfitting to a single class. On the other hand, accuracy metric offers high scores by classifying most of the records in the majority class.

Figure 3.8 shows the graphical representation of AUC scores for all four categories of crime where raw and eng. indicate AUC scores using raw only features and raw with engineered features respectively. We visualize the observed and predicted alcohol-related crime distribution in Figure 3.9. Red grid represents positive alcohol-related crime and blue grid for negative alcohol-related crime.

Table 3.6: Results for AUC. The * symbol indicates statistically insignificant findings

	LR		**RF**		**SVM**		**Ensemble**	
Crime type	**raw**	**raw + eng.**	**raw**	**raw + eng.**	**raw**	**raw + eng.**	**raw**	**raw + eng.**
Alcohol-related	.575	.723	.649	.818	.635	.747	.661	.825
Assault	.528	.613	.457	.545	.504	.533*	.459	.567
Property damage	.519	.651	.531	.646	.501	.505*	.534	.657
Motor vehicle	.515	.686	.488	.682	.494	.536	.490	.694

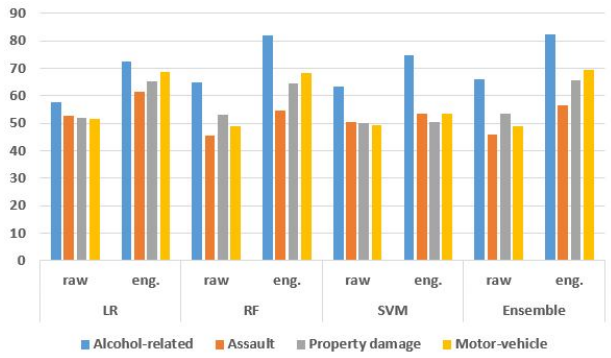


Figure 3.8: AUC scores for four different categories of crime (based on Table 3.6)

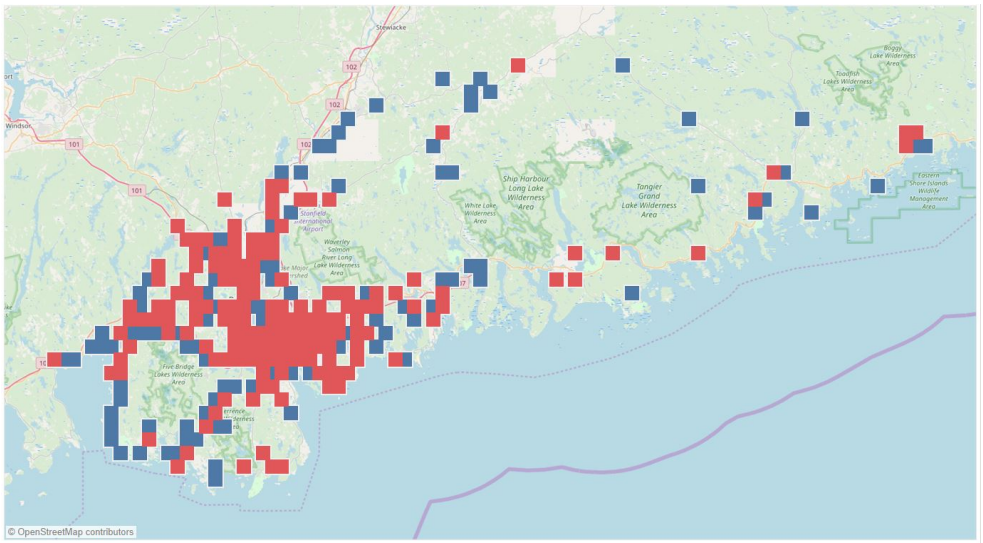
3.7 Conclusions

In this phase, we explored the creation of spatial features derived from geolocated data. We created two types of spatial features. The first used a geocoding service that can query OSM data and return a category and a type of information regarding where the crime occurred. The second used the HDSCAN algorithm to create *hotspots* grouped by type of crime, extracted a *hotpoint* from each *hotspot*, and finally returned the shortest distance for a *hotpoint* as a feature to feed a classifier.

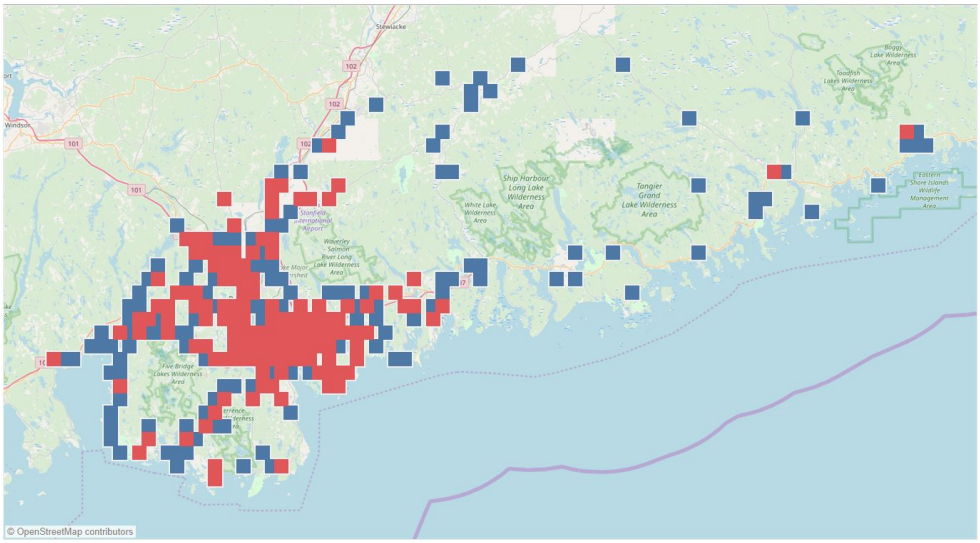
The new features were evaluated using four different crime types (alcohol-related, assault, property damage and motor vehicle) using only the information provided

in the UCR forms as features for a classifier as the baseline. The results show that significant improvements in accuracy and AUC were found when the newly engineered features were added to the tested classifiers.

However, the analysis performed in this phase was based on crime incidents only. Considering data only from crime events may raise the question of the presence of bias. Therefore, adding data with 'no crime' points (or regions) might be beneficial to get the real picture of crime incidents. Moreover, incorporating socioeconomic and human behavioral factors with crime data can help to discover the implicit characteristics of crime and criminal activities. In addition, according to the research gaps revealed in Chapter 2, examining the influence of streetlight on crime occurrence prediction is likewise important. In our next phase (Chapter 4), we attempt to address some of the issues. In Chapter 4, we propose a data-driven approach for future crime incidents prediction by integrating features from different perspectives.



(a) Observed crimes.



(b) Predicted crimes.

Figure 3.9: Alcohol-related crime distribution. (a) observed alcohol-related crime distribution, (b) predicted alcohol-related crime distribution. Red: positive alcohol-related crime; Blue: negative alcohol-related crime.

Chapter 4

Phase II: Data-Driven Approach on Single Domain

4.1 Introduction

Most of the studies that presented data-driven approaches for crime pattern detection and prediction have focused on mega-cities like Chicago, New York, Greater London, etc. However, the physical characteristics, human impact characteristics, and their interactions are totally different for different regions and cities. Therefore, applying those models for predicting crime in a smaller city is very challenging and may lead to different outcomes. Our study aims to build data-driven models for future crime incidents prediction for smaller cities. The main hypothesis is that the relative paucity of data, compared to mega-cities, can be compensated by the use of non-traditional datasets that can be derived from social media and from the Internet-of-Things (IoT) infrastructure of a modern city. We extract five different categories of features from six different data sources. We propose to explore traditional demographics data with commuting features (e.g., commuting mode and time), IoT-like streetlight poles position data, as well as human mobility data with dynamic features from location-based social networks. To the best of our knowledge, employing demographics data with human mobility features for future crime prediction is the first attempt for a small city such as Halifax, Canada [14]. For model building, we use ensemble learning methods such as Random Forest and Gradient boosting. We conduct a performance comparison for all five categories of features. We also compare the prediction results generated from ensemble learning methods with a baseline method called DNN-based feature level data fusion [58].

The rest of the chapter is organized as follows. Section 4.2 provides the details of feature engineering approaches to improve the prediction accuracy in Halifax. Next, in Section 4.3, the data source and data preparation activities are presented. Section 4.4 delivers the details of the experimental setup including the prediction models as well as the baseline model. Section 4.5 presents the experimental results derived

from our proposed features and compares the performance for all models. Finally, Section 4.6 draws some concluding remarks along with the research directions of our next phase.

4.2 Feature Extraction

Aiming at predicting future crime incidents, we extract features for each dissemination area (DA). According to Statistics Canada, a DA is the smallest standard geographic area in their data, which consists of one or more adjacent dissemination blocks [3]. Our extracted features are brought into four different sections. Section 4.2.1 details the temporal and historical features used in this work. The demographics and streetlight features are explained in Section 4.2.2, while Section 4.2.3 shows the POI features used in this work. Finally, Section 4.2.4 shows some human mobility dynamic features extracted from social networks.

4.2.1 Temporal and Historical Features

According to criminology research, crime may change over a long period of time (e.g., season) as well as in a short period of time (e.g., day or week) [85]. Thus, the temporal features we extracted are month, day of the week, time interval in a day, and season. We arrange crime records in 8 three-hour time intervals and 4 seasons (winter, fall, summer, and spring) for each DA. On the other hand, some research analyzed the relation of future crime incidents with the past crime history [116]. Therefore, we calculate crime frequency and crime density for each region based on historical crime data. As the area and population sizes are different for different regions, we normalize the crime frequency using the area and population size to obtain the crime densities (D_{crp} and D_{cra}).

$$D_{crp}(r) = \frac{CR(r)}{P(r)}, \tag{4.1}$$

$$D_{cra}(r) = \frac{CR(r)}{A(r)}, \tag{4.2}$$

where $CR(r)$ addresses the total number of crimes in DA r, $P(r)$ is the total number of population in region r, and $A(r)$ is the area of that region. We also compute the crime distribution based on each season.

4.2.2 Demographic and Streetlight Features

Demographic and socioeconomic features have been widely used by researchers for crime rate estimation [101] and crime occurrence prediction [17]. The main demographic features we consider for our study are population density, dwelling characteristics, income, mobility, the journey to work, aboriginals and visible minorities, age, and sex. The journey to work features measure two main things: (i) the time people leave for work and (ii) the primary mode of commute for residents aged more than 15 years. We consider 6 different measures for the time people leave for work, such as between 5 a.m. and 5:59 a.m, 6 a.m. and 6:59 a.m., 7 a.m. and 7:59 a.m., 8 a.m. and 8:59 a.m., 9 a.m. and 11:59 a.m., and 12 p.m. and 4:59 a.m. For the mode of commute, public transit, walk, bicycle, and other methods are considered. Mobility indicates the geographic movement of a population over a period of time, for instance, it shows the information if a person moved to the current place of residence or is living at the same place as 1 year or 5 years ago. Mobility features include 2 different types of status: (i) non-movers, and (ii) movers. Non-movers refer to the persons who are living in the same place since 1 or 5 years ago. On the other side, movers refer to the persons who did not live in the same residence 1 or 5 years ago. Movers include non-migrants and migrants condition. Movers who moved within the same census subdivision are referred as non-migrants. Migrants include persons who moved from a different city or different country.

Besides demographic features, we observe the effect and graveness of streetlight distribution on future crime incidents prediction motivated by [111].

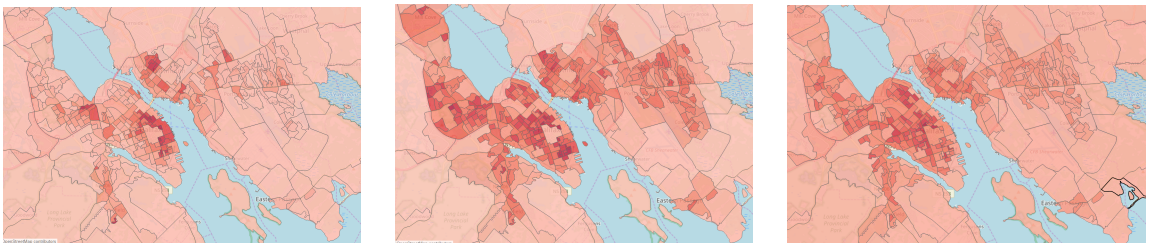
Given a dataset of streetlight locations, for each DA we propose the use of 3 streetlight features: (i) the total number of streetlights, (ii) the streetlight density, and (iii) the average minimum distance between crime data points and streetlight poles. The streetlight density of region r is computed as follows:

$$D_{st}(r) = \frac{St(r)}{A(r)}, \tag{4.3}$$

where $St(r)$ denotes the total number of streetlights in DA r. To calculate the average minimum distance from crime location to streetlight poles, we use the Haversine distance metric with scikit-learn [80]. The Haversine distance formula uses geographical latitudes and longitudes of two points on the earth to determine the shortest distance

between those points.

Figure 4.1 visualizes the crime (year 2013), population, and streetlight densities by most observable DAs in Halifax. Dark red color indicates high density, and light red indicates low density. The bin sizes for population and streetlight densities are same; on the other hand, for crime density, we choose smaller bin size to get a clear picture. As shown in the picture, most of the criminal incidents happen in downtown Halifax.



(a) Crime density. (b) Population density. (c) Streetlight density.

Figure 4.1: Crime, population density and streetlight density by most observable DAs in Halifax.

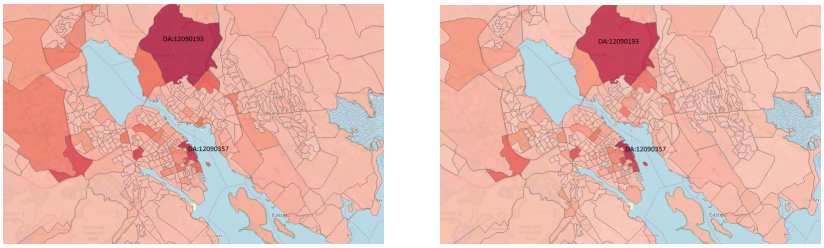
4.2.3 POI Features

In this work, we propose the use of Point-Of-Interest (POI) features that can be obtained from location-based social networks (e.g. Foursquare). POI indicates a specific venue information including geographic location which people find useful and may have a unique value due to its dynamism such as a pub, a restaurant, and a train station. Our extracted POI features include (i) the total number of POIs, (ii) the POI frequency, and the density for different POI categories. Foursquare identifies 10 major POI categories such as food, arts & entertainment, college & university, nightlife spot, outdoors & recreation, professional & other places, residence, shop & service, event, and travel & transport. The density of each POI category is defined as follows:

$$D_{cp}(r) = \frac{N_c(r)}{N(r)}, \tag{4.4}$$

$$D_{ca}(r) = \frac{N_c(r)}{A(r)}, \tag{4.5}$$

where, $N_c(r)$ is the total number of POIs of category c in a DA r, $N(r)$ is the total number of POIs in region r, and $A(r)$ is the area of that region. Figure 4.2 (a) shows the POI distribution of most observable dissemination areas (DAs) in Halifax. As depicted in figure, the distribution of POIs are uneven, for instance, downtown area (DA 12090357) includes most of the POIs except Burnside area (DA 12090193). Burnside area is a commercial and industrial area, which encompasses large land area as compared to downtown area.



(a) Poi count distribution. (b) Check-in count distribution.

Figure 4.2: The total POI and check-in count distributions by most observable DAs in Halifax. DA 12090193 and 12090357 contain more than 800 and 675 venues respectively. The majority of Halifax's dissemination areas have less than 50 venues. On the other hand, DA 12090193 and 12090357 contain more than 675 and 1300 check-ins respectively.

4.2.4 Human Mobility Dynamic Features

Our study also explores dynamic human mobility data from location-based social networks, Foursquare in order to find if there is any relation with crime context. Social networks often have location data of their users, including their visits to different POIs in a city (datasets used in this context are discussed in Section 4.3). We extract 10 features for each DA based on the total number of user check-ins, and check-in frequency for each POI category. Moreover, the check-in count for each DA at a time interval, the check-in density, region popularity, and visitor count are also computed. For DA r at time interval t, the check-in density is defined as follows:

$$D_{ckc}(r,t) = \frac{Ck(r,t)}{Ck(r)}, \tag{4.6}$$

$$D_{cka}(r,t) = \frac{Ck(r,t)}{A(r)}, \tag{4.7}$$

where, $Ck(r,t)$ is the number of check-ins in DA r at time interval t, and $Ck(r)$ is the total number of check-ins in that region. Visitor count refers to the number of unique users that visited a DA at time interval t (i.e., region popularity).

$$R_{rp}(r,t) = \frac{Ck(r,t)}{Ck(t)}, \tag{4.8}$$

where, $Ck(t)$ is the total number of check-in at time interval t for all regions. Figure 4.2 (b) depicts the check-in count distributions for some dissemination areas in Halifax. Here, check-in count computes the total number of check-ins in a specific DA at a specific time.

We extract total 153 features for each dissemination area. We tried univariate feature selection approaches with scikit-learn library to obtain the best features. We mainly focus on SelectKBest method where the score function computes the ANOVA F-value between input feature and class label. However, the improvement was not satisfactory after applying the chosen feature selection methods. Therefore, for the experiment, we consider removing the redundant features from each feature group. Later, we applied 'trial and error' method by adding a different feature combination to obtain optimal feature set. Finally, we select 65 features that are more relevant for crime prediction problem. The details of the total selected features for each category appear in Table 4.1.

Table 4.1: Details of the selected features

Feature category	**Extracted features**	**Selected features**	**Selected feature names**
Temporal and historical	12	8	Month, weekday, time interval, season, crime frequency, crime density based on population, crime density based on area, crime density for season
Demographic	101	32	Population, population density, dwelling characteristics (11) mobility movers, mobility non movers, mobility migrants, mobility non migrants, aboriginals and visible minorities, primary mode of commute for residents (5), journey to work: the time people leave for work (5), low income (3), age and sex
Streetlight	3	2	streetlight frequency, streetlight density
Foursquare POI	21	19	Total POI, food count, residence count, nightlife count, arts & entertainment count, college & University count, outdoors & recreation count, professional & other places count, shop & service count, travel & transport count, and the densities of all POI categories (9)
Foursquare dynamic	16	4	Total check-in for each time interval, check-in density, visitor count, region popularity

4.3 Datasets

We use crime data provided by the Halifax Regional Police (HRP) department, which includes records for all Dissemination Areas (DAs) in the Halifax Regional Municipality (HRM) in Nova Scotia, Canada. As mentioned in Chapter 3, the dataset was extracted from the UCR survey, which measures the incidence of crime and its characteristics in Canadian society. In this phase, we explore all crime occurrences from 2012 to 2014 for the experiments. The crime attributes extracted from the dataset include the geographic location, incident start time, month, weekday, and UCR descriptions (incident type). We have a total of 201,086 crime observations (excluding invalid and null information), where 69,340 data points happened in 2012, 65,785 in 2013, and 65,961 in 2014. We map all crime records to one of the 599 DAs collected for Halifax from statistics Canada 2016 census, based on their geographic location. We group and index crime occurrences based on the DA where they happened, the year, month, day of the week, and the time interval of the day (we partition a day into 8 three-hour time intervals).

In addition to the raw crime data, we collected demographic data for each DA from the Canadian census analyser [2]. We also extracted POI and dynamic features for Halifax from a dataset of Foursquare check-ins around the world, collected between April 2012 and January 2014 [112]. The total collected POI venue for Halifax is 13,195. We have a total of 12,171 dynamic check-in data which indicate the user check-ins at different locations. Lastly, streetlight information was obtained from the Streetlight Vision (SLV) API of HRM, which contains the location of 42,653 streetlight poles after removing null values and invalid data. We then computed the streetlight features proposed in Section 4.2 and mapped them to each DA.

Given that there are only records of crime occurrences in the dataset, we augment it to include 'no crime' records. Thus, if there was no crime for a specific time interval, we labeled that observation as 'no crime'. The final size of the dataset, including crime and no crime records, is 1,207,584 (3*12*7*8*599).

As the occurrence of crime event is not frequent, most of the data (around 87%) are labeled with 'no crime'. To address this issue, we apply the under-sampling technique for 'no crime' records in order to obtain a more balanced dataset [61]. We use the random under-sampling technique which randomly selects a subset of

DA	Year	Month	Weekday	Time_ interval	Population_ density	Food_ density	Checkin_ density	------	Class
12090103	2012	1	1	0	0.11	.02	.01	--	1
				1	0.11	.02	0	--	0
				2	0.11	.02	0	--	0
				3	0.11	.02	.02	--	1
				4	0.11	.02	.01	--	0
				5	0.11	.02	.008	--	0
				6	0.11	.02	0	--	0
				7	0.11	.02	.03	--	1

Figure 4.3: Data sample after integration and mapping based on dissemination area (DA) - 12090103 for different time slots.

observations from the major class (no crime) of the dataset. Applying random under-sampling might lead to a biased dataset; also the deleted data points could have an useful or adverse impact to fit the model. However, this under-sampling approach is compatible for our study as we are employing it for artificially creating 'no crime' records only and the number of records for 'crime' occurrences are sufficient in spite of the fact of having class imbalance. Table 4.2 shows the details of the dataset. Tables 4.3 and 4.4 list the total number of venues for each POI category and the total number of check-ins in different time intervals respectively. The number of check-ins is very low at time intervals 0 and 1 i.e, between 12 am and 6 am. The highest amount of check-ins occur between 9 am and 9 pm. We removed the category 'event' from the experiment due to a small number of records as well as missing venue information. Figure 4.3 depicts an example of sample dataset (eight rows) after data indexing, mapping and integration [1]. After data indexing, similar to crime data, we map each demographic profile, streetlight pole, POI venue and check-in location to the corresponding dissemination area using their geographic coordinates. Later, we join all groups of data with crime data by following our formatted index. We use GeoPandas library (https://geopandas.org/) and QGIS (Quantum GIS) tool (https://qgis.org/en/site/) to conduct these spatial operations.

[1]Here, we represent some dummy data instead of introducing the original data due to the privacy issue.

Table 4.2: Details of the datasets

Dataset	**Source**	**Total data**
Historical crime data	Halifax Regional Police	201,086
Dissemination area data	Statistics Canada	599
Demographic data	Canadian Census Analyser	599
Streetlight data	Halifax Regional Municipality	42,653
Foursquare POI data	Foursquare	13,195
Foursquare checkin data	Foursquare	12,171

4.4 Experimental setup

We run experiments with well-known ensemble learning classifiers, Random Forest (RF) [18] and Gradient Boosting (GB) [43], with scikit-learn [80] in Python. We used randomized grid-search in preliminary experiments for the hyper-parameter optimization of each classifier evaluated.

Besides evaluating the effect of each group of features, we compare our results to a DNN-based feature level data fusion baseline method [58]. Since the environmental context feature group used in the literature [58] is unavailable for Halifax, we implement the DNN without those features. We use the same parameter settings reported in the corresponding paper for the baseline model, except for the activations of the DNN, which were replaced by sigmoid functions as they resulted in a better performance. We train the DNN for 300 epochs and select the best test scores.

For evaluating the effectiveness of each feature group, we analyze the AUC and Geometric mean (Gmean) scores of the classifiers. The AUC score mainly calculates the area under the ROC curve. ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at all possible threshold settings [39]. On the other hand, for two-class classification problem, Gmean computes the square root of the product of the sensitivity (TPR) and specificity (TNR) [44]. This metric increases the accuracy of both classes as large as possible while maintaining the balanced accuracies. The computation of Gmean is defined as follows:

$$Gmean = \sqrt{TPR \times TNR} \tag{4.9}$$

Table 4.3: Total POIs for each category

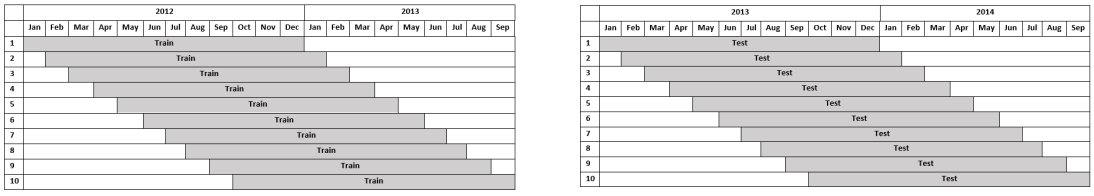
POI category	**Total count**
Food	1646
residence	484
arts & entertainment	414
college & university	365
nightlife spot	379
outdoors & recreation	1112
professional & other places	2688
shop & service	4525
travel & transport	1197
event	19

Table 4.4: Total check-ins for each time interval

Time interval	**Total check-ins**
0 (0-3)	268
1 (3-6)	228
2 (6-9)	1575
3 (9-12)	2211
4 (12-15)	2482
5 (15-18)	2050
6 (18-21)	2479
7 (21-24)	878

where, TPR and TNR represent the true positive rate and true negative rate respectively. At the same time, for the comparison with the baseline method, we report the same evaluation metrics. As our test data is class-imbalanced, we do not consider using accuracy in this phase we used in our previous phase for performance evaluation.

We developed a 10-fold time-constrained validation approach that is relatively similar to Out-of-sample (OOS) approach [25]. However, regarding train-test split point, we try to preserve seasonal patterns for both parts. In this approach, we guarantee that the records in the training set happened before the ones used for testing, and so we are effectively using data from the past to predict the future. We consider a sliding time window of 2 years, where the first 12 months are taken for training the models and the subsequent 12 months are used for testing. Thus, the models are still capable of capturing seasonality patterns as the training split always contains a full year of data. As our dataset includes three consecutive years of crime records from 2012 to 2014, for the first fold we take all records from January 2012 to December 2012 for training, and the test split goes from January 2013 to December 2013. Next, for the second fold we slide the window one month forward so that the training set spans from February 2012 to January 2013, and the test spans from February 2013 to January 2014. We repeat this process for 10 different folds. Figure 4.4 depicts the process of train and test splits.



(a) Training split.

(b) Testing split.

Figure 4.4: Train-test split procedure: for each split, training samples hold crime records that occurred in the past.

4.5 Results and Discussion

4.5.1 Results for our Proposed Features

In Table 4.5, we show the classification results with various feature combinations where the model predicts if there will be a crime or not. We tested the addition of four different groups of features to the Raw dataset (temporal + historic crime) (R): Demographic (D), Streetlight (S), Foursquare dynamic (F), and Foursquare POI (P) features.

We compare 12 different models by adding all feature categories one by one with the raw features. Our first model is implemented based on the raw features only, named as model MR. We built models MD, MS, MF, and model MP by adding demographic, streetlight, foursquare check-in, and foursquare poi data, respectively, with the raw data. Similarly, by combining two consecutive feature groups with the raw data, we built the models MDS, MDF, MDP, MSF, MSP, and model MFP. Finally, model MA is implemented based on all of the feature combinations. Both RF and GB classifiers share a similar trend for all models based on the AUC score and Gmean. As GB performs better than RF for all combinations, in our discussions, we only consider the GB method. Model MR, trained only with raw features, is resulting in a low AUC score of 59.94% and 58.42% Gmean score. Such behavior is expected since criminal behavior is affected by many different variables other than simple spatial and temporal factors [120].

By analyzing the addition of each group of features individually (top part of Table 4.5), the inclusion of demographic features (model MD) exhibits the best results, for which GB shows an improvement of almost 10% in AUC (70.02%) and about 11% in Gmean (70.01%) compared to only raw features. Similarly, streetlight features in model MS show an approximate 9% and 10% improvement for AUC and Gmean, respectively. Demographic variables reveal most of the characteristics of different regions, including social and economic factors, which are commonly correlated with criminality. Likewise, the installment of streetlight poles that reflects streetlight density feature also considers the same demographic profile for each corresponding area. Interestingly, Foursquare dynamic features (F) achieve less accuracy individually as compared to demographic and streetlight features. One of the reasons for this may

be that there is missing information for check-in data for some dissemination areas. However, dissimilar to dynamic features, Foursquare POI features (P) perform better, for instance, 69.92% in AUC and 69.87% in Gmean.

Table 4.5: Results for average AUC and Gmean scores for 12 different models based on five feature categories combination

		Features					Random Forest		Gradient boosting	
No.	**Model**	**R**	**D**	**S**	**F**	**P**	**AUC (%)**	**Gmean (%)**	**AUC (%)**	**Gmean (%)**
1	MR	✓					59.50	58.40	59.94	58.42
2	MD	✓	✓				69.15	69.12	**70.02**	**70.01**
3	MS	✓		✓			68.51	68.44	68.70	68.70
4	MF	✓			✓		64.04	63.64	64.68	64.02
5	MP	✓				✓	69.27	69.21	69.92	69.87
6	MDS	✓	✓	✓			69.13	69.09	70.02	70.01
7	MDF	✓	✓		✓		69.19	69.15	70.02	70.00
8	MDP	✓	✓			✓	69.06	69.02	**70.07**	**70.05**
9	MSF	✓		✓	✓		68.01	67.86	69.07	68.92
10	MSP	✓		✓		✓	69.26	69.21	69.93	69.89
11	MFP	✓			✓	✓	69.28	69.21	69.89	69.83
12	MA	✓	✓	✓	✓	✓	69.13	69.07	**70.11**	**70.10**

Models 6 to 11 (MDS, MDF, MDP, MSF, MSP, and MFP) show the evaluation results for three feature categories combination. The AUC and Gmean scores are better and almost consistent for all models including the model with Foursquare dynamic features. The reason behind this is that all of them contain either demographic, streetlight or POI features. In model MA, we combine all five categories of features. It gives us the best results compared to every other model which are 70.11% AUC and 70.10% Gmeans scores. As Foursquare dynamic features do not lead to performance loss while combining others, in our study, we used all feature categories for building a model.

4.5.2 Comparison with a Baseline

Table 4.6 reports the AUC and Gmean scores for our one of the best performing ensemble-based models, Model MA with Gradient Boosting (GB-MA) and the baseline DNN model. Our proposed model performs significantly better than the baseline model based on AUC and Gmean scores. Though DNN can handle non-linear relationships and data dependencies among different sources, it is very challenging for the model to perform accurately for smaller domains or domains that suffer from data scarcity. This is the most likely reason for the baseline model to degrade performance.

Table 4.6: Performance evaluation for GB-MA and the baseline DNN model

Model	**AUC (%)**	**Gmean (%)**
DNN (baseline) [58]	50.28	48.56
GB-MA	**70.11**	**70.10**

4.6 Conclusions

In this phase, we study a fundamental problem of crime incidents prediction for the future time interval. We have presented a data-driven approach to see how prediction performance can be improved by integrating multiple sources of data. Specifically, we focus on exploring population-centric features with streetlight and Foursquare-based features for each dissemination area in Halifax. Our problem also considers the temporal dimension of the crime profile in depth. We compare all 5 categories of feature combinations differently and unitedly. The results show that demographic, streetlight and Foursquare POI features have strong correlations with crime. All of them show significant performance improvement for crime prediction individually and jointly. Though Foursquare dynamic data does not outperform demographic or streetlight data, it presents a satisfactory performance for crime prediction after adding with other features. Additionally, we compare our best ensemble model (i.e., Model MA with Gradient Boosting in Table 4.5) with the DNN-based baseline model. Our results show that GB outperforms the DNN baseline for the same groups of features. Therefore, applying ensemble based method leads to a better performance in predicting future crime for smaller cities, such as Halifax.

However, as it is very challenging to get accurate results for future crime prediction when sufficient data is unavailable, performing domain adaptation, as well as different transfer learning techniques using available data from a big city would be advantageous. In our next phase (Phase III), we try to work on this issue. Chapter 5 defines the third phase of our research. It introduces some ideas to extract and transfer the knowledge from multiple sources and apply those ideas for crime occurrence prediction.

Chapter 5

Phase III: Domain Adaptation and Transfer Learning

5.1 Introduction

Urban crime analysis, particularly future crime prediction is challenging due to the complex behavioral patterns of crime and urban configuration. Additionally, for smaller cities like Halifax, it is hard to get sufficient crime data and their corresponding factors for model building. Therefore, exploiting domain adaptation and transfer learning approaches have a considerable potential to reinforce the prediction problem by utilizing big cities crime data. The distributions among different cities might be the same or somewhat different. The terms, 'city' and 'domain' are used interchangeably in this chapter. We assume that our source and target domains are different but related, and the feature spaces between domains and the tasks are the same. Considering we have some labeled data available in the target domain, we perform supervised domain adaptation techniques and hence transferring knowledge between source and target domains. The task of knowledge transfer raises some issues about what knowledge can be transferred, how and when to transfer that knowledge across domains. This study is mainly devoted to resolving the first issue, i.e., which or what knowledge to be transferred.

We divide this work into five different sections. Section 5.2 provides a review of the existing research on transfer learning approaches regarding what knowledge to be learned. Section 5.3 describes the datasets used in this phase for cross-domain learning. Section 5.4 illustrates our proposed methods and scenarios for domain adaptation and transfer learning. The experimental results based on our proposed scenarios are presented in Section 5.5. Finally, conclusions are stated in Section 5.6.

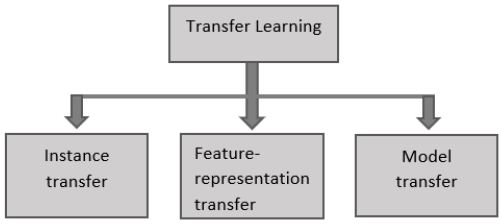


Figure 5.1: Transfer learning approaches

5.2 Literature Review

In this section, three different approaches in transfer learning (e.g., instance-transfer, feature-representation-transfer and model-transfer) based on what type of knowledge is transferred across domains [79] are discussed.

Instance transfer

In general, the instance-based transfer learning setting uses instance re-weighting and resampling techniques to obtain the relevant source instances, which can then be used with the labeled target data. In recent years, many extended boosting based ensemble learning methods have been proposed for this setting. TrAdaBoost is a widely used boosting based transfer learning algorithm that addresses the instance transfer learning problem [31]. The main goal of this method was to train a classifier using both the old (source) and new (target) domain data and transfer knowledge between different distribution instance spaces. In this approach, old data, which is very dissimilar from new data and incorrectly classified, get reduced weight. On the other hand, new target data get higher weights for misclassified examples to intensify their impacts. In 2010, Yao et al. [113] proposed an extension of TrAdaBoost, called MultiSourceTrAdaBoost by leveraging multiple sources data for knowledge transfer. The author states that using a single source domain for knowledge transfer may lead to negative transfer and performance degradation due to the weak relationships between source and target. MultiSourceTrAdaBoost follows the same strategy as TrAdaBoost by employing weights to the source and target training data, except in the weak classifier selection. In each iteration of MultiSourceTrAdaBoost, a weak classifier is chosen based on the close relationships between source and target training

data. Later, Liu et al. [64] designed a weighted resampling-based transfer learning framework (TrResampling) to improve the classification accuracy from TrAdaBoost. The algorithm resamples higher weights data in the source domain and adds this with the labeled target domain data. Then, the TrAdaBoost algorithm is applied for model building by adjusting source and target weights. Besides the resampling strategy, the study also assembled bagging-based [66] and MultiBoosting-based [65] transfer learning algorithms.

In addition to the boosting based methods, a variety of techniques exist to utilize the instances from source data. The work by Tianyang et al. [104] proposed an instance-based deep transfer learning approach for image classification problems. The authors mainly pre-trained a model using source domain data and then applied that model to labeled target training data. This strategy helps to find the optimized target training set by estimating and removing the less influential target training data. Later, this optimized target data is used for building a new model or fine-tune the previous pre-trained model. In 2016, Shuang et al. [124] proposed a source subset selection method by estimating the close relationships between source and target instances. The study employed an extension of Vovk's conformity test for this purpose.

Feature representation transfer

The Feature transfer learning setting assumes that there might be an inclusive relationship between source and target domains, and this approach tries to learn a new feature representation for the target domain. A cross-domain sentiment classification problem has been studied by Pan et al. [77] through the feature alignment approach. The authors first identify the domain-independent and mutually dependent features and then build a spectral feature alignment (SFA) algorithm to reduce the difference between domain-specific features. In another work, Xia et al. [110] presented a feature ensemble method for sentiment classification where domain-independent features get higher weights, and domain-specific features get lower weights. The key point in feature representation transfer learning is finding a good feature representation between domains with a different distribution. Pan et al. [78] proposed such a learning method named Transfer Component Analysis (TCA) for cross-domain WiFi

localization and text classification.

Model transfer

Model transfer learning is also referred to as parameter-transfer learning. This approach finds out some shared parameters of the model for related source and target domains. Parameter-transfer methods are mainly effective for multi-task learning, where the adapted model is employed to the target tasks. TaskTrAdaBoost [113] is an extension of TrAdaBoost algorithm for parameter-transfer based setting. The model identifies the shared parameters from different sources and target training part and reuses them to learn the target classifier. Another parameter-transfer method was proposed by Chattopadhyay [27] for detecting muscle-fatigue in various stages. The proposed framework relies on the conditional probability distribution differences of multi-source data, which is named as Conditional Probability based Multi-Source Domain Adaptation (CP-MDA). Differently, Segev et al. [92] proposed two model transfer learning algorithms: structure expansion/reduction (SER) and structure transfer (STRUT), based on a local transformation of a decision tree structure.

In this study, we focus on instance-based knowledge transfer. This approach is mainly motivated by importance sampling where relevant source domain data are re-weighted or/and target training subset selection before training the model. Our study is the first of its kind in utilizing such knowledge transfer approach in crime prediction.

5.3 Datasets

For cross-domain transfer learning approach, we consider crime incidents from three different cities: Halifax, Toronto and Vancouver. We consider Halifax as target domain; Toronto and Vancouver cities as the source domains.

5.3.1 Halifax Data

The data collection and mapping of raw crime data and different feature groups for Halifax city are the same described in Section 4.3. We added crime data from January 2015 to December 2015 with the previous data. For this year, we have total 17,744

records of crime incidents.

5.3.2 Toronto Data

We use Toronto Major Crime Indicators (MCI) 2014 to 2018 occurrences as source data, which is obtained from the public safety data portal of Toronto police service [4]. For the experiments, we explore crime incidents from 2014 to 2015. A total of 138,668 crime points are reported after deducting invalid and null points, where 2014 includes 26,507 records and 2015 consists of 26,796 records. On the other hand, there are 3702 DAs collected for Toronto from statistics Canada 2016 census. Similar to Section 4.3, we map all crime records to one of the 3702 DAs based on their geographic location. The summary of the datasets for Toronto is given in Table 5.1. Tables 5.2 and 5.3 present total number of venues for each POI category and total check-ins for each time interval respectively. Figure 5.2 highlights the downtown Toronto area based on four different features extracted from various feature categories.

Table 5.1: Details of the dataset for Toronto

Dataset	**Source**	**Total data**
Historical crime data	Toronto public safety data portal	138,668
Dissemination area data	Statistics Canada	3702
Demographic data	Canadian Census Analyser	3702
Foursquare POI data	Foursquare	17004
Foursquare checkin data	Foursquare	123,397

5.3.3 Vancouver Data

Besides Toronto dataset, we explore crime occurrences for the city of Vancouver as source data obtained from the Vancouver Open Data Catalogue. The data includes 24,573 crime records for the year 2014. After mapping the no crime records for each time interval and DA, we have a total of 49,146 records. According to the statistics Canada 2016 census, the city contains 993 Dissemination Areas. Like Halifax and Toronto, Vancouver's demographic data is picked up from the Canadian census analyser [2]. On the other hand, the Foursquare POIs and check-in data are missing

Table 5.2: Total POIs for each category (Toronto)

POI category	**Total count**
Food	4800
residence	1197
arts & entertainment	664
college & university	513
nightlife spot	851
outdoors & recreation	1236
professional & other places	2813
shop & service	3102
travel & transport	771
event	2

Table 5.3: Total check-ins for each time interval (Toronto)

Time interval	**Total check-ins**
0 (0-3)	4474
1 (3-6)	1758
2 (6-9)	11604
3 (9-12)	16761
4 (12-15)	24744
5 (15-18)	22972
6 (18-21)	28451
7 (21-24)	12633

(a) Crime density.

(b) Population count.

(c) Population density.

(d) POI count.

Figure 5.2: Toronto's crime and demographic information focusing on downtown area. (a) Crime density, (b) Population count, (c) Population density, and (d) POI count. The bin size for all four images is 100. The dark red color indicates high concentrated area and light red indicates the opposite.

for Vancouver. Table 5.4 shows the total data and data sources for Vancouver city. Figure 5.3 outlines some features derived from the crime data and demographic data. The dark red and light red color define high density and low density respectively. The bin size for the presented features is identical.

5.4 Experimental setup

As discussed in Section 5.3, we use Halifax data as target domain data where a subset of year 2014 data is used for training and year 2015 data are for testing. On the other

Table 5.4: Details of the datasets for Vancouver city

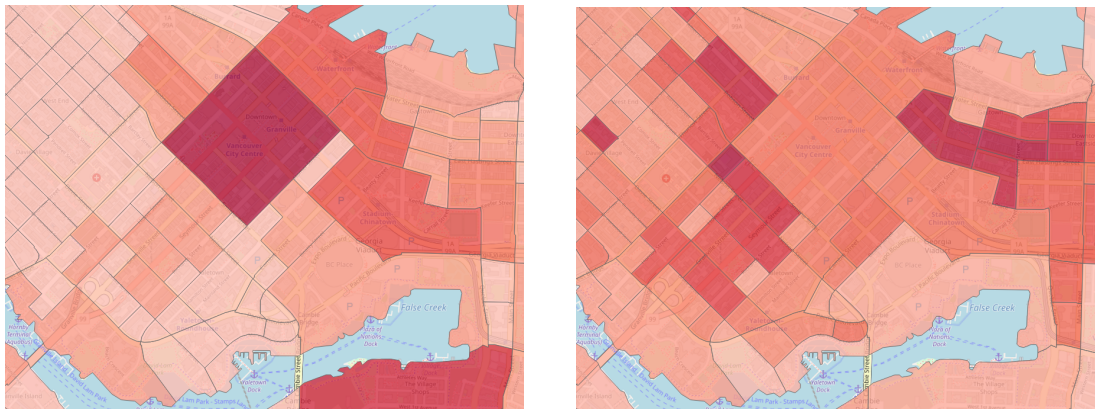
Dataset	**Source**	**Total data**
Historical crime data	Vancouver Open Data Catalogue	24,573
Dissemination area data	Statistics Canada	993
Demographic data	Canadian Census Analyser	993

hand, 2014 Toronto and Vancouver data are employed for training as source data.

5.4.1 Multi-source Domain Adaptation

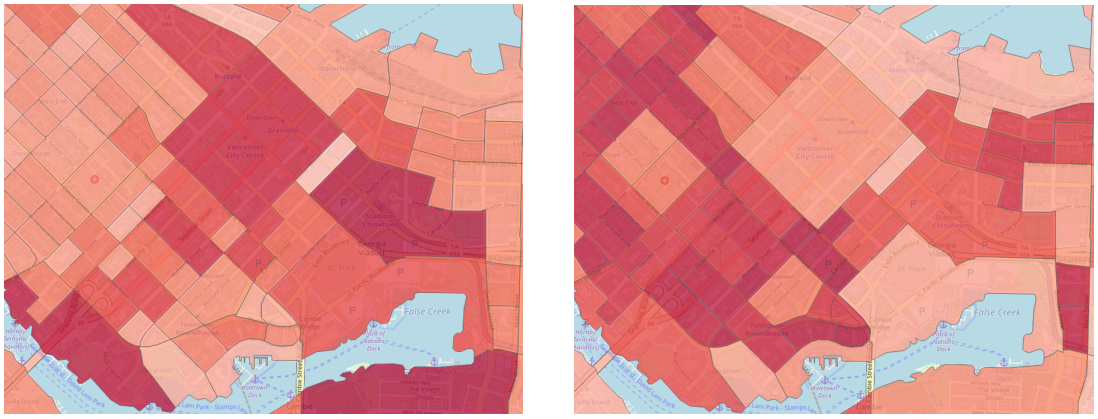
In general, most of the existing transfer learning algorithms rely on a single source domain. However, transferring knowledge from only one source may lead to negative transfer occurrences and hence, performance degradation. In transfer learning, negative knowledge transfer implies that the knowledge learned from the model by utilizing the source domain adversely affects the performance [24]. The model confronts the negative knowledge transfer problem if the source and target domains are distantly related. The efficiency of transferring positive knowledge mainly depends on the relationship between source and target domains. Therefore, leveraging the multi-source domain helps find the related source and target domains and reduce the possibility of negative transfer by importing knowledge from the closely related source. Before approaching knowledge transfer among domains, we analyze the distribution differences of three domains in feature space and label space. Afterward, the domains are adapted to an individual representation by reducing the distances among them. The distributions of three different cities based on population density and mobility migrants rate are shown in Figure 5.4 (a) and (b), respectively. The distributions of population density are nearly related among the three domains. However, for the mobility migrants rate, the distributions between Halifax and Vancouver domains are roughly different.

For domain adaptation, we propose to apply two approaches: (i) local min-max



(a) Crime frequency.

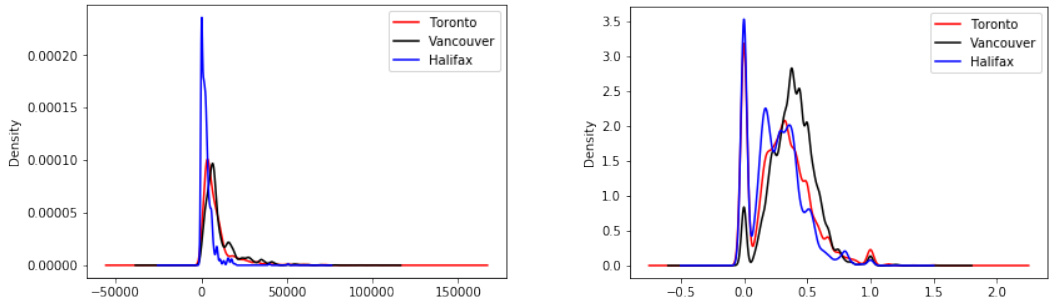
(b) Crime density.



(c) Population count.

(d) Population density

Figure 5.3: Vancouver's crime and demographic information focusing on downtown area. (a) Crime frequency, (b) Crime density, (c) Population count, and (d) Population density. The bin size for all four images is 100.



(a) Population density.

(b) Mobility migrants rate.

Figure 5.4: Distribution differences among domains

normalization, and (ii) global min-max normalization, motivated by the work of cross-building energy forecasting [88]. The general min-max normalization formula is represented as:

$$N(x_{ij}) = \frac{x_{ij} - min(X_j)}{max(X_j) - min(X_j)}, \tag{5.1}$$

where, i = 1,..., n; j = 1,..., m; n is the total number of instances; and m is the total number of features.

Local min-max normalization focuses on the relative relationship between an input feature (X_j) and an output variable (Y). This approach considers each domain locally and uses the local minimum and maximum values for normalization. Suppose, $N()$ is a normalization function and we have 2 source domains ($S = 2$). For local normalization, we calculate $N(X_j^1)$ and $N(X_j^2)$ separately.

Global min-max normalization gives particular attention to the absolute relationship between X_j and Y. This approach considers each domain as a subset of a global domain (D) where the feature (X_j) of each domain belongs to a superset J. We calculate global normalization as $N(X_j^D, J)$.

For knowledge transfer, instead of using single modality data, we utilized all feature categories extracted in Section 4.2 with multimodal characteristics. Figure 5.5 exhibits an example of transferring knowledge from Toronto to Halifax city with four feature categories for the task of crime occurrence prediction. Here, R, D, P and F indicate raw, demographics, Foursquare POI and Foursquare dynamic feature categories respectively. In this figure, we assume that the data from Foursquare POI (P) and dynamic (F) sources are not sufficient for Halifax city in comparison with Toronto. In such situations, we can learn knowledge from the source domain, Toronto regarding the inherent relationships between people's movement in any specific POI and crime occurrences. Later, we can use this knowledge to predict crime occurrences in Halifax based on the dynamic features despite the fact that there exists a data insufficiency problem.

We build six different models of knowledge transfer based on cross-domain data fusion. Model 1 is implemented based on the available target training data. However, there might be an overfitting problem if the available target training set is limited. Given no labeled data from the target domain, model 2 is built upon the Toronto source only. Similarly, model 3 is based on the Vancouver data only. On the other

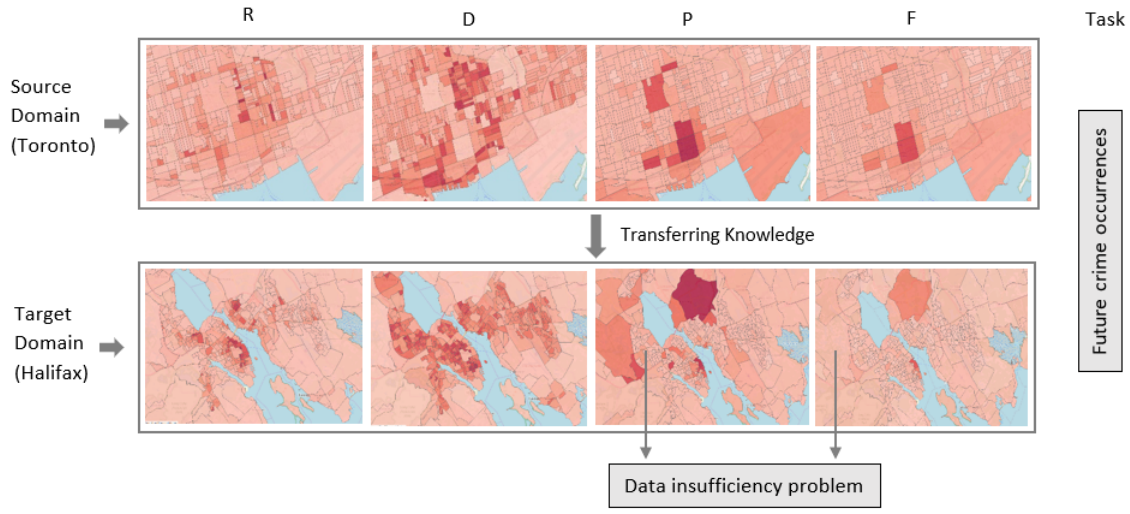


Figure 5.5: Transferring knowledge from Toronto (source) to Halifax (target). R: Raw features, D: Demographic features, P: Foursquare POI features and F: Foursquare dynamic features.

hand, model 4 belongs to the Toronto source data as well as available Halifax data. Model 5 includes Vancouver source along with the available data from the Halifax target. Model 6 imports knowledge from both the Toronto and Vancouver sources to find the relatedness with the target domain. The model representation for multi-source data is writing below:

$$\text{Model } 1 : H \rightarrow H$$

$$\text{Model } 2 : T \rightarrow H$$

$$\text{Model } 3 : V \rightarrow H$$

$$\text{Model } 4 : T \cup H \rightarrow H$$

$$\text{Model } 5 : V \cup H \rightarrow H$$

$$\text{Model } 6 : T \cup V \cup H \rightarrow H$$

Here, H, T, and V indicate Halifax, Toronto, and Vancouver domains, respectively. We tested models 2 and 4 using all feature categories extracted before except street-light features due to the lack of this category in the Toronto domain. Models 3, 5, and 6 are implemented based on raw (R) and demographic (D) features only. For our current study, foursquare feature categories (P and F) are unavailable for Vancouver

domain. The feature representation for different models is given below:

$$\text{Model 1} : R \cup D \cup S \cup F \cup P$$

$$\text{Model 2} : R \cup D \cup F \cup P$$

$$\text{Model 3} : R \cup D$$

$$\text{Model 4} : R \cup D \cup F \cup P$$

$$\text{Model 5} : R \cup D$$

$$\text{Model 6} : R \cup D$$

5.4.2 Seasonal-subset Selection

Besides source domain data, our model representation relies on a small amount of target domain data that we call target training set. Considering the impact of seasonality on crime occurrences as well as the fact of seasonal target predictive set, our first approach of selecting target training set focuses on seasonal aspects. Figure 5.6 shows six different scenarios based on seasonal perspectives. Model 1, which only considers a small amount of target (Halifax) data for training, uses consecutive seasons for knowledge transfer. Similarly, model 2 and 3 utilize consecutive seasons from Toronto and Vancouver domains respectively for model building. However, it assumes there is no available target training data for the study. On the other hand, model 4 transfers consecutive seasonal instances from the Halifax domain and all 4 seasons from the Toronto domain. Likewise, in model 5, all successive seasonal instances from Halifax domain and all 2014 data from Vancouver domain are used for instance transfer. In model 6, we assume all seasonal instances from Toronto and Vancouver domains are available along with the consecutive Halifax seasonal instances.

5.4.3 Prediction Model

Motivated by the fact that ensemble learning methods can adopt generalization [53] on different domains where the distributions are also different, we consider using ensemble based machine learning methods for our transfer learning phase. We mainly focus on Gradient Boosting (GB) [43] classifier to run the experiment under instance-transfer learning paradigm. The key reasons to choose GB over other popular machine learning algorithms are its enticing qualities as well as the challenging crime data

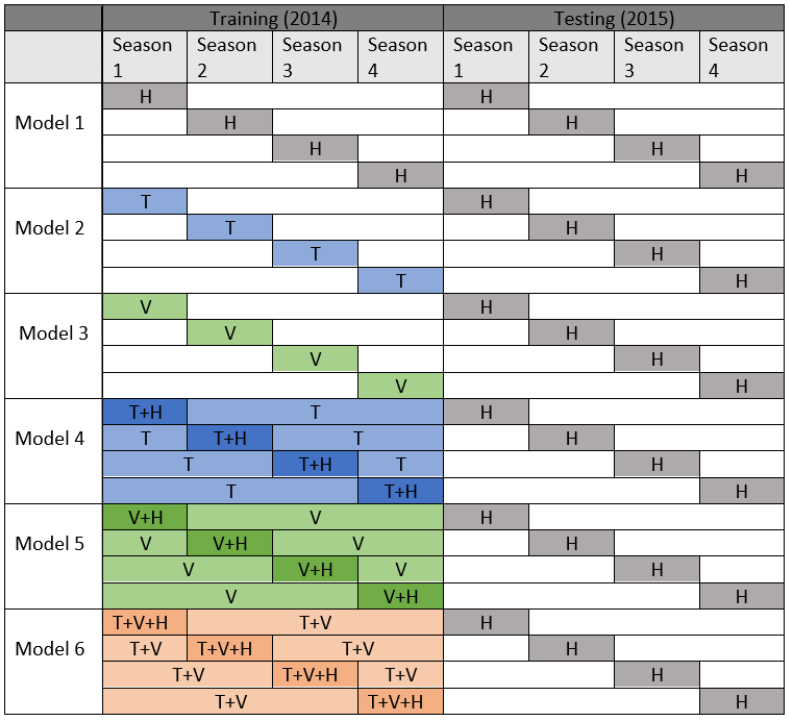


Figure 5.6: Six different scenarios based on seasonal perspective

characteristics. Other than generalization capability on unseen data, GB can handle nonlinear relationships among diverse sources of data. Moreover, the model does not require large datasets to evade overfitting problem, as well as maximum effort and attention for data cleaning and preparation. Similar to phase II, we applied randomized grid-search technique to find out the optimized parameter settings for the selected classifiers. We compare our results with a popular bagging ensemble method: Random Forest (RF) and some well-known boosting ensemble based transfer learning methods: TrAdaBoost and TrResampling. Tables A.1 and A.2 in Appendix A present the hyper parameter settings used for Random Forest and Gradient Boosting methods respectively.

Gradient Boosting (GB)

Gradient boosting is a boosting ensemble learning framework which learns from the previous mistakes. Instead of updating weights of misclassified crime points like AdaBoost, GB calculates the residual errors of the model trained on decision tree.

Let the training data set be $T = (x_i, y_i)$, where $i = 1$ to n and y_i is the class label with value 0 or 1. To calculate the residual, we have to identify a differentiable loss function, $L(y_i, F(x))$ which measures the difference between the observed class and the predicted class. The steps for model building with the generic gradient boosting method are:

- compute the base model with decision tree. Here, the model is initialized with a constant value.

$$F_0(x) = \arg\min_{y_p} \sum_{i=1}^{n} L(y_i, y_p) \tag{5.2}$$

  Here, y_p is the predicted value and for the loss function log loss can be used for classification problem. The goal of this setting is to find out a predicted value such that the whole loss would be minimized.

- iterate the following steps for $m = 1$ to M:

  (i) calculate the pseudo-residuals:

$$error_i = -[\frac{\delta L(y_i, F(x_i))}{\delta F(x_i)}] \tag{5.3}$$

  (ii) fit the decision tree using the above pseudo-residuals as target variables

  (iii) add the predicted residuals from (ii) to the previous predictions

- Get the final prediction $F_M(x)$

5.5 Results and Discussion

Similar to Phase II, as we have class-imbalanced test data, for method's performance evaluation we analyze the AUC and Gmean scores.

Table 5.5 evaluates the performance metrics out of GB classifier with RF classifier for six different models (described in Section 5.4.1). This evaluation is based on the season specific training set. Though RF and GB classifiers exhibit similar patterns, GB performs better than RF for the majority of the models. According to the results from GB classifier, model 1 reports 69.25% AUC and 68.95% Gmean scores based on season 1. The rest of the seasons also show a similar performance except season 3.

On the other hand, model 2, trained with Toronto source only data is resulting in a low AUC and Gmean scores for almost all of the seasons in comparison with model 1. For instance, model 2 degrades approximately 4% AUC (65.17%) and Gmean (65.14%) scores for season 1. Likewise, model 3, built on Vancouver source only data reduces approximately 6% performance for AUC and Gmean scores. The performance loss may happen due to the weak connections between Vancouver and Halifax domains. Moreover, the outcomes highlight the significance of comprising accessible target specific instances for model building. Model 4, which incorporates the target training set with Toronto source, exhibits the best performance among all six seasons. For instance, model 4 presents around 1%, 5% and 7% performance improvement with AUC (70.25%) and Gmean (69.73%) scores compared to model 1, 2 and 3 respectively for season 1. It tells that adding instances from Toronto data with Halifax promotes positive knowledge transfer. Model 5 and 6 show almost similar results as model 4 for each season, i.e., adding Toronto and Vancouver sources together does not help to enhance performance. The reason might be (1) the data scarcity of foursquare feature categories in the source domain, and/or (2) negative knowledge transfer due to the distant relationships between source and target domains.

Table 5.6 compares the AUC and Gmean scores achieved from GB classifier with two base transfer learning algorithms: TrAdaBoost and TrResampling for model 4. The AUC scores learned from GB classifier show approximately 4.8% and 3.5% improvement compared to TrAdaBoost and TrResampling methods respectively for season 1. Similarly, for Gmean our proposed algorithm promotes 4.3% against TrAdaBoost and 4.8% against TrResampling. Rest of the seasons also show the similar patterns. The most probable reasons for base learners to degrade performance are, TrAdaBoost is sensitive to the quality of instances from the different distributions, as well as it can not handle multi-source different distributions data. Similar to TrAdaBoost, TrResampling faces the same kinds of issues, i.e., the negative knowledge transfer problem though it performs slightly better than TrAdaBoost. The results of model 5 learnt from Vancouver and Halifax sources are shown in Table 5.7. Table 5.8 evaluates the results from multi-source (Toronto and Vancouver) data based on model 6. Gradient Boosting performs better than base algorithms for both models 5 and 6 in the same way as model 4.

Figure 5.7 depicts the overall picture of models 4, 5, and 6 applying seasonal subset selection based on AUC scores.

Table 5.5: Performance evaluation based on six different models

		Random Forest		Gradient boosting	
Season	Model	AUC (%)	Gmean (%)	AUC (%)	Gmean (%)
1	1	69.69	69.25	69.25	68.95
	2	64.34	63.39	65.17	65.14
	3	63.18	62.61	63.17	62.69
	4	69.55	69.13	**70.25**	**69.73**
	5	70.00	69.48	70.05	69.74
	6	69.69	69.11	70.10	69.71
2	1	69.61	69.57	69.25	69.24
	2	64.47	64.35	65.04	64.94
	3	63.09	62.10	63.74	63.54
	4	69.54	69.50	**70.63**	**70.59**
	5	69.41	69.35	70.35	70.35
	6	69.28	69.20	70.52	70.52
3	1	67.57	67.55	67.54	67.49
	2	63.74	63.65	63.86	63.66
	3	62.89	62.72	62.54	62.06
	4	67.42	67.41	**68.70**	**68.68**
	5	67.57	67.56	67.99	67.87
	6	67.52	67.51	68.12	68.04
4	1	68.76	68.68	68.12	68.08
	2	64.20	64.08	64.80	64.54
	3	62.85	62.84	62.41	62.37
	4	68.65	68.54	**69.23**	**69.09**
	5	68.79	68.66	68.99	68.96
	6	68.80	68.66	69.11	69.07

Table 5.6: AUC and Gmean scores based on Toronto and Halifax data (model 4)

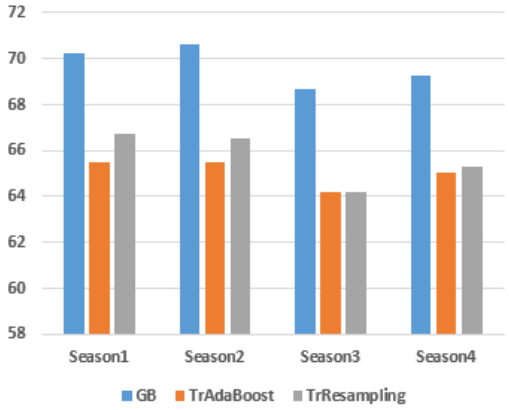
Model 4	Gradient boosting		TrAdaBoost		TrResampling	
Season	AUC (%)	Gmean (%)	AUC (%)	Gmean (%)	AUC (%)	Gmean (%)
1	70.25	69.73	65.46	65.40	66.70	64.88
2	70.63	70.59	65.46	65.16	66.50	65.14
3	68.70	68.68	64.17	63.80	64.20	63.40
4	69.23	69.09	65.02	64.94	65.30	64.58

Table 5.7: AUC and Gmean scores based on Vancouver and Halifax data (model 5)

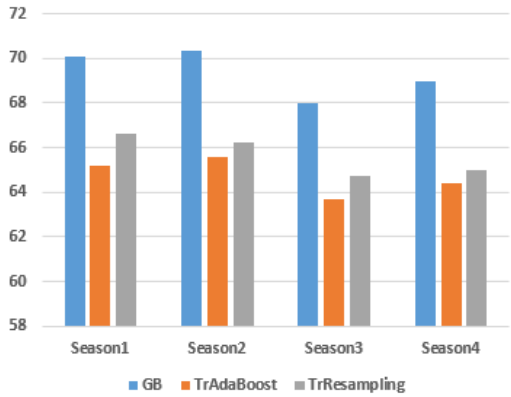
Model 5	Gradient boosting		TrAdaBoost		TrResampling	
Season	AUC (%)	Gmean (%)	AUC (%)	Gmean (%)	AUC (%)	Gmean (%)
1	70.05	69.74	65.17	65.15	66.6	65.17
2	70.35	70.35	65.59	65.49	66.2	65.11
3	67.99	67.87	63.70	63.57	64.7	63.56
4	68.99	68.96	64.37	64.26	65.0	63.82

5.6 Conclusions

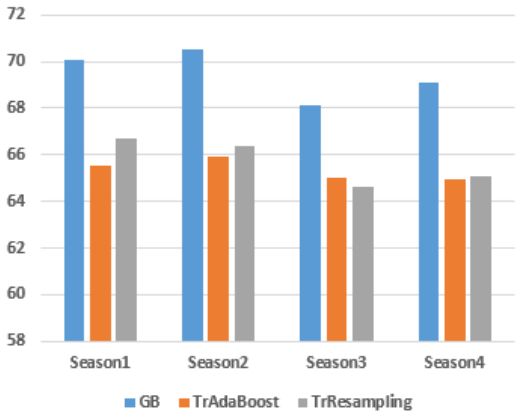
In this phase, we exploit transfer learning framework on urban crime research by adapting domains with different distributions and transferring knowledge among



(a) model 4 based on seasonal subset



(b) model 5 based on seasonal subset



(c) model 6 based on seasonal subset

Figure 5.7: Comparison of AUC scores. (a) Model 4 seasonal subset, (b) Model 5 seasonal subset, (c) Model 6 seasonal subset.

Table 5.8: AUC score and Gmean based on multisource data (model 6)

Model 6	Gradient boosting		TrAdaBoost		TrResampling	
Season	AUC (%)	Gmean (%)	AUC (%)	Gmean (%)	AUC (%)	Gmean (%)
1	70.10	69.71	65.53	65.52	66.70	64.98
2	70.52	70.52	65.98	65.84	66.40	64.90
3	68.12	68.04	65.00	64.88	64.60	63.40
4	69.11	69.07	64.95	64.92	65.10	63.65

them. We propose a number of approaches for training sample selection using multi-source domains to deal with the crime prediction problem. Our experiments illustrate that the performance vary over different models with seasonal perspectives. Gradient Boosting with transfer learning settings outplays the base transfer learning algorithms, TrAdaBoost and TrResampling for each model.

Chapter 6

Conclusions and Future Work

Our concluding remarks and some future research directions are presented in this chapter. The chapter is divided into two sections. Section 6.1 summarizes the contribution and outcomes of our research, Section 6.2 presents several future research ideas regarding crime prediction based on cross-domain learning.

6.1 Summary

Crime research plays an impactful role in developing more effective policies and improving the quality of urban life. This thesis studied a data-driven approach for crime occurrence prediction by fusing multimodal data from different domains. The problems are demonstrated in three different phases.

In the first phase, we review the literature on crime pattern detection and prediction for both single and cross-domain aspects. We introduce the factors that might influence crime scope and quantitative research for crime prediction. This phase extracts knowledge regarding crime and criminal activities' spatial and temporal relationships and solves the crime prediction problem by examining Halifax's single domain. As the crime distribution and behavioral patterns are diverse for different types of crime, we learn each type of crime individually in this phase. We explored the creation of spatial features derived from geolocated data, and created two types of spatial features. The first used a geocoding service that can query OSM data and return a category and a type of information regarding where the crime occurred. The second used the HDSCAN algorithm to create *hotspots* grouped by type of crime, extracted a *hotpoint* from each *hotspot*, and finally returned the shortest distance for a *hotpoint* as a feature to feed a classifier. The new features were evaluated using four different crime types (alcohol-related, assault, property damage, and motor vehicle) using only the UCR forms' information as features for a classifier as the baseline. The results showed a significant improvement in accuracy and AUC when the newly

engineered features were added to the tested classifiers.

In the second phase, we investigated streetlight infrastructure, demographic profiling as well as human behavioral patterns along with geographic profiling. This phase intends to build data-driven models to solve crime prediction problems focusing on smaller cities. We examine the spatial and temporal relationships of streetlight density with crime occurrences, including sociodemographic measures. We consider exploring commuting patterns of different residents' groups with traditional demographic measures. To observe human behavior across the city, we proposed using Foursquare POI and check-in features. We tested the effect and significance of all investigated feature combinations based on each Dissemination Area (DA) in Halifax for each time interval. The experiments are conducted, including ensemble-based machine learning methods: Random Forest and Gradient Boosting. The results reveal a strong correlation between extracted features and crime. Adding demographic measures, POI, and streetlight features significantly improves the prediction performance. After comparing the results with the DNN-based baseline, concludes that the DNN-based model fails with the current data-driven setup.

Finally, the third phase addresses domain adaptation and transfer learning paradigms for the crime prediction problem. We implement a data-driven approach by investigating all feature combinations for cross-domain learning. As it is challenging to prepare enough labeled training data based on a small city like Halifax, we examine multi-source domain adaptation by leveraging knowledge from two other domains: Toronto and Vancouver. We propose to apply instance-based transfer learning techniques for transferring knowledge between source and target domains. For instance, we propose different settings based on season-specific subset selection with cross-domain data fusion. We mainly focus on ensemble learning methods for cross-domain learning because of its generalization ability with new data. We evaluate the GB classifier for all proposed setups and compare the results with two base transfer learning algorithms: TrAdaBoost and TrResampling. Based on our experiments, the ensemble-based GB classifier improves the AUC scores by average 4% with TrAdaBoost and 3.8% with TrResampling for multi-source data. From all the experiments on instance transfer learning, we can conclude that the GB classifier works better when available target specific instances are added to the Toronto source.

6.2 Future Research

This section discusses a number of future research works and ideas regarding multi-source domain adaptation in crime pattern detection and prediction.

Predicting Specific Types of Crime. Identifying specific types of crime that might happen in the near future is our immediate concern for cross-domain learning. From our analysis based on the Halifax domain, we observe that different types of crime exhibit different spatial and temporal distributions. Correspondingly, behavioral patterns, mobility, and networking might be different for individual crime types. Therefore, investigating the prediction performance and the significance of individual features for cross-domain study capturing various crime categories is of great importance.

Feature-Representation and Model Transfer Learning. In our study, we mainly focus on the instance transfer learning problem. For feature representation, we highlight the common features among source and target domains. However, learning a good feature representation and transferring that knowledge to the target domain is simultaneously important. Particularly, when a full structure is missing for any specific modalities (e.g., POI data is missing in Vancouver domain), rather than picking just common features learning feature knowledge from the other domains will be advantageous.

Adding particular types of crime might experience having different but related tasks for source and target domains. In such cases, exploring parameter-transfer and a relational-knowledge transfer would be interesting.

Incorporating Crime Data with Multimodal Data. A direction we want to follow is to integrate environmental context information, including images and videos to the current dataset and explore the performance of models when such information is available. Features extracted from image and video data may provide significant perceptions about a region and its characteristics. For instance, an image of a filthy or clean spot tells the story about its surroundings and inhabitants. Similarly, a video of anomalous situations helps us identify various abnormal activities such as

road accidents, shoplifting, robbery, and fighting. We can obtain image data from a location-based social networking site, Flickr ($https://www.flickr.com/$). Moreover, we can learn knowledge utilizing real-world surveillance videos by following the UCF-Crime dataset [93].

Other than image and video data, the correlation between meteorology and crime will also be explored in our study. We also plan to connect text data from our previous study with all geolocated crime data.

Moreover, for distance calculation between a crime point and a hotpoint (discussed in Section 3.3.2), we plan to add proxy-distance along with the haversine great-circle distance. For instance, considering travel time and travel mode to determine the actual distance, we could utilize the web map services (e.g., google map) and fetch the walking distance, driving, and/or road closures between two points; later, use this as a new feature.

Exploring Discrimination Prevention Techniques. Investigating discrimination in socially-sensitive decision records is state-of-the art research to avoid biased classification learning. In a societal context, discrimination indicates unjust or unequal action of people based on preconception. If protected attributes such as gender, race have an explicit contribution to decision-making or dependency on other correlated features, discrimination may also occur in the trained model. As we are using real-world crime data for our study, investigating and preventing discrimination are highly crucial before decision making. We plan to investigate a pre-processed discrimination prevention technique on our crime data by following the idea from Calmon et al. [22]. The study includes three properties: discrimination control, distortion control, and utility preservation. Apart from the pre-processed discrimination prevention, we will also explore the post-processing approach [50] for discrimination prevention.

Model Interpretability and Explaining Individual Prediction. As many advanced machine learning algorithms act like a black box model, their trustworthiness has come into question. The literature provides a detailed review of different

approaches to uncover black-box model considering the significance of model interpretability for the real-world problem [48]. It is very intuitive to explain any classifier's individual predictions to take action based on that prediction. In 2016, an interpretable model called Local Interpretable Model-Agnostic Explanations (LIME) has been successfully applied on an image classification problem [87]. LIME helps to explain individual predictions of any classifier in a faithful way. Therefore, it would be interesting to investigate this approach on crime prediction problem, particularly its transfer learning part, to identify which features lead to the positive contribution of transfer learning.

Deep learning for Complex Pattern Detection. Though the DNN based prediction model employed in our research did not perform well, incorporating image data for the next phase directs us re-investigate the model for a cross-domain data-driven approach. Besides, the study for detecting parking hotspots among cross-city [121] motivates us to explore Convolutional City Domain Adaptation Network (ConvCDAN) on crime research.

To generalize deep learning model on cross-domain crime data with distributions discrepancy, we intend to explore a sampling-based method called Implicit Class-Conditioned Domain Alignment [55]. The method assumes that there is no labeled data in target domain, as well as the source and target tasks might be different. Therefore, the labeled source domain data is aligned with unlabeled target domain data through uniform alignment distributions. This way the algorithm selects class-aligned instances for training domain adaptation model. As we have class imbalanced source and target domains, investigating implicit alignment approach might be useful.

On the other hand, a prospective approach to deal with the data insufficiency problem in urban crime data is applying meta-learning with medium-shot learning [54]. This approach utilizes both the gradient-based and metric-based meta-learning methods to advance the performance.

Bibliography

[1] Census Profile - Halifax (population centre). Statistics Canada. Archived from the original on February 11, 2017. Accessed: 2017-02-08.

[2] Profile of Census Dissemination Areas. Accessed: 2019-07-02.

[3] Statistics canada. 2016 census - boundary files. Accessed: 2019-07-02.

[4] Toronto police service - public safety data portal. Accessed: 2019-05-30.

[5] Urban population by city size. OECD. Accessed: 2018-09-08.

[6] *Canadian alcohol and drug use monitoring survey*. Statistics Canada, 2012.

[7] *Crimes, by type of violation, and by province and territory (2012)*. Statistics Canada, 2014.

[8] Monsuru Adepeju and Tao Cheng. Determining the optimal spatial scan of Prospective space-time scan statistics (PSTSS) for crime hotspot prediction. In *25th GIS Research UK Conference (GISRUK 2017), Geographical Information Science Research UK*, 2017.

[9] Monsuru Adepeju, Gabriel Rosser, and Tao Cheng. Novel evaluation metrics for sparse spatio-temporal point process hotspot predictions - a crime case study. *International Journal of Geographical Information Science*, 2016.

[10] Mohammad Al Boni and Matthew S. Gerber. Automatic optimization of localized kernel density estimation for hotspot policing. In *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 2017.

[11] Tahani Almanie, Rsha Mirza, and Elizabeth Lor. Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots. *International Journal of Data Mining & Knowledge Management Process*, 2015.

[12] Fateha Khanam Bappee. Identification and classification of alcohol-related violence in nova scotia using machine learning paradigms. In *Advances in Artificial Intelligence - 30th Canadian Conference on Artificial Intelligence, Canadian AI 2017, Edmonton, AB, Canada, May 16-19, 2017, Proceedings*, pages 421–425, 2017.

[13] Fateha Khanam Bappee, Amilcar Soares Junior, and Stan Matwin. Predicting crime using spatial features. In *Canadian Conference on Artificial Intelligence*, pages 367–373. Springer, 2018.

[14] Fateha Khanam Bappee, Lucas May Petry, Amilcar Soares, and Stan Matwin. Analyzing the impact of foursquare and streetlight data with human demographics on future crime prediction, 2020.

[15] Alexandros Belesiotis, George Papadakis, and Dimitrios Skoutas. Analyzing and predicting spatial crime distribution using crowdsourced and open data. *ACM Trans. Spatial Algorithms Syst.*, 3(4):12:1–12:31, April 2018.

[16] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Emmanuel Letouzé, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Moves on the Street: Classifying Crime Hotspots Using Aggregated Anonymized Data on People Dynamics. *Big Data*, 2015.

[17] Andrey Bogomolov, Bruno Lepri, Jacopo Staiano, Nuria Oliver, Fabio Pianesi, and Alex Pentland. Once upon a crime: Towards crime prediction from demographics and mobile data. *CoRR*, abs/1409.2983, 2014.

[18] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[19] Rosemary D.F Bromley and Audrey L. Nelson. Alcohol-related crime and disorder across urban space and time: evidence from a british city. *Geoforum*, 33(2):239–254, 2002.

[20] Aaron M. Brower and L. Carroll. Spatial and temporal aspects of alcohol-related crime in a college town. *Journal of American College Health*, 55:267–275, 2007.

[21] Anna L. Buczak and Christopher M. Gifford. Fuzzy association rule mining for community crime pattern discovery. In *ACM SIGKDD Workshop on Intelligence and Security Informatics*, ISI-KDD '10, pages 2:1–2:10, New York, NY, USA, 2010. ACM.

[22] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc., 2017.

[23] Ricardo J. G. B. Campello, Davoud Moulavi, Arthur Zimek, and Jörg Sander. Hierarchical Density Estimates for Data Clustering, Visualization, and Outlier Detection. *ACM Transactions on Knowledge Discovery from Data*, 2015.

[24] Bin Cao, Sinno Pan, yu Zhang, Dit-Yan Yeung, and Qiang Yang. Adaptive transfer learning. 01 2010.

[25] Vítor Cerqueira, Luís Torgo, and Igor Mozetic. Evaluating time series forecasting models: An empirical study on performance estimation methods. *CoRR*, abs/1905.11744, 2019.

[26] Spencer Chainey, Lisa Tompson, and Sebastian Uhlig. The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal*, 21(1):4–28, 2008.

[27] Rita Chattopadhyay, Qian Sun, Wei Fan, Ian Davidson, Sethuraman Panchanathan, and Jieping Ye. Multisource domain adaptation and its application to early detection of fatigue. *ACM Transactions on Knowledge Discovery from Data*, 6(4), December 2012.

[28] Xinyu Chen, Youngwoon Cho, and Suk Young Jang. Crime prediction using twitter sentiment and weather. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 63–68, 2015.

[29] Corinna Cortes and Vladimir Vapnik. Support-Vector Networks. *Machine Learning*, 20(3):273–297, 1995.

[30] Michael Cusimano, Sean Marshall, Claus Rinner, Depeng Jiang, and Mary Chipman. Patterns of urban violent injury: A spatio-temporal analysis. *PLoS ONE*, 2010.

[31] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. volume 227, pages 193–200, 01 2007.

[32] Thomas G. Dietterich. Machine-Learning Research – Four Current Directions. *AI MAGAZINE*, 18:97–136, 1997.

[33] Grant Drawve. A Metric Comparison of Predictive Hot Spot Techniques and RTM. *Justice Quarterly*, 2016.

[34] Bowen Du, Chuanren Liu, Wenjun Zhou, Zhenshan Hou, and Hui Xiong. Catch me if you can: Detecting pickpocket suspects from large-scale transit records. pages 87–96, 08 2016.

[35] Lian Duan, Tao Hu, En Cheng, Jianfeng Zhu, and Chao Gao. Deep convolutional neural networks for spatiotemporal crime prediction. In *Proceedings of the 2017 International Conference on Information and Knowledge Engineering*, IKE '17, pages 61–67, 2017.

[36] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

[37] M. Lyn Exum. Alcohol and aggression: An integration of findings from experimental studies. *Journal of Criminal Justice*, 34(2):131–145, 2006.

[38] Masoomali Fatehkia, Dan OBrien, and Ingmar Weber. Correlated impulses: Using facebook interests to improve predictions of crime rates in urban areas. *PLOS ONE*, 14(2):1–16, 2019.

[39] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[40] Marcus Felson, Ronald V Clarke, and Barry Webb. *Opportunity Makes the Thief: Practical theory for crime prevention*. 1998.

[41] Jian Feng, Ying Dong, and Leilei Song. A spatio-temporal analysis of urban crime in Beijing: Based on data for property crime. *Urban Studies*, 2016.

[42] J. Fitterer, T. A. Nelson, and F. Nathoo. Predictive crime mapping. *Police Practice and Research*, 2015.

[43] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001.

[44] Vicente Garcia, Ramon Mollineda, and Josep Sanchez. Theoretical analysis of a performance measure for imbalanced data. pages 617–620, 08 2010.

[45] Geocoder. Simple and consistent geocoding library written in python, January 2013.

[46] Matthew S. Gerber. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[47] Corina Graif and Robert J. Sampson. Spatial heterogeneity in the effects of immigration and diversity on neighborhood homicide rates. *Homicide Studies*, 2009.

[48] Riccardo Guidotti, Anna Monreale, Franco Turini, Dino Pedreschi, and Fosca Giannotti. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 2019.

[49] David J. Hand and Robert J. Till. A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. *Machine Learning*, 2001.

[50] Moritz Hardt, Eric Price, None, and Nati Srebro. Equality of Opportunity in Supervised Learning. *Nips*, 2016.

[51] David E. Hojman. Inequality, unemployment and crime in Latin American cities. *Crime, Law and Social Change*, 2004.

[52] Geoffrey Holmes, Bernhard Pfahringer, and Richard Kirkby. Multiclass Alternating Decision Trees. *Proceedings of the 13th European Conference on Machine Learning (ECML '02)*, 2002.

[53] Chidubem Iddianozie and Gavin McArdle. A transfer learning paradigm for spatial networks. pages 659–666, 04 2019.

[54] Xiang Jiang, Liqiang Ding, Mohammad Havaei, Andrew Jesson, and Stan Matwin. Task adaptive metric space for medium-shot medical image classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 147–155, Cham, 2019. Springer International Publishing.

[55] Xiang Jiang, Qicheng Lao, Stan Matwin, and Mohammad Havaei. Implicit class-conditioned domain alignment for unsupervised domain adaptation. In *In International Conference on Machine Learning (ICML)*, 06 2020.

[56] Joint Committee for Guides in Metrology (JCGM/WG 2). International Vocabulary of Metrology - Basic and General Concepts and Associated Terms (VIM) 3rd Edition. *English*, 2006.

[57] Cristina Kadar, J. Iria, and Irena Pletikosa. Exploring Foursquare-derived features for crime prediction in New York City. In *KDD - Urban Computing WS '16*, 2016.

[58] Hyeon-Woo Kang and Hang-Bong Kang. Prediction of crime occurrence from multi-modal data using deep learning. *PLOS ONE*, 12(4):1–19, 04 2017.

[59] Schlesinger LB, Kassen M, Mesa VB, and Pinizzotto AJ. Ritual and signature in serial sexual homicide. *J Am Acad Psychiatry Law*, 38(2):239–46, 2010.

[60] Yann A. LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 2015.

[61] Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[62] Kelvin Leong and Anna Sung. A review of spatio-temporal pattern analysis approaches on crime analysis. *International E-Journal of Criminal Sciences*, 2015.

[63] Renjie Liao, Xueyao Wang, Lun Li, and Zengchang Qin. A novel serial crime prediction model based on Bayesian learning theory. In *Machine Learning and Cybernetics (ICMLC), 2010 International Conference on*, 2010.

[64] X. Liu, Z. Liu, G. Wang, Z. Cai, and H. Zhang. Ensemble transfer learning algorithm. *IEEE Access*, 6:2389–2396, 2018.

[65] Xiaobo Liu, Guangjun Wang, Zhihua Cai, and Harry Zhang. A multiboosting based transfer learning algorithm. *JACIII*, 19:381–388, 2015.

[66] Xiaobo Liu, Guangjun Wang, Zhihua Cai, and Harry Zhang. Bagging based ensemble transfer learning. *Journal of Ambient Intelligence and Humanized Computing*, 7:29–36, 02 2016.

[67] Abish Malik, Ross Maciejewski, Sherry Towers, Sean McCullough, and David S. Ebert. Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement. *IEEE Trans. Vis. Comput. Graph*, 20(12):1863–1872, 2014.

[68] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. London: Chapman and Hall, 3rd edition, 1997.

[69] Prem Melville and Raymond J. Mooney. Constructing diverse classifier ensembles using artificial training examples. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-2003)*, pages 505–510, Acapulco, Mexico, August 2003.

[70] P. Mohan, S. Shekhar, J. A. Shine, and J. P. Rogers. Cascading Spatio-Temporal Pattern Discovery. *IEEE Transactions on Knowledge and Data Engineering*, 2012.

[71] G. O. Mohler, M. B. Short, P. J. Brantingham, F. P. Schoenberg, and G. E. Tita. Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 2011.

[72] Yang Mu, Wei Ding, Melissa Morabito, and Dacheng Tao. Empirical discriminative tensor analysis for crime forecasting. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2011.

[73] Tomoki Nakaya and Keiji Yano. Visualising crime clusters in a space-time cube: An exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *T. GIS*, 14(3):223–239, 2010.

[74] Shyam Varan Nath. Crime pattern detection using data mining. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IATW '06, pages 41–44, Washington, DC, USA, 2006. IEEE Computer Society.

[75] Andrew Newton and Marcus Felson. Editorial: crime patterns in time and space: the dynamics of crime opportunities in urban areas. *Crime Science*, 2015.

[76] J. K. Ord and Arthur Getis. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. *Geographical Analysis*, 1995.

[77] Sinno Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Cross-domain sentiment classification via spectral feature alignment. pages 751–760, 01 2010.

[78] Sinno Jialin Pan, Ivor W. Tsang, James T. Kwok, and Qiang Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 2011.

[79] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Trans. on Knowl. and Data Eng.*, 22(10):1345–1359, October 2010.

[80] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[81] Débora V.S. Pereira, Martin A. Andresen, and Caroline M.M. Mota. A temporal and spatial analysis of homicides. *Journal of Environmental Psychology*, 2016.

[82] Cheng Qian, Yubo Wang, Jinde Cao, Jianquan Lu, and Jürgen Kurths. Weighted-traffic-network-based geographic profiling for serial crime location prediction. *EPL*, 2011.

[83] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems 32*, pages 3347–3357. Curran Associates, Inc., 2019.

[84] Matthew Ranson. Crime, weather, and climate change. *Journal of Environmental Economics and Management*, 67(3):274–302, 2014.

[85] Jerry Ratcliffe. The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. *Police Practice and Research*, 5(1):5–23, 2004.

[86] Jerry Ratcliffe. Near repeat calculator. *Philadelphia, PA: Temple University and National Institute of Justice.*, 2009.

[87] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why Should I Trust You?Explaining the Predictions of Any Classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, 2016.

[88] Mauro Ribeiro, Katarina Grolinger, Hany F ElYamany, Wilson A Higashino, and Miriam A M Capretz. Transfer learning with seasonal and trend adjustment for cross-building energy forecasting. *Energy & Buildings*, 165:352–363, 2018.

[89] Joel Rubin. Stopping crime before it starts. *The Los Angeles Times*, 2010.

[90] Shakila Khan Rumi, Ke Deng, and Flora Dilys Salim. Crime event prediction with dynamic features. *EPJ Data Science*, 2018.

[91] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 3rd edition, 2009.

[92] Noam Segev, Maayan Harel, Shie Mannor, Koby Crammer, and Ran El-Yaniv. Learn on Source, Refine on Target: A Model Transfer Learning Framework with Random Forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[93] Waqas Sultani, Chen Chen, and Mubarak Shah. Real-world anomaly detection in surveillance videos. *CoRR*, abs/1801.04264, 2018.

[94] Mohammad A. Tayebi, Martin Ester, Uwe Glasser, and Patricia L. Brantingham. CRIMETRACER: Activity space based crime location prediction. In *ASONAM 2014 - Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2014.

[95] Jameson L. Toole, Nathan Eagle, and Joshua B. Plotkin. Spatiotemporal correlations in criminal offense records. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[96] Martin Traunmueller, Giovanni Quattrone, and Licia Capra. Mining mobile phone data to investigate urban crime theories at scale. In *SocInfo*, volume 8851 of *Lecture Notes in Computer Science*, pages 396–411. Springer, 2014.

[97] B Wang, D Zhang, D Zhang, P. J. Brantingham, and A. L. Bertozzi. Deep learning for real time crime forecasting. 2017.

[98] Bao Wang, Penghang Yin, Andrea L. Bertozzi, P. Jeffrey Brantingham, Stanley J. Osher, and Jack Xin. Deep learning for real-time crime forecasting and its ternarization. *CoRR*, abs/1711.08833, 2017.

[99] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In Toby Walsh, editor, *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pages 1541–1546. IJCAI/AAAI, 2011.

[100] Dawei Wang, Wei Ding, Henry Lo, Melissa Morabito, Ping Chen, Josue Salazar, and Tomasz Stepinski. Understanding the spatial distribution of crime based on its related variables using geospatial discriminative patterns. *Computers, Environment and Urban Systems*, 2013.

[101] Hongjian Wang, Daniel Kifer, Corina Graif, and Zhenhui Li. Crime rate inference with big data. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 635–644, 2016.

[102] Mingjun Wang and Matthew S. Gerber. Using Twitter for next-place prediction, with an application to crime prediction. In *Proceedings — 2015 IEEE Symposium Series on Computational Intelligence, SSCI 2015*, 2015.

[103] Ping Wang, Rick Mathieu, Jie Ke, and H. J. Cai. Predicting criminal recidivism with support vector machine. In *International Conference on Management and Service Science, MASS 2010 International Conference*, 2010.

[104] T. Wang, J. Huan, and M. Zhu. Instance-based deep transfer learning. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 367–375, 2019.

[105] Tong Wang, Cynthia Rudin, Daniel Wagner, and Rich Sevieri. Learning to detect patterns of crime. In *ECML/PKDD (3)*, volume 8190 of *Lecture Notes in Computer Science*, pages 515–530. Springer, 2013.

[106] Xiaofeng Wang, Donald E. Brown, and Matthew S. Gerber. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *ISI 2012 — 2012 IEEE International Conference on Intelligence and Security Informatics: Cyberspace, Border, and Immigration Securities*, 2012.

[107] Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown. Automatic crime prediction using events extracted from twitter posts. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[108] Karl R. Weiss, Taghi M. Khoshgoftaar, and Dingding Wang. A survey of transfer learning. *Journal of Big Data*, 3:1–40, 2016.

[109] Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706, 2014.

[110] Rui Xia, Chengqing Zong, Xuelei Hu, and Erik Cambria. Feature ensemble plus sample selection: Domain adaptation for sentiment classification. *Intelligent Systems, IEEE*, 28:10–18, 05 2013.

[111] Yanqing Xu, Cong Fu, Eugene Kennedy, Shanhe Jiang, and Samuel Owusu-Agyemang. The impact of street lights on spatial-temporal patterns of crime in Detroit, Michigan. *Cities*, 2018.

[112] Dingqi Yang, Daqing Zhang, and Bingqing Qu. Participatory cultural mapping based on collective behavior data in location-based social networks. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 7(3):30, 2016.

[113] Y. Yao and G. Doretto. Boosting for transfer learning with multiple sources. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1855–1862, 2010.

[114] Chung-Hsien Yu, Wei Ding, Ping Chen, and Melissa Morabito. Crime forecasting using spatio-temporal pattern with ensemble learning. In *Advances in Knowledge Discovery and Data Mining - 18th Pacific-Asia Conference, PAKDD 2014, Tainan, Taiwan, May 13-16, 2014. Proceedings, Part II*, volume 8444 of *Lecture Notes in Computer Science*, pages 174–185. Springer, 2014.

[115] Chung-Hsien Yu, Wei Ding, Melissa Morabito, and Ping Chen. Hierarchical Spatio-Temporal Pattern Discovery and Predictive Modeling. *IEEE Transactions on Knowledge and Data Engineering*, 2016.

[116] Chung-Hsien Yu, Max W. Ward, Melissa Morabito, and Wei Ding. Crime Forecasting Using Data Mining Techniques. In *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011.

[117] Qingzhao Yu, Bin Li, and Richard Allen Scribner. Hierarchical additive modeling of nonlinear association with spatial correlations — an application to relate alcohol outlet density and neighborhood assault rates. *Journal of Statistics in Medicine*, 28(14):1896–1912, 2009.

[118] Xiangyu Zhao and Jiliang Tang. Exploring transfer learning for crime prediction. In *IEEE International Conference on Data Mining Workshops, ICDMW*, 2017.

[119] Xiangyu Zhao and Jiliang Tang. Modeling temporal-spatial correlations for crime prediction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM '17, pages 497–506, New York, NY, USA, 2017. ACM.

[120] Xiangyu Zhao and Jiliang Tang. Crime in urban areas: A data mining perspective. *CoRR*, abs/1804.08159, 2018.

[121] Liu Zhaoyang, Yanyan Shen, and Yanmin Zhu. Where will dockless shared bikes be stacked?: — parking hotspots detection in a new city. pages 566–575, 07 2018.

[122] Joey Zhou, Sinno Pan, Ivor Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. volume 3, 07 2014.

[123] Joey Zhou and Ivor Tsang. Heterogeneous domain adaptation for multiple classes. April 2014.

[124] Shuang Zhou, Gijs Schoenmakers, Evgueni Smirnov, Ralf Peeters, Kurt Driessens, and Siqi Chen. Largest source subset selection for instance transfer. In Geoffrey Holmes and Tie-Yan Liu, editors, *Asian Conference on Machine*

Learning, volume 45 of *Proceedings of Machine Learning Research*, pages 423–438, Hong Kong, 20–22 Nov 2016. PMLR.

Appendix A

Additional Results for Cross-domain Learning

This chapter presents some additional information for Chapter 5.

Table A.1: Hyper-parameter settings for Random Forest

Random Forest							
Season	**Model**	**no. of estimators**	**max depth**	**max features**	**min sample leafs**	**min sample splits**	**random state**
1	1	300	**None**	6	10	10	100
	2	300	6	-	-	-	100
	3	300	**None**	10	10	10	100
	4	300	**None**	6	10	10	100
	5	300	**None**	6	10	10	100
	6	300	**None**	6	10	10	100
2	1	300	**None**	6	10	10	100
	2	300	6	-	-	-	100
	3	300	**None**	10	10	10	100
	4	300	**None**	6	10	10	100
	5	300	**None**	6	10	10	100
	6	300	**None**	6	10	10	100
3	1	300	**None**	6	10	10	100
	2	300	6	-	-	-	100
	3	300	**None**	10	10	10	100
	4	300	**None**	6	5	6	100
	5	300	**None**	6	10	10	100
	6	300	**None**	6	10	10	100
4	1	300	**None**	6	10	10	100
	2	300	6	-	-	-	100
	3	300	**None**	10	10	10	100
	4	300	**None**	6	10	10	100
	5	300	**None**	6	10	10	100
	6	300	**None**	6	10	10	100

Table A.2: Hyper-parameter settings for Gradient Boosting

Gradient Boosting				
Season	**Model**	**no. of estimators**	**max depth**	**random state**
1	1	300	4	100
	2	300	3	100
	3	400	3	100
	4	300	4	100
	5	300	4	100
	6	400	3	100
2	1	300	4	100
	2	300	3	100
	3	400	3	100
	4	300	4	100
	5	300	4	100
	6	400	3	100
3	1	300	4	100
	2	300	3	100
	3	400	3	100
	4	300	4	100
	5	300	4	100
	6	400	3	100
4	1	300	4	100
	2	300	3	100
	3	200	5	100
	4	300	4	100
	5	400	3	100
	6	400	3	100