applied sciences

MDPI

Article

Spatiotemporal Analysis of Web News Archives for Crime Prediction

Areeba Umair [1], Muhammad Shahzad Sarfraz [1,*], Muhammad Ahmad [1,*], Usman Habib [1], Muhammad Habib Ullah [2] and Manuel Mazzara [3]

1 Department of Computer, National University of Computer and Emerging Sciences, Islamabad, CFD Campus, Chiniot, Punjab 35400, Pakistan; f180802@nu.edu.pk (A.U.); usman.habib@nu.edu.pk (U.H.)
2 Department of Electrical Engineering and Information Technology, University of Naples Federico II, 80125 Naples, Italy; Muhammad.habibullah@unina.it or habib.wattoo@nu.edu.pk
3 Institute of Software Development and Engineering, Innopolis University, 420500 Innopolis, Russia; m.mazzara@innopolis.ru
* Correspondence: shahzad.sarfraz@nu.edu.pk (M.S.S.); mahmad00@gmail.com (M.A.)

Received: 2 November 2020; Accepted: 14 November 2020; Published: 20 November 2020

Abstract: In today's world, security is the most prominent aspect which has been given higher priority. Despite the rapid growth and usage of digital devices, lucrative measurement of crimes in under-developing countries is still challenging. In this work, unstructural crime data (900 records) from the news archives of the previous eight years were extracted to predict the behavior of criminals' networks and transform it into useful information using natural language processing (NLP). To estimate the next move of criminals in Pakistan, we performed hotspot-based spatial analysis. Later, this information is fed to two different classifiers for possible identification and prediction. We achieved the maximum accuracy of 92% using K-Nearest Neighbor (KNN) and 62% using the Random Forest algorithm. In terms of crimes, the results showed that the most prevalent crime events are robberies. Thus, the usage of digital information archives, spatial analysis, and machine learning techniques can open new ways of handling a peaceful and sustainable society in eradicating crimes for countries having paucity of financial resources.

Keywords: crime prediction; spatial analysis; sustainable e-governance; sustainable information management

1. Introduction

Crimes are the most common social issues nowadays, affecting the economic growth, quality of life, and economy of any country. Crimes affect the reputation of a country on an international scale and affect the economy of the country by placing a financial burden on the government in hiring additional police forces. For the eradication of crimes, the government needs to adopt some optimized strategy [1] and sustainable e-governance information systems. Algorithms that predict the occurrence of crimes based on time and location can help the government to deploy law enforcement in highly dangerous areas [2].

Internet-based news resources, such as online newspapers and news channel archives, have been tremendously increased in number, volume, and coverage, and they contain useful as well as authentic data [3]. Nevertheless, the data of the archives are not so arranged and categorized, so it can be quite challenging to extract useful information about specific or interesting crime events [4–6]. According to the Pakistan Bureau of Statistics, the crime rate of Pakistan is increasing constantly, and among all the crimes, the rate of murder, kidnapping, robbery, accidents, and blasts is high. News archives provide a valuable source of information. It contains rich and purposeful content which is recorded carefully by specialists and it portrays some principal aspects related to the specific article [7]. The most popular

and authentic newspaper's archives of Pakistan are Dawn News, Dunya News, Ary News, The News, Daily Times, Pakistan Press Foundation, The Nations, and Journalism Pakistan [8]. The purpose of this research work is to utilize free of cost data available in news archives and perform the spatiotemporal analysis for crime prediction. NLP is an efficient mechanism to extract the keywords as representative of the whole text of the news body and researchers have used different Natural Language Processing (NLP) techniques for mining the data of news web archives [7].

Similarly, geostatistical-based approaches have been used by different researchers to identify the high-risk regions [1,2,9]. The development in Geographical Information Systems (GIS) approaches has enabled the analysis of spatial data in different domains. GIS-based approaches provide the visualization and exploration of incidences by creating map layers as visualization of spatial data which can help detect the patterns and trends of criminal networks. Hence, the spatial distribution of crime data for the prediction of future crime events using data mining and machine learning on the spatial dataset can provide accurate distribution of crimes [10]. These types of novel methods for crime mapping can be helpful in many aspects of society, such as decreasing the probability of accidents, crime ratio, and murder cases. Moreover, it can secure the nation from blasts, kidnappers, and murders [11]. This study focuses on information retrieval from the news archives, extraction of attributes from the news headlines, and applying spatial analysis as well as machine learning to predict future crimes.

Crime-solving is a complex task that requires human efforts and intelligence for the processing of criminal data. Therefore, data mining can assist researchers in crime identification problems [12–14]. Researchers have done extensive research on the usage of data mining and machine learning techniques in the identification and prediction of crime events and criminal networks [15]. There are many data mining and machine learning tools available for researchers. Weka is one of the tools which can assist researchers in mining the data and applying certain machine learning algorithms [16]. It has the capabilities of performing preprocessing, feature selection, clustering, and classification on data [17]. In this study, the algorithm of KNN predicted the crime type with 92% accuracy.

In today's world, security is the most promising aspect which should be provided by the government to their citizens. The principal objective of crime mapping is to estimate the probability and ratio of any mishap happening in the country. The objectives of this study include:

- To predict the crime patterns through news archives data and extract the crime information from the news text using freely available tools for the developing and under-developed countries that have a paucity of resources; considering Pakistan as an example.
- To help law enforcement agencies, in anticipation of the crime rate by analyzing the spatial distribution trends promptly.
- To predict the behavior of criminal networks by estimating their next move using machine learning algorithms.

In a nutshell, this study presents the feasibility to apply geospatial methods and machine learning approaches in order to predict the crimes/criminal activities using the eight years of data available in web archives.

The rest of the paper is structured as follows. Section 2 presents the state-of-the-art methods proposed in the literature for the targeted problem. Section 3 represents the overall methodology of our contribution to crime prediction. Spatial Crime analysis is performed in Section 4 while crime prediction is performed in Section 5. We perform evaluations in Section 6. The results are discussed in Sections 7 and 8 concludes the study.

2. Related Work

With the advancement in technology, criminal behavior is becoming more and more channeled and complex [4]. For crime control, the nature of crime must be understood [18]. The spatial analysis helps to decode the spatial behavior of criminal activities [18] and assist law enforcement in making predictions about future crimes that may occur [1].

Many crime prediction approaches have been proposed earlier by different researchers. Agarwal et al. proposed the framework of crime prediction in which crime analysis is performed on crime datasets by k-means clustering using a rapid miner tool [4]. However, there is a need to apply machine learning. Kiani et al. proposed a new framework for clustering and crime prediction in which they used the Genetic Algorithm (GA) for the detection of outliers. Their main focus was to classify the crime cases based on the frequency of crime occurrence during different years [19]. Reddya et al. used the tools of R such as Rgoogle maps [1], googleVis [1], ggplot2 [1], and ggmap for visualization of criminal data. They used the K-Nearest Neighbor (KNN) algorithms and Naïve Bayes algorithms to help the prediction of crimes [1], however, they can use more advanced methods of machine learning and apply spatial analysis as well. I. Matijosaitiene et al. proposed the method of crime prediction using land-use data with the help of machine learning algorithms. They identified the exact hours of crime occurring using hotspot analysis by using logistic regression and determining the precise time of the next crime [2], but the prediction results can be enhanced using advanced methods of machine learning.

Malathi et al. proposed the model of crime prediction using data mining techniques. The model consisted of data cleaning, data clustering, classification, and outlier detection [12]. In [20], Ivan et al. used GIS to visualize the spatial distribution of accidents along with the road networks. They identified the spatial patterns of road-side accidents along with its occurrence in different moments such as hours, days, seasons, and years, etc. Thakali et al. used kernel density estimation and kriging for identifying the hotspot of crashing incidents and estimating the collision frequency, respectively [9]. Haan used kernel density for estimation of concentration at a given point in space [21]. Xue et al. proposed the method of spatial analysis with a latent decision for crime prediction. They designed two different spatial models for crime prediction such as uniform and distinct. Both models helped in understanding the spatial pattern of crimes and criminal behaviors [11]. In [22] the authors design a deep neural network for the crime prediction by utilizing the New York crime dataset while in [23] Duan et al. predicted the crime suspect location using spatiotemporal analysis. Hu et al. designed a Bayesian model for urban crime prediction based on regional statistics [24].

Pflueger used a random forest algorithm to predict criminal activities by offenders having a mental illness. This approach can be helpful not only for the judiciary but also for designing new strategies for risk management [25]. Almanie et al. found crime patterns using decision trees and Naïve Bayesian classifiers. They predicted the future crime events in a particular location (latitude, longitude) within a specific time interval. They combined demographic information with the findings of the crime dataset of cities and then estimated which factor is affecting the neighbors the most [26]. The crime hotspot and spatial analysis can help to identify the spatial crime patterns. Jangra et al. compared the prediction rate and accuracy of KNN with the Naïve Bayes over the crime dataset. They used the previous scenarios of KNN over crime prediction and compared with their proposed scenario of Naïve Bayes and found out that both the techniques showed different accuracy rates. Jangra et al. reported that the accuracy of Naïve Bayes is higher than the KNN algorithm. Moreover, they emphasized that such types of techniques in combination with spatial datasets can predict crime-related data in an efficient manner [10]. Table 1 gives a summary of the related work.

From the literature, it has been determined that several approaches to GIS have been proposed to identify crime patterns and trends. However, there is a lack of research that predicts location-based crimes in adjoining areas of Pakistan utilizing free of cost data available in news archives. The freely available data can be transformed into useful information using natural language processing algorithms and prediction can be performed using supervised and unsupervised learning. Such kind of research

can help identify future crimes cases in the developing and under-developed countries having a paucity of financial resources.

Table 1. Summery of the state-of-the-art works.

References	Dataset	Techniques	Outcome
[1]	Official site of the U.K Police	KNN and Naïve Bayes	Prediction of crimes
[2]	Crime and land use data	Logistic regression	Exact hours of crime occurring
[4]	Local crime dataset	K-means	Clusters
[9]	Historical crash data of USA	Kernel density estimation and Kriging	Identify hotspot in roads network
[10]	Crime data on India	KNN and Naïve Bayes	Prediction of crime type
[11]	Regional Crime Analysis Program (ReCAP)	Hotspot	Criminal event prediction
[12]	Surveys	KNN	Outlier detection
[19]	Real crime dataset recorded by police	Genetic Algorithm (GA)	Detection of crimes
[25]	Mentally ill offenders' data	Random Forest	forensic-psychiatric risk-assessment
[26]	Denver crimes dataset	Decision Tree and Naïve Bayes	Identify crime patterns

3. Methodology

The crime mapping with the help of spatial analysis uses data from news archives; the news records are then processed with the help of various Python text processing modules to extract the valuable information from the text of news records. Spatial characteristics of crimes such as dispersed, clustered, or random will be extracted for analyzing the distributions of crimes [12]. Various GIS visualization techniques are used to give a better visualization of crimes.

3.1. Data Collection and Data Processing

Data has crawled through the news archives of almost all popular and authentic newspapers mainly from Dawn News, Dunya News, Ary News, The News, Daily Pakistan, Pakistan Press Foundation, The Nation, and Journalism Pakistan, with the help of a Data Miner Tool.

The data miner tool was selected for crawling because it gathers data from a specific website and represents it in tabular form. It classifies the news further in the title, description, date, URL, etc. The attributes of news records selected for this study were title, description, date, URL, location, latitude, longitude, type of crime, mishap, and the number of people affected. The archives of news data from the years 2011–2019 were mined for the development of the research model. In total, 920 records related to crimes from archives were extracted and were screened for duplication removal. Out of 920, it was found that twenty records were extracted multiple times and these were removed. The data consists of 900 records, and the number of records of each specific crime type is shown in Table 2. According to Table 2, the records of robbery are highest among all, i.e., 320 after that accident has a 230 number of records. Then blast has 150, Kidnapping has 90, Murder has 55, a shot has 20, suicide has 20, and arrest has 15 number of records. The data were further processed using various modules of Python such as Natural Language Processing Toolkit (NLTK), regular expression (RE), pandas, geo-py, etc. for data normalization.

Table 2. Number of records of each crime category and their evaluation accuracy.

Crime Type	No. of Records	Precision	Recall
Robbery	320	74.32	69.88
Accident	230	67.48	55.29
Blast	150	52.09	45.68
Kidnapping	90	49.11	42.22
Murder	55	43.18	47.44
Shot	20	44.20	56.43
Suicide	20	43.33	55.21
Arrest	15	41.93	61.89

3.1.1. Punctuation Removal

Many Python modules do not process text if it contains punctuation. Considering this, as a first step, punctuation from the title and description of news records was removed, so that it becomes easy for the algorithms to work over it.

3.1.2. Extraction of Attributes

We used modules of NLP to extract the attributes from the news articles. NLP enables the computer to understand the human language and derive the meaning from the long paragraphs of text. NLP helps in finding entities and sentiments in the sentences. We have used NLTK of Python for NLP. The NLTK is the suite of programs and built-in libraries for natural language processing (NLP) symbolically and statistically [27]. NLTK 3.5 module of Python is used to extract the location from the text of news records. The most common algorithms of natural language processing such as sentiment analysis, tokenizing, topic segmentation, stemming, and part-of-speech tagging, named entity recognition can be handled with NLTK. NLTK analyzes, pre-processes, and understands the written text and helps computer in interpretation of the text. NLTK was selected for this study because it uses named-entity recognition, which uses predefined categories for information extraction. Most of the studies used the N-gram approach for the extraction of information from the text [28], but the accuracy of Regular Expression (RE) 2019.05.25 module is found more than the N-gram approach. RE modules of Python are used to extract the crime type, i.e., robbery, murder, blast, etc. and to extract the information about mishaps. RE module has been used for this activity because it is powerful and provides the best results in parsing text.

3.1.3. Geo-Coding

For spatial data analysis, geocoding plays a vital role. It assigns the respective latitude and longitude to the location for better visualization and better prediction of crime patterns. There exist many techniques of geo-coding in Python such as geocoder, geo-py, etc. We used the geo-py 2.0.0 module along with the Pandas module of Python for geo-coding because it is the best tool to deal with columns and rows while geo-coding [29]. The flowchart given below shows the whole information extraction process of news in Figure 1.

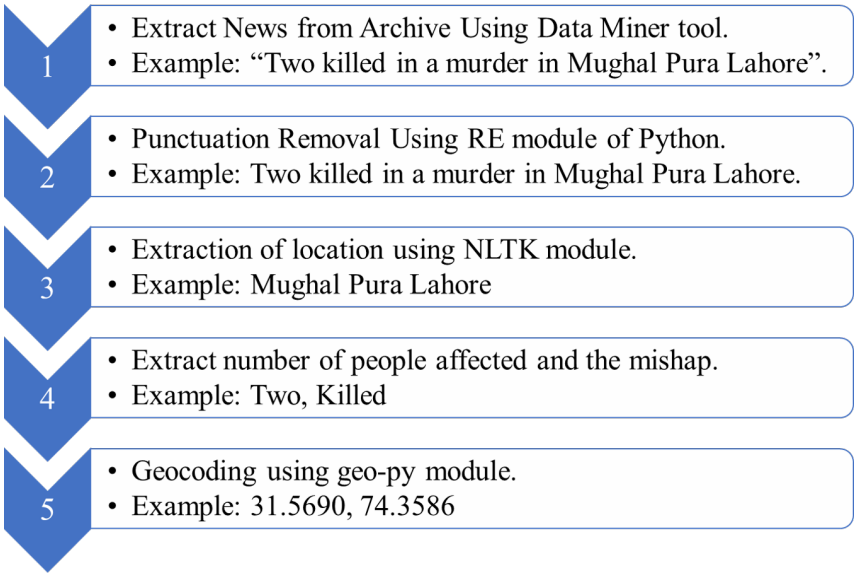

Figure 1. Flowchart of information extraction process with example.

3.1.4. Evaluation of Information Extraction

We have used precision and recall to evaluate the accuracy of crime information extraction, especially the crime event as shown in Table 2.

3.1.5. Attributes of Data

The total attributes of data are given in Table 3 below which shows that total attributes of the data are title, description, date, location, crime type, number of people affected, latitude, and longitude along with their description, data type, and examples.

Table 3. Attributes of data.

Attribute	Data Type	Description	Examples
Title	String	Title of the news	Lahore: 3 killed in blast
Description	String	Description of the news	A wounded man receives treatment...
Date	Date	Date of the event	3 October 2019
Location	String	Location of the mishap	Madina town, Faisalabad
Crime Type	String	Type of the crime	Murder
Number of people affected	Integer	Count of people affected in mishap	3 Killed
Lat	Float	Latitude	34.33464
Long	Float	Longitude	72.46536

3.1.6. Preprocessing

Initially, all the instances were integrated, and attributes were extracted from the description. Later on, data cleaning was performed and the attributes of URL and description were removed as these fields were not required for prediction. Attributes were selected based on extensive literature and there was no need to reduce the parameter, so keeping in view data reduction techniques were not applied. Preprocessed data attributes were in the acceptable format as per the input of KNN and Naïve Bayes algorithms. Keeping this in view, there was no need to do extra data transformations and discretization steps. We have applied one-hot encoding in order to convert the categorical attributes into the numeric form.

3.2. Geo-Spatial Data Mapping

Data visualization is the visual art that is used to represent data and information graphically. It is used to analyze the trends and patterns in data [30]. In this paper, data visualization has been used to analyze the crime with the help of the map and to predict the crime rate in the future. ArcGIS modules and tools were used to display information on the map.

3.2.1. Visualization of Crime Data Using Arcgis

ArcGIS is a tool which is widely used to visualize spatial datasets. In this research, the shapefile of Pakistan is loaded along with the extracted dataset in the ArcGIS to show crime records based on latitude and longitude. Figure 2a is the representation of the dataset on ArcGIS and it indicates that Punjab has been highly susceptible to crimes since the last decade. Moreover, Figure 2b is built using ArcMap software and shows the distribution of crimes based on its categories. Through such types of geographical representation, the areas or locations which may be susceptible to crime in the future can be easily identified.

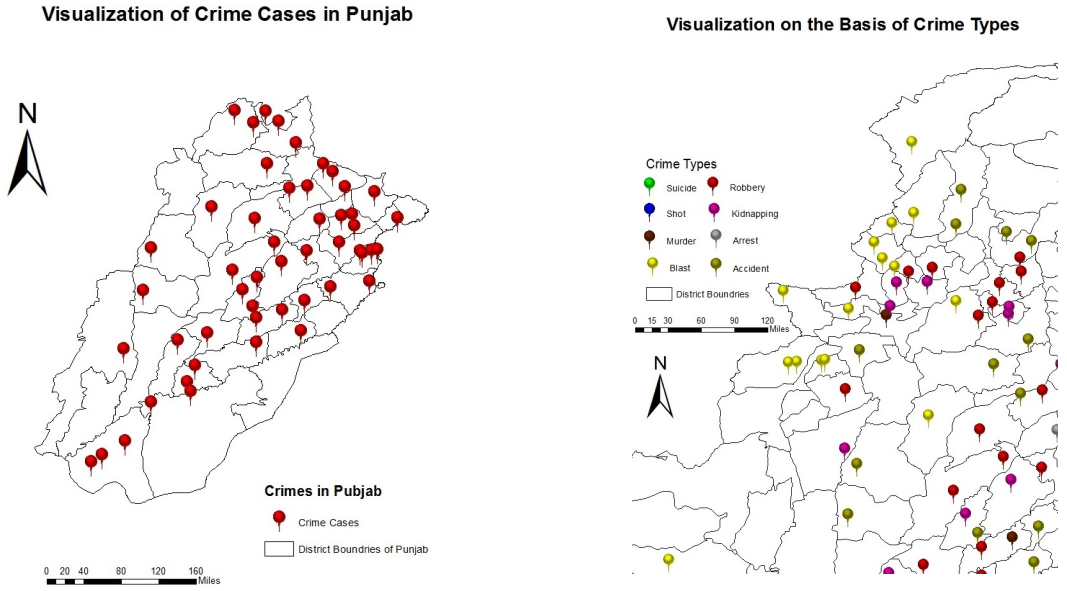

Figure 2. Visualization of (**a**) crime cases and (**b**) crime types during last eight years using ArcGIS.

3.2.2. Visualization By Creating Hotspots

Hotspot shows the areas of high crime rate, where the possibility of crime to occur is very high, as shown in Figure 3. It means the concentration of crime in a particular area [31,32] by count which finds the hot and cold spots in each aggregation area across the study area. Creating hotspots on maps helps law enforcement agencies to indicate the area of high crime rate, consequently predicting the reasons for crime in that specified area and prevention of further crime by high alerting the security requirements in that area [2,33,34]. We used an optimized hotspot analysis tool with the IDW tool in ArcGIS to get the results. The results show that the areas of Karachi and Hyderabad are more prone to crime in the future as shown in Figure 3. IDW is usually used with hotspots to identify the affected and unaffected patterns more clearly [35]. In Figure 3, the red portion represents the area of high crime rate whereas the blue portion represents the areas of the low crime rate. Hence, according to a hotspot, the crime rate is high in the areas of Karachi and Hyderabad.

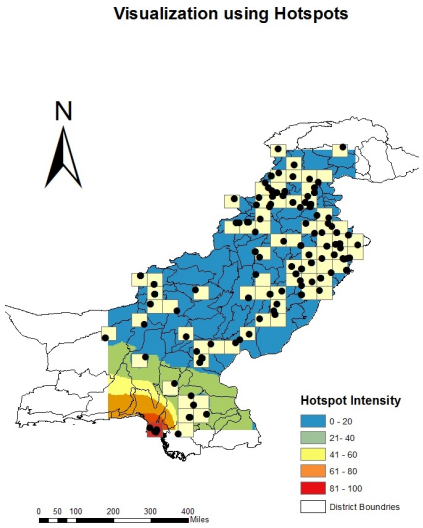

Figure 3. Visualization of crime cases on the basis of each crime type using ArcGIS.

4. Spatial Crime Analysis

Crime analysis is defined as the analytical process that identifies crime patterns and trends related to crime data, which assist in deploying strategies and planning for future crime prediction [4]. We have performed the spatial crime analysis using the spatial data we extracted from the web news archives in order to investigate the trends of crime geographically. Spatial crime analysis means to study the spatial distribution of the crime rate that either the crime features are clustered, random, or dispersed. It shows the spatial correlation between the features points of crimes and identifies the trends among the crime patterns. Spatial crime analysis involves a collection of statistical techniques to discover spatial patterns, spatial clusters, and spatial trends in criminal data. Researchers declared that crime is not a random activity; instead, it is spatially concentrated in most of the cases [12]. The objectives of spatial analysis are to identify the relocation patterns of the criminals. We can estimate the next move of the criminals with the help of various geospatial methods such as hotspot analysis etc.

It is necessary to know how crime data is spatially distributed. To investigate this, we identified the relationship between crime features using the average nearest neighbor using the spatial dataset of crime that we extracted from the news archives. Cluster analysis is also used to study the distribution of crimes. We have used k-means [36,37] clustering for cluster analysis over the spatial data of crime. Clusters are formed in a region where there is a greater tendency of the crime rate. Pattern analysis also gives the spatial interaction between the locations, which is used in the estimation of heterogeneity and dependence of crime over other factors [12].

4.1. Analysis Using Average Nearest Neighbor

According to the Routine Activity Theory, the behavior patterns of people and their environment has a significant impact on criminal activities. Therefore, to identify and explain the relationship between neighborhood and crime characteristics is a key aspect [12]. It is necessary to know how the crime data is spatially distributed, i.e., either clustered, random, or dispersed. We have used the average nearest neighbor, a statistical tool in ArcGIS, to measure the autocorrelation between crime features in our spatial crime dataset. Average nearest neighbor is the tool that measures the distance from the center of each point to the centers of its neighbors. Further, it calculates the average of all the calculated nearest distances. The average distance is compared with the average of the hypothetically which gives a random distribution using the formula given in Equation (1). The average nearest neighbor is calculated as the ratio between the observed distances of each feature to the expected distance.

$$average\ nearest\ neighbor = DO/DE \quad (1)$$

In average nearest neighbor, if the value of the nearest neighbor ratio is less than one, it indicates that the patterns are clustered as shown in Figure 4. Figure 4 is obtained and built using the average nearest neighbor tool in ArcMap. In the case that the value is greater than one, it shows that the patterns are dispersed in the relationship. Hence, Figure 4 shows that in our data, the spatial correlation of feature points is clustered.

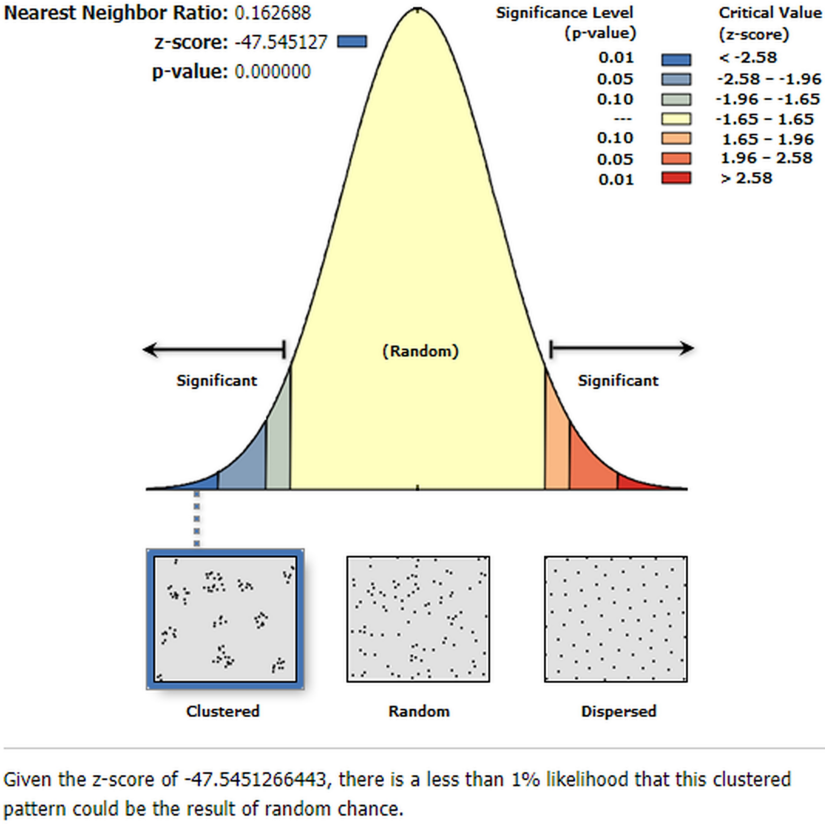

Figure 4. Visualization of crime cases on the basis of each crime type using ArcGIS.

4.2. Clustering Using K-Means Clustering

Clustering is the technique of data mining that groups the objects in sets of similar features or properties and each set differs from others in its behavior [4]. It can help in the prediction of crimes based on spatial distribution by analysis of the clusters [12]. In this study, we used the k-Means algorithm to perform clustering using a crime dataset because it is applicable over the large datasets and has less complexity as compared to other clustering algorithms [19]. The Weka tool is used in this study for performing k-Means clustering. In k-Means clustering, k clusters are formed from n observation based on the nearest mean. The process of k-Means clustering involves:

1. Declaring the number of clusters as *k*.
2. Choose the centers of each cluster.
3. Each instance is assigned to the cluster, which is the nearest.
4. The centroids of clusters are recalculated.
5. The process is iterated.

Table 4 shows the centroids of each cluster formed through the k-Means algorithm. The total data is divided into eight clusters numbered as 0 to 7. Table 5 shows the distribution of clusters based on the crime type. Clusters' names are assigned based on the centroid. Figure 5 has been built using the Weka tool and illustrates the clusters of crime with respect to their latitude. We just included the centroid of the central cluster in Table 4 which is obtained as a result of K-means clustering in Weka and helps to identify the ratio of different crimes among cities of Pakistan.

Table 4. Clusters of crimes in Pakistan on the basis of k-Means algorithm.

Clusters	Data Points	Location	LAT	LONG	Date
Centroid	900	Karachi	30.38	71.04	5 January 2019
Cluster 0	261	Islamabad	33.80	72.42	5 January 2019
Cluster 1	220	Lahore	31.47	73.90	25 February 2019
Cluster 2	45	Gujranawala	32.28	74.06	6 December 2018
Cluster 3	30	Hyderabad	26.35	68.57	1 December 2018
Cluster 4	65	Multan	30.32	71.53	5 February 2019
Cluster 5	63	Quetta	30.50	67.20	6 January 2019
Cluster 6	205	Karachi	25.02	67.17	9 February 2019
Cluster 7	11	Malakand	29.31	66.48	13 July 2018

Table 5. Description of our selected feature selection methods for comparison.

Cluster #	Crime Type
Cluster 0	Kidnapping
Cluster 1	Accident
Cluster 2	Arrest
Cluster 3	Shot
Cluster 4	Murder
Cluster 5	Blast
Cluster 6	Robbery
Cluster 7	Suicide

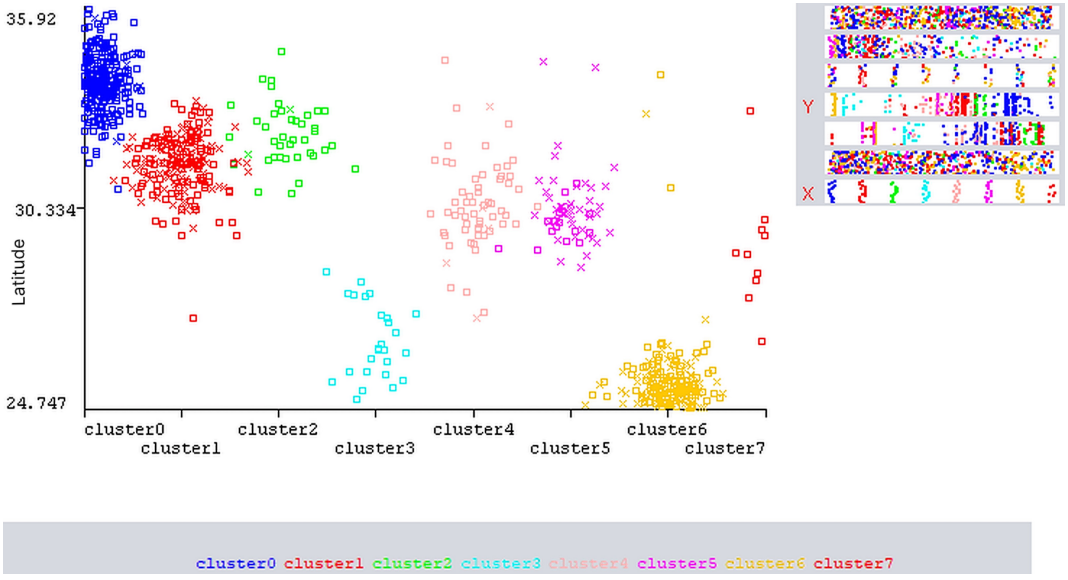

Figure 5. Clusters of crimes cases of Pakistan formed as a result of k-Means clustering.

5. Crime Prediction

Crime-solving is considered a complex task that requires human intelligence and experience. Researchers use data mining techniques over large crime datasets, which help to identify the hidden patterns present in a huge dataset. There are various tools and software for applying data mining and extracting useful information to facilitate our use of a massive amount of data. Srivastava et al. used the Weka tool for applying data mining algorithms over crime data for prediction purposes [17]. KNN and random forest classifiers are considered highly accurate for this kind of prediction. Inspired by the successful applications of the above-mentioned machine learning algorithms, we used these two algorithms in our research to predict the crime in Pakistan.

5.1. Supervised Learning

In supervised learning, several examples are required to train the model with the help of a training dataset. The labeled data is provided to the model, according to which the classifier trains itself [19]. It is necessary to train the machine learning model on the data similar to the target data [38–40]. The target data has some unlabeled values which can be predicted by machine learning models based on training datasets. In this study, we used 5-folds cross-validation in order to train and test our model. The class label of this research is a crime type.

We selected the KNN model for our research because it is a nonparametric algorithm and uses similarity matrices to compare the labels of test data with the training data. Each instance of data is considered as a record with n features. To predict the class labels of the test data, KNN selects those records of training data that are closest to the unlabeled records [41]. KNN uses a distance-based approach, which is beneficial to deal with the data having a clustered set of features. The random forest algorithm was preferred, as it is free from any types of parametric assumptions. Moreover, this algorithm is quite suitable for datasets that are nonlinear and have high-order complexity in nature [25]. As we are dealing with spatial data with slightly complex and nonlinear data types, it is quite suitable for such types of scenarios.

5.1.1. Prediction Using KNN

KNN is a model-free algorithm and gives an n training vector in dimensions of feature place. It identifies the k nearest neighbor of the feature vector that was being estimated using Equation (2). KNN works by looking at the history of past crimes and finds a similar crime based on the matched number of neighbors. Its output is class membership, which means the maximum votes identify an object from its neighborhood. Crime prediction can be made most efficiently using KNN because the neighbors of a victim house are more susceptible to the next theft. Therefore, the nearby areas of recent crime locations are considered more vulnerable to the next crime. Considering the similarity between the testing and training data, the distance was calculated to predict the classes of test data. KNN algorithm predicts the test data based on the nearest neighbor method [41]. In this study, KNN is trained using a crime dataset with the help of the Weka tool, and our generated test set was used to predict the event occurring at the specified location. KNN involves the factor of distance, so the distance between the training feature and test feature is computed using the formula in Equation (2):

$$d_i = \sqrt{(x_i - x)^2 + (y_i - y)^2} \tag{2}$$

After distance computation, the nearest neighbors are identified using sorting techniques and are assigned the crime type based on the voting of its neighbors. We can implement a KNN model by the following steps [42]:

1. Input the dataset.
2. The value of k is initialized.
3. Data points are iterated from 1 to the total number of training in order to get the predicted class.
4. Calculate the distance between testing data and each row of training data; considering Euclidean distance as the distance metric because it measures the distance between the pair of samples p and q in an n-dimensional feature space.
5. The values of calculated distances are sorted in ascending order.
6. Top k rows are returned from the sorted array.
7. The most commonly used class of selected rows is returned.
8. Prediction results are returned.

5.1.2. Prediction Using Random Forest

Random Forest is the most famous and powerful supervised machine learning algorithms. The algorithm creates several decision trees within a forest [43]. In general, the more trees in the forest, the more robust the prediction and high the prediction accuracy rate. It predicts the new class based on features of previous classes of old trees. When a new tree is introduced, each tree gives votes for the new tree; the forest chooses the classification of having the most votes of all the other trees in the forest [25]. We used the Weka tool for prediction of crime events by random forest on the crime dataset. Random forest algorithms can be implemented based on the following steps [44]:

1. It Randomly selects k features from total m features on random basis where $m \gg k$.
2. It calculates the node d, among the k features, using the method of best split point.
3. It Splits the node into daughter nodes using the method of best split.
4. It repeats steps 1 to 3 until, it reaches l number of nodes.
5. It builds forest by repeating steps 1 to 4 for n number of times to create n number of trees.
6. It takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target).
7. For each predicted class, the votes are calculated.
8. Final predictions are made based on the high voted predicted class using the random forest algorithm.

6. Evaluation

Different evaluation matrices are used to measure the performance of any algorithm. The performance criteria for evaluations are accuracy, precision, recall, F-measure, ROC curve, root mean square error, absolute error, etc. Accuracy is defined as the ability to predict categorical class labels. This means that it calculated the proportion of correctly predicted instances [17]. Accuracy measurements were done using the formula in Equation (3):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Precision is the measure of closeness of instances with each other and is calculated as per the formula given in Equation (4).

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

A recall is the measure of actual positive instances in the dataset that have been correctly classified as positive by the classifier. It is calculated using the formula given in Equation (5).

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

F-measure is calculated by taking the weighted harmonic mean of the precision and the recall as in Equation (6).

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

where

- False Negative (FN): If the record is positive in the dataset, but the classification outcome is negative;
- True Negative (TN): If the record is negative in the dataset but the classification outcome is negative;
- False Positive (FP): If the data record is negative, but the classification outcome is positive;
- True Positive (TP): If the instance is positive but the classification outcome is positive.

7. Results and Discussion

Crime prediction is one of the most challenging tasks, especially when data availability of criminal reports is not up to the mark [45,46]. Electronic media is one of the most powerful tools, which can provide accurate data and remains useful for the conduction of the research. Data mining tools helped in managing the data in an understandable format which led to meaningful information for answering crime patterns and their relationships.

In our prediction model, we used two machine learning algorithms for the prediction of crime events on the archive dataset. The results of these two algorithms were compared in terms of accuracy and prediction. The average accuracy of the KNN and Random Forest was observed as 92% and 62%, respectively. This indicates that the prediction of KNN remained high and efficient as compared to Random Forest. Tables 6 and 7 also show the results of both of the algorithms along with the parameters. Table 6 shows the different values of accuracy, precision, recall, and F-measure against different parameters of the KNN algorithm. The results show that the values of matrices are being high when the number of K is increasing and we got maximum values when K was equal to 9. Table 7 represents the number of trees as the parameters and the values of accuracy, precision, recall, and F-measure against them. We have achieved the higher values of matrices with a higher number of trees in the random forest algorithm.

Table 6. Results of KNN showing Accuracy, Precision, Recall, F-measure against different parameters of the algorithm obtained using Weka tool.

k	Precision	Recall	F-Measure (Macro-Averaged)	Accuracy	Micro-Averaged
3-NN	0.92	0.87	0.82	0.89	0.89
5-NN	0.91	0.85	0.84	0.93	0.89
7-NN	0.91	0.86	0.87	0.93	0.90
9-NN	0.91	0.85	0.89	0.94	0.90

Table 7. Results of Random Forest showing Accuracy, Precision, Recall, F-measure against different parameters of the algorithm obtained using Weka.

Trees	Precision	Recall	F-Measure (Macro-Averaged)	Accuracy	Micro-Averaged
10 Trees	0.54	0.52	0.47	0.60	0.50
20 Trees	0.55	0.49	0.51	0.55	0.50
30 Trees	0.60	0.57	0.52	0.59	0.52
40 Trees	0.62	0.55	0.57	0.60	0.55

KNN predicts the most accurate result because it can reduce the adverse effects caused by improper classification of features and reduce the errors of classification [5]. In this method, surrounding samples play their role to classify each sample. Therefore, considering the class of nearest neighbour samples, the class of unknown sample can be predicted. In the test and training datasets, distances between unknown samples of the test data and samples of training data were computed. The unknown sample of the test data has been assigned the value of the smallest distance corresponding to the sample in the training set [6]. The reason for getting high accuracy by KNN may be because this algorithm selects the features based on a distance between points, considering the points having nearby crimes occurring in the archived datasets may lead to higher accuracy.

An automated duplication removal process can increase the data extraction process. Similarly, usage of advanced machine learning such as reinforcement learning and deep learning algorithms may give better results. Moreover, automatic geo-coding methods for the extraction of precise locations can identify the exact location of the crime. Such type of integrated model will help decision-makers and law enforcement agencies predict the more precise location of crimes for getting fruitful results. As mentioned earlier, the challenge in this research was to extract data without any

cost or ground survey in an efficient manner. Such an automated process is quite useful for developing and under-developing countries where geospatial data is not being maintained or shared.

8. Conclusions

The usage of digital information archives is a cost-effective way of predicting crime events occurring in a country. Data related to crimes extracted through automated tools such as Python can be converted into useful information for the prediction of fruitful results. The location-based geo-coded data, processed through GIS-based software, i.e., ArcMap provided locations based statistical information, which helped in the identification of the patterns, trends, and relationships between crime features. Furthermore, the hotspots analysis assisted in identifying the areas and regions of high susceptibility. Such types of research can be quite helpful to law-enforcement agencies to monitor highly sensitive areas and to remain in high alert in terms of security. KNN and the Random Forest algorithm concluded that Pakistan has the worst condition in a robbery as compared to other crimes. Such types of a robust method can be an effective way to keep an eye on risk-prone areas. In conclusion, such types of automated processes can open new ways of handling a peaceful and sustainable society in eradicating crimes for the developing and under-developed countries having a paucity of financial resources.

Due to limitations of time, availability of data, and lack of resources, we were only be able to extract limited datasets, i.e., 900 crime records at the city level. There is a possibility of uncertainty in the number of crime cases because the data has been extracted from particular news archives. Adding other electronic resources, mainly from the local language, can increase the accuracy of the dataset. Moreover, the unpredictability and uncertainty in the crime rate is still a challenge for researchers and decision-makers. This is because various other factors affect the crime rate simultaneously such as criminal mental state, poverty, low income, unemployment, illiteracy, family pressure, bad company, etc. [47,48]. By adding socioeconomic data, precise locations of crimes, and data from other electronic resources, a useful prediction model can be developed. In addition to that, the demographic data (population density) of Pakistan can help us to improve the crime prediction. It can show how the population distribution is associated with the crime rate in Pakistan. Similarly, some other potential biases such as information bias can produce more fruitful results for crime prediction.

Author Contributions: Conceptualization, A.U., M.S.S., U.H., and M.H.U.; Data curation, A.U., M.S.S., and M.H.U.; Formal analysis, A.U., M.S.S., and M.M.; Funding acquisition, M.A. and M.M.; Investigation, A.U., M.A., M.H.U., and M.M.; Methodology, A.U.; Supervision, M.S.S., M.A., and U.H.; Validation, A.U., M.S.S., M.A., and U.H.; Visualization, A.U., M.S., and U.H.; Writing—original draft, A.U., M.S.S., M.A., U.H., M.H.U., and M.M.; Writing—review and editing, A.U., M.S.S., M.A., U.H., M.H.U., and M.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received funding from a research project under national research program for universities, No: 10537/Federal/NRPU/R&D/HEC/2017. This research was partially supported by Innopolis University, Innopolis, Russia.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Toppireddy, H.K.R.; Saini, B.; Mahajan, G. Crime Prediction & Monitoring Framework Based on Spatial Analysis. *Procedia Comput. Sci.* **2018**, *132*, 696–705, [CrossRef]
2. Matijosaitiene, I.; Zhao, P.; Jaume, S.; Gilkey, J.W. Prediction of hourly effect of land use on crime. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 16, [CrossRef]
3. Ristea, A.; Al Boni, M.; Resch, B.; Gerber, M.S.; Leitner, M. Spatial crime distribution and prediction for sporting events using social media. *Int. J. Geogr. Inf. Sci.* **2020**, *34*, 1708–1739, [CrossRef] [PubMed]
4. Agarwal, J.; Nagpal, R.; Sehgal, R. Crime Analysis using K-Means Clustering. *Int. J. Comput. Appl.* **2013**, *83*, 1–4, [CrossRef]
5. Sun, C.C.; Yao, C.L.; Li, X.; Lee, K. Detecting crime types using classification algorithms. *J. Digit. Inf. Manag.* **2014**, *12*, 321–327.

6. Imandoust, S.B.; Bolandraftar, M. Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background. *Int. J. Eng. Res. Appl.* **2013**, *3*, 605–610.
7. Za'in, C.; Pratama, M.; Lughofer, E.; Anavatti, S.G. Evolving type-2 web news mining. *Appl. Soft Comput. J.* **2017**, *54*, 200–220, [CrossRef]
8. Yousaf, Z.; Yasmeen, G.; Ali, E. Sensationalizing the News Events by Pakistani Media. *J. Media Stud.* **2019**, *34*, 53–75.
9. Thakali, L.; Kwon, T.J.; Fu, L. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. *J. Mod. Transp.* **2015**, *23*, 93–106, [CrossRef]
10. Jangra, M.; Kalsi, S. Naïve Bayes Approach for the Crime Prediction in Data Mining. *Int. J. Comput. Appl.* **2019**, *178*, 33–37, [CrossRef]
11. Xue, Y.; Brown, D.E. Spatial analysis with preference specification of latent decision makers for criminal event prediction. *Decis. Support Syst.* **2006**, *41*, 560–573, [CrossRef]
12. Malathi, A.; Baboo, S.S. Enhanced Algorithms to Identify Change in Crime Patterns. *Int. J. Comb. Optim. Probl. Inform.* **2011**, *2*, 32–38.
13. Brayne, S.; Christin, A. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Soc. Probl.* **2020**, 1–17, [CrossRef]
14. Manzanares, M.C.S.; Diez, J.J.R.; Sánchez, R.M.; Yáñez, M.J.Z.; Menéndez, R.C. Lifelong learning from sustainable education: An analysis with eye tracking and data mining techniques. *Sustainability* **2020**, *12*, 1970, [CrossRef]
15. Kotevska, O.; Kusne, A.G.; Samarov, D.V.; Lbath, A.; Battou, A. Dynamic Network Model for Smart City Data-Loss Resilience Case Study: City-to-City Network for Crime Analytics. *IEEE Access* **2017**, *5*, 20524–20535, [CrossRef] [PubMed]
16. Witten, I.H.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. *SIGMOD Record* **1999**, *31*, 371. [CrossRef]
17. Srivastava, S. Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. *Int. J. Comput. Appl.* **2014**, *88*, 26–29, [CrossRef]
18. Ingilevich, V.; Ivanov, S. Crime rate prediction in the urban environment using social factors. *Procedia Comput. Sci.* **2018**, *136*, 472–478, [CrossRef]
19. Kiani, R.; Mahdavi, S.; Keshavarzi, A. Analysis and Prediction of Crimes by Clustering and Classification. *Int. J. Adv. Res. Artif. Intell.* **2015**, *4*, 11–17, [CrossRef]
20. Ivan, K.; Haidu, I. The spatio-temporal distribution of road accidents in Cluj-Napoca. *Geogr. Tech.* **2012**, *7*, 32–38.
21. de Haan, P. On the use of density kernels for concentration estimations within particle and puff dispersion models. *Atmos. Environ.* **1999**, *33*, 2007–2021, [CrossRef]
22. Duan, L.; Ye, X.; Hu, T.; Zhu, X. Prediction of suspect location based on spatiotemporal semantics. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 185. [CrossRef]
23. Duan, L.; Hu, T.; Cheng, E.; Zhu, J.; Gao, C. Deep Convolutional Neural Networks for Spatiotemporal Crime Prediction. In Proceedings of the International Conference on Information and Knowledge Engineering (IKE), Las Vegas, NV, USA, 17–20 July 2017; pp. 61–67.
24. Hu, T.; Zhu, X.; Duan, L.; Guo, W. Urban crime prediction based on spatiotemporal Bayesian model. *PLoS ONE* **2018**, *13*, 1–18, [CrossRef] [PubMed]
25. Pflueger, M.O.; Franke, I.; Graf, M.; Hachtel, H. Predicting general criminal recidivism in mentally disordered offenders using a random forest approach. *BMC Psychiatry* **2015**, *15*, 1–10, [CrossRef]
26. Almanie, T.; Mirza, R.; Lor, E. Crime Prediction Based on Crime Types and Using Spatial and Temporal Criminal Hotspots. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1–19, [CrossRef]
27. Loper, E.; Bird, S. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*; Association for Computational Linguistics: Philadelphia, PA, USA, 2002.
28. Gervás, P. Constrained creation of poetic forms during theme-driven exploration of a domain defined by an N-gram model. *Conn. Sci.* **2016**, *28*, 111–130, [CrossRef]
29. Bosch, M. Swisslandstats-geopy: Python tools for the land statistics datasets from the Swiss Federal Statistical Office. *J. Open Source Softw.* **2019**, *4*, 1511–1515, [CrossRef]

30. Carrabine, E. Just images: Aesthetics, ethics and visual criminology. *Br. J. Criminol.* **2012**, *52*, 463–489, [CrossRef]
31. Butt, A.; Ahmad, S.S.; Shabbir, R.; Erum, S. GIS based surveillance of road traffic accidents (RTA) risk for Rawalpindi city: A geo-statistical approach. *Kuwait J. Sci.* **2017**, *44*, 129–134.
32. Jana, M.; Sar, N. Modeling of hotspot detection using cluster outlier analysis and Getis-Ord Gi* statistic of educational development in upper-primary level, India. *Model. Earth Syst. Environ.* **2016**, *2*, 1–10, [CrossRef]
33. Tabangin, D.R.; Flores, J.C.; Emperador, N.F. Investigating Crime Hotspot Places and their Implication to Urban Environmental Design: A Geographic Visualization and Data Mining Approach. *Int. J. Hum. Soc. Sci.* **2008**, *2*, 4004–4012.
34. Russnák, J.; Ondrejka, P.; Herman, L.; Kubíček, P.; Mertel, A. Visualization and spatial analysis of police open data as a part of community policing in the city of Pardubice (Czech Republic). *Ann. GIS* **2016**, *22*, 187–201, [CrossRef]
35. Majumder, R.; Bhunia, G.S.; Patra, P.; Mandal, A.C.; Ghosh, D.; Shit, P.K. Assessment of flood hotspot at a village level using GIS-based spatial statistical techniques. *Arab. J. Geosci.* **2019**, *12*, 409, [CrossRef]
36. Ahmad, M.; Haq, I.; Mushtaq, Q.; Sohaib, M. A New Statistical Approach for Band Clustering and Band Selection Using K-Means Clustering. *Int. J. Eng. Technol.* **2011**, *3*, 606–614.
37. Ahmad, M.; Mazzara, M.; Raza, R.A.; Distefano, S.; Asif, M.; Sarfraz, M.S.; Khan, A.M.; Sohaib, A. Multiclass Non-Randomized Spectral–Spatial Active Learning for Hyperspectral Image Classification. *Appl. Sci.* **2020**, *10*, 4739. [CrossRef]
38. Wei, Q.; Dunbrack, R.L. The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics. *PLoS ONE* **2013**, *8*, 1–12, [CrossRef] [PubMed]
39. Ceylan, H.; Parlakyildiz, S. An approach to estimate occupational accidents using least-squares support vector machines. *Kuwait J. Sci.* **2017**, *44*, 83–91.
40. Wang, S.; Minku, L.L.; Chawla, N.; Yao, X. Learning from data streams and class imbalance. *Conn. Sci.* **2019**, *31*, 103–104, [CrossRef]
41. Shatnawi, M.K.A. Stock Price Prediction Using K -Nearest Neighbor (k NN) Algorithm. *Int. J. Bus. Humanit. Technol.* **2013**, *3*, 32–44.
42. He, Q. P.; Wang, J. Fault detection using random projections and k-nearest neighbor rule for semiconductor manufacturing processes. *IEEE Trans. Semicond. Manuf.* **2007**, *20*, 345–354, [CrossRef]
43. Ngarambe, J.; Irakoze, A.; Yun, G.Y.; Kim, G. Comparative performance of machine learning algorithms in the prediction of indoor daylight illuminances. *Sustainability* **2020**, *12*, 4471, [CrossRef]
44. Zakariah, M. Classification of large datasets using Random Forest Algorithm in various applications: Survey. *Int. J. Eng. Innov. Technol.* **2014**, *4*, 189–198.
45. Lim, M.; Abdullah, A.; Jhanjhi, N.; Khurram Khan, M.; Supramaniam, M. Link prediction in time-evolving criminal network with deep reinforcement learning technique. *IEEE Access* **2019**, *7*, 184797–184807, [CrossRef]
46. Pan, C.; Li, B.; Wang, C.; Zhang, Y.; Geldner, N.; Wang, L.; Bertozzi, A.L. Crime modeling with truncated Lévy flights for residential burglary models. *Math. Model. Methods Appl. Sci.* **2018**, *28*, 1857–1880, [CrossRef]
47. Dollar, C.B.; Donnelly, E.A.; Parker, K.F. Joblessness, Poverty, and Neighborhood Crime: Testing Wilson's Assertions of Jobless Poverty. *Soc. Curr.* **2019**, *6*, 343–360, [CrossRef]
48. Tang, Y.; Zhu, X.; Guo, W.; Wu, L.; Fan, Y. Anisotropic diffusion for improved crime prediction in urban China. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 234, [CrossRef]