

Prediction of Extremist Behaviour and Suicide Bombing from Terrorism Contents Using Supervised Learning

Nasir Mahmood* and Muhammad Usman Ghani Khan

Department of Computer Science, University of Engineering and Technology, 54890, Lahore

*Corresponding Author: Nasir Mahmood. Email: nasir202003@yahoo.com

Received: 30 August 2020; Accepted: 16 March 2021

Abstract: This study proposes an architecture for the prediction of extremist human behaviour from projected suicide bombings. By linking ‘dots’ of police data comprising scattered information of people, groups, logistics, locations, communication, and spatiotemporal characters on different social media groups, the proposed architecture will spawn beneficial information. This useful information will, in turn, help the police both in predicting potential terrorist events and in investigating previous events. Furthermore, this architecture will aid in the identification of criminals and their associates and handlers. Terrorism is psychological warfare, which, in the broadest sense, can be defined as the utilisation of deliberate violence for economic, political or religious purposes. In this study, a supervised learning-based approach was adopted to develop the proposed architecture. The dataset was prepared from the suicide bomb blast data of Pakistan obtained from the South Asia Terrorism Portal (SATP). As the proposed architecture was simulated, the supervised learning-based classifiers naïve Bayes and Hoeffding Tree reached 72.17% accuracy. One of the additional benefits this study offers is the ability to predict the target audience of potential suicide bomb blasts, which may be used to eliminate future threats or, at least, minimise the number of casualties and other property losses.

Keywords: Extremism; terrorism; suicide bombing; crime prediction; pattern recognition; machine learning; supervised learning

1 Introduction

Crime is a politico-socio-economic problem that adversely affects people worldwide, marring the social welfare and progress of the masses. Law enforcement agencies need to formulate crime policies and strategic plans to prevent crimes and reduce crime rates. However, they face the challenge of effectively extracting relevant knowledge from a large volume of criminal data and reports [1]. Knowledge discovery (KD) and data mining from this mass of data require sophisticated analytical processing. This KD is ultimately used to provide practical decision-making support to law enforcement agencies. Nevertheless, the analytical processing of large amounts of data is complicated for humans [2]. Therefore, scholars have proposed numerous techniques to



This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

automate a significant part of or the entire analytical process. Crime data mining is a well-known set of measures to automatically extract hidden knowledge from dimensional databases [3].

In comparison to other crimes, terrorism—an aggregate of brutality—is a more complex and versatile framework that directly affects harmony, the daily schedule of nations, and social orders. It is utilised to produce widespread, global dread among nations' citizenry [4]. Acts of terrorism encompass a broad range of activities, including criminal communication, planning, transportation, logistics, reconnaissance of targets, harbouring, providing materials and items such as weapons, and supporting other relevant materials and finances. Acts of terrorism by extremist or terrorist organisations create psychological pressure amongst civilians and governments; the military, due to the casualties; and property losses, social unrest, and economic disruption [5]. There is no universal consensus on the cause and roots of terrorism [6]. According to the Global Terrorism Database (GTD) maintained by the University of Maryland, more than 61,000 incidents occurred between 2000 and 2014, resulting in 140,000 casualties. Meanwhile, the Global Terrorism Index (GTI) reported an increase in terrorism from 65 countries in 2015 to 77 countries in 2016, indicating more devastation than over the last 17 years [7].

After the terrorist attacks on the World Trade Centre and the Pentagon on 11 September 2001, the former President of the United States, George Walker Bush, introduced the phrase 'war against terror'. The campaign was launched with the US and UK invasion of Afghanistan. The attack engendered mass displacement in Afghanistan and compelled the country's people to take refuge in the religiously, ethnically, linguistically and culturally associated tribal areas of Pakistan. This cross-border displacement of people from Afghanistan to Pakistan caused social, political and economic upheaval, leading to a rise in militancy and extremism, drone attacks, internally displaced persons (IDPs) and suicide bombing attacks [1].

Pakistan has been an active target of terrorist activities for the past 18 years. Between 2001 and 2010, the country's anti-terrorist activities incurred a total of \$68 billion. Moreover, in 2009, 300,000 IDPs were recorded in the wake of different forms of terrorist acts, including target killings, military operations, planted bombs and suicide attacks [2]. Suicide bombing is a terrorist practice targeting military personnel, famous personalities, religious sites and civilians. These attacks are usually carried out using vehicles or by individuals wearing vests and carrying explosives [1]. [Tab. 1](#) illustrates the data regarding suicide attacks in Pakistan from 8 May 2002 to 17 June 2018.

Crime and terrorism are common problems in almost every society because they affect the quality of life and economic growth. They bring fear and disrupt the population's unity by breaking social associations. The discipline of criminology involves the study of crime and criminal behaviour and a process that aims to identify a crime's characteristics, motives and hidden patterns. The emergence of modern techniques, such as machine learning (ML) and data mining, and the availability of a high volume of crime and terrorist datasets, have enabled the identification and prediction of crimes [8]. The predictive capability of crime, facilitated by the effective implementation of security policies, can assist crime prevention [9].

There is substantial statistical proof that crime and terrorism are predictable not because criminals and terrorists operate in their comfort zone; rather, a frequency of variables make their methods work well. The most significant theories supporting this hypothesis include criminal behaviour theory, routine activity theory, rational choice theory, and crime pattern theory. These theories are consolidated to form a blended theory. The proposed research is based on the crime pattern theory [8].

Table 1: Suicide bomb blast incidences in Pakistan. Source: South Asia terrorism portal (SATP)

Year	Incidents	Killed	Injured
2002	1	15	34
2003	2	69	103
2004	7	89	321
2005	4	84	219
2006	7	161	352
2007	54	765	1677
2008	59	893	1846
2009	76	949	2356
2010	49	1167	2199
2011	41	628	1183
2012	39	365	607
2013	43	751	1411
2014	35	336	601
2015	20	188	410
2016	19	401	935
2017	22	369	1052
2018	11	61	132
Total	479	7291	15428

Data is characterised as an assortment of facts and figures, statistics, and measurements that can be utilised for references and examinations to reach determinations. Information assortment is a pivotal and deliberate way to deal with data from various sources to obtain an exact image of a region of intrigue. It assists in answering research questions, formulating a hypothesis and drawing conclusions. The objective of information assortment is to collect high-quality evidence that will then be converted to allow for an information-rich investigation, permitting the structure of persuasive and tenable responses to explore questions. Accurate data collection is key to ensuring the morality of the study. Information assortment is of particular significance in the domain of terrorism and related fear-based oppressive exercises. Exact and reliable information can help to stop psychological warfare exercises and execute security approaches that can forestall the development of fear-based oppressor gatherings [10].

ML is a sub-domain of artificial intelligence that uses the computational statistical model. It is widely used to design intelligent algorithms that can learn from previous data or knowledge to make future predictions or decisions. ML answers two fundamental questions related to artificial intelligence: namely, how can computer systems automatically improve themselves through experience, and what are the basic statistical, computational information laws that govern all the learning systems? Learning problems can be defined as problems surrounding the improvement of performance measures through training experience. Applications of ML include computer vision, email filtering, predictive analytics, natural language processing, optical character recognition and pattern recognition [11].

Supervised learning (SL) techniques require a sufficient amount of labelled training data for classification or to label unseen test data. In contrast, deep learning (DL) techniques emerged recently from artificial neural networks (ANNs), requiring minimal engineering by hand. Thus,

the latter methods can benefit from an increase in the amount of available computation and data compared with classical ML techniques and external neural networks. Supervised ML algorithms are frequently used to recognise, understand and translate human languages to extract meaningful information [11], and ML and analytics have contributed to the development of many medical, financial, technological, and business-and science-related applications. ML has also proven to be a vital means of understanding, analysing and predicting criminal and terrorist behaviour [12]. Previous studies have concentrated on theoretical models to develop a hypothesis about causes and consequent effects. ML algorithms are innovative and have predictive capabilities: ML can add robustness to the variables of a sample by validating actual predictive capabilities. Furthermore, ML can rank variables by the influence of predictive accuracy, giving a sense of the importance of a particular variable [6].

Understanding crime is the objective of many types of research and studies. While numerous benchmark datasets are available, it is difficult to extract some attributes, such as the number of casualties and expected injuries, from the crime and terrorism data. Some non-linear models have been proposed to find a correlation between crime data and urban matrices. However, because of non-Gaussian distributions and multi-correlation in urban indicators, it is common to find controversial conclusions about the influence of some urban indicators on crime [13]. ML methods frequently rely on supervised classification learning, which includes support vector machines (SVMs), ANNs, the naïve Bayes classifier (NB), and maximum entropy [14]. The knowledge gained from the ML and data mining approaches and techniques can help law enforcement agencies prevent or decrease criminal and terrorist activities in society [8].

This study adopts and recommends the SL-based approach to predict the target audience (target class) of suicide bombers using the suicide bomb blast data of Pakistan available on the SATP. The study has the following objectives:

- Analysing Pakistan's suicide bombing data to understand and extract valid attacker behavioural variables available in the data.
- Designing an accurate dataset to train and test the system efficiently (useful for future studies).
- Finding the maximum results through the twin operations of training and testing by applying simple SL algorithms.

2 Literature Review

The literature review for this study covers three significant dimensions of terrorist events and data: suicide bombing, previously proposed suicide bombing and crime prediction techniques, and the use of ML for crime prediction.

The primary aim of qualitative research by Abbasi et al. [1] was to analyse the social, economic and physiological implications and repercussions for Pakistan after 9/11 and the ensuing war against terror. This study's major findings included a relationship between religious extremist behaviour and suicide bombing, external invasion, and internal displacement. The study provided a better understanding of suicide bombing culture and other useful statistical details of terrorist events.

Rasheed et al. [10] discussed the existing sources of suicide bombing data and datasets, presenting a vital case study of a data collection related to suicide bombings in Pakistan. Important contributions made by the study included new variables, such as explosion types, explosives, perpetrators, motives, etc. These variables, of course, explain the phenomena of suicide attacks.

Agarwal et al. [7] provided useful insights by tracking patterns and trends [10] in their analysis of a historical dataset of the GTD. They predicted the factors that can potentially correlate with the menace of terrorism. Different data mining and ML techniques, including SVMs, random forest (RF), and logistic regression (LR), were employed to analyse the dataset and predict the would-be terrorist groups, the success or failure of the attacks, and their effects on external factors. In the implementation, k-means clustering and dummy classifiers showed an improvement, while RF with the GTD and dummy classifiers with the GTD peaked to the marks of 0.82 and 0.56, respectively [7].

Gao et al. [5] also used the GTD to compare five classification ML algorithms: the decision tree (DT), LR, a Gaussian Bayesian Network (GBN), RF and AdaBoost. The experiment results showed that classification based on the DT had the highest precision at 94.8%. Moreover, the GBN could list all the possibilities according to the probabilities and showed 94.7% of the results [5]. In another study, Mehmood et al. [2] acquired data concerning terrorism in Pakistan between 1998 and 2012 from the SATP. The methodology included a cluster analysis based on statistical correlation, followed by data pre-processing. The clusters were discovered over event and target; event and method were used in terrorism, and a more significant clustering grouped the distinct combination into separate clusters. The clusters were analysed according to three dimensions: the period, geography and type of terrorist events. The authors found that some critical terrorist activities, including suicide bomb attacks after 2012, reshaped the architecture of terrorism networks and events in Pakistan.

Soliman et al. [4] proposed a hybrid computational intelligent algorithm as a decision support tool for the phenomenon of terrorism. The algorithm was based on different decision support tools and data mining techniques and aimed to improve the previously proposed algorithms inspired by meta-heuristics. The algorithm could predict the terrorist groups responsible for terrorist attacks on different regions of Egypt from 1996 to 2017. The accuracy of the prediction model with the neural network (NN) decision-maker was recorded at 74.77% with a mean square error (MSE) = 0.018 at iteration 10 = 0.0860. Although the proposed algorithm provided a marked accuracy, it involved complex implementation details [4].

Basuchoudhary et al. [6] argued that ML could explain the phenomenon of terrorism since it can replace the missing data in scientifically validated ways. ML can help to reduce the multidimensionality of the most commonly used variables with no causal effect, explaining why it can identify causal variables. In another study, Basuchoudhary et al. [6] outlined how ML could be a vital part of the iterative knowledge-building process. Greitzer et al. [15] asserted that insider attacks could be detected based on psychological, behavioural, physical and sociotechnical indicators and factors mapped into a domain ontology. The proposed solution incorporated the technical indicators from the previous work. The ontology was derived from the taxonomy of the domain knowledge [15].

Moreover, Nguyen et al. [12] put forward a crime forecasting method that predicted crimes based on location and time. The techniques comprised data acquisition, pre-processing, linking data with demographic data, and prediction using ML techniques including SVMs, RF, gradient boosting machines (GBMs), and NNs. The results obtained by the classifiers SVMs, RF, GBMs and NNs using scaled conjugate back-propagation and resilient back-propagation were 79.39%, 65.79%, 61.67%, 74.02%, and 74.24%, respectively. The data was provided by the Portland Police Bureau and the public government source American Factfinder [12].

Kang [9] proposed a feature-level data fusion technique with the environmental context based on a deep neural network (DNN). The dataset consisted of online databases of crime statistics, demographics, meteorological data and images in Chicago, Illinois. Experimental performance results showed that this DNN model was more accurate in predicting crime than other prediction models [9]. Meanwhile, a study by Ahishakiye et al. [8] considered developing a crime prediction prototype model using the DT (J48) algorithm since the related literature has argued that it is the most efficient ML algorithm for predicting crime data. From the experimental results, the J48 algorithm predicted the unknown category of crime data at an accuracy of 94.25%, a rate high enough for the system to be relied upon to predict future crimes. The dataset, entitled ‘Crime and Communities’, was acquired from the UCI Machine Learning Repository.

Azizan et al. [14] developed a terrorism detection technique using ML through sentiment analysis on the microblogging social website Twitter. Terrorists and people who support terrorism demonstrate patterns in these sentiments, which run through the very fabric of the comments, tweets, or messages they post. This study built upon the current sentimental analysis methods and techniques by using ML for crime prediction. The NB accuracy was recorded at between 85% to 96% [14]. Alves et al. [13] obtained accurate predictions through statistical learning, suggesting that crime prediction depends on urban matrices and indicators. The proposed model provided a better solution to predicting crime with good accuracy and identifying the importance of the feature. It also held, even under small perturbation, on the training dataset [15]. This approach showed up to 97% accuracy using RF classifiers. Furthermore, the importance of urban indicators was ranked and clustered in groups of equal influence in the data sample analysed [13].

Using ML, Singh et al. [16] were able to predict terrorist attacks by country and region. This study was carried out upon the GTD. Six ML algorithms were applied to the selected dataset to achieve 82% accuracy using NB and LR. Gerber [17] argued that the importance of GPS-tagged tweets for crime prediction has been ignored in the literature; fewer types of crime have been discussed, and the performance comparison of previously proposed and currently used hot-spot models on Twitter have not been addressed. Finally, Lim et al. [18] conducted experiments to demonstrate that a criminal network link prediction model based on deep reinforcement learning (DRL) outperformed the GBM model with a relatively smaller dataset. AUC scores were 0.85, 0.82, and 0.76 for the DRL criminal network link prediction model, compared with the GBM model scores of JUANES, MAMBO, and JAKE, respectively. The experiments indicated that the DRL method was capable of a better predictive performance than conventional SL under the same hyper-parameter setting. Further research should focus on confirming whether the SL technique’s predictive precision would be stronger over a larger dataset and number of training iterations [18].

3 System Modelling

Fig. 1 shows the steps involved in the current study. Its details are discussed below.



Figure 1: Proposed model

3.1 Data Collection

3.1.1 Multilayer Perceptron

The data were collected from the SATP, which launched in March of 2000. The SATP is the largest comprehensive, searchable and continually updated database on terrorism, low-intensity warfare, and ethnic/communal/sectarian strife in South Asia. The project is the initiative of the Institute for conflict management (ICM). The ICM provides consultancy services on terrorism and internal security to various governments. It was established in 1997 in New Delhi, India, and was registered as a non-profit, non-governmental organisation supported by voluntary contributions and project aid [10]. SATP has data of 479 suicide bomb blast incidents from 2002 to 2018. [Tab. 2](#) contains sample data of the first five suicide bomb blasts in 2017.

Table 2: Suicide bomb blast incidences in Pakistan. Source: SATP

Sr. No.	Date	Place/District	Incidents	Killed	Injured
1	7 February	Mandan/Bannu/KP	At least two policemen were injured in an explosion at the main gate of Mandan Police Station in the Bannu District of Khyber Pakhtunkhwa (KP).	1	2
2	13 February	Mall road/Lahore/Punjab	At least 14 persons, including six police officers, were killed and 85 injured when a suicide bomber struck around 6 pm outside the Punjab Assembly on Mall Road of Lahore, Punjab's provincial capital, during a protest.	15	85
3	15 February	Ghalanai/Mohmand agency/FATA	Three levies personnel and two civilians were killed while eight others were injured in twin suicide attacks in Ghalanai Tehsil of Mohmand Agency in FATA.	7	8
4	15 February	Hayatabad/Peshawar/KP	A civil judge vehicle driver was killed when a motorcycle-borne suicide bomber rammed his bike into the vehicle in the Hayatabad area of Peshawar, the provincial capital of KP. Civil Judge Asif Jadoon and three female judges of the lower judiciary were travelling in an official car in the Phase 5 area when an attacker on a motorcycle struck the front of the vehicle, causing an explosion, which killed the driver and injured the four judges.	2	4
5	16 February	Sehwan Sharif/Jamshoro/Sindh	At least 88 people were killed when a suicide bomber attacked the crowded Sufi shrine of Lal Shahbaz Qalandar in the Sehwan Sharif town of Jamshoro District in Sindh, injuring at least 343 others.	89	343

3.2 Dataset Preparation

There were four steps to the dataset preparation phase, all of which are shown in Fig. 2 and explained in further detail below.

3.2.1 Multilayer Perceptron

The data's initial investigation revealed that 479 suicide bombing incidents took place in Pakistan from 2002 to 2018. These incidents claimed the lives of more than 7,291 people, while 15,428 people were injured. The preliminary analysis indicated five attributes to the data: the year and date, the location (district), incident detail, number of people killed, and the number of people injured.

3.2.2 Data Preprocessing

The dataset preparation was accomplished in four steps. The phases involved in the data preparation are illustrated in Fig. 2.



Figure 2: Steps involved in dataset preparation

3.2.3 Knowledge Discovery and Extracted Variables

After pre-processing, important behavioural information and attributes were identified and extracted from the available data. For example, the ‘selected date’ attribute provided a further valuable discovery: that the attacker chose specific days of the week for the attacks. Similarly, the chosen incident filed (alpha-numeric text) presented some novel and critical attributes, such as blast type, target type, and attack space, as illustrated in Fig. 3.

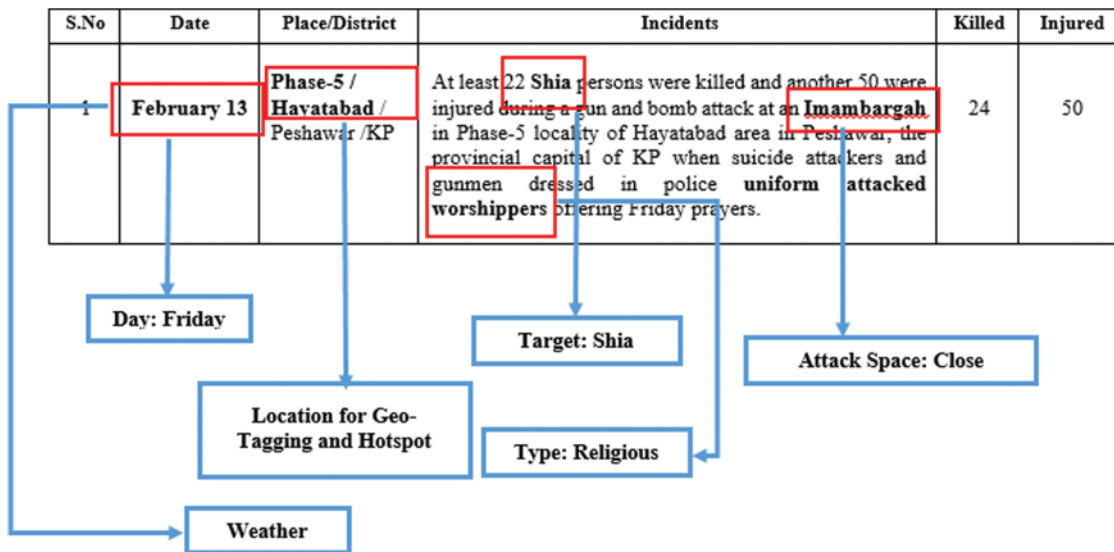


Figure 3: Novel attributes discovered in collected data

The behavioural attributes of the attacker were extracted from the dataset, as shown in [Tab. 3](#).

Table 3: Behavioural attributes of the attacker along dataset

Sr. No.	Behavioural variables	Sample data	Description
1	Day selection	Mon, Tue, Wed, Thu, Fri, Sat, Sun	Day of the week
2	Month selection	Jan, Feb, Mar, April, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec	Month of the year
3	Location selection	(x, y) coordinates (numeric)	Map location
4	District/city selection	Isl., Lah., Kar., Pes., etc.	Cities of Pakistan
5	Province/state selection	Pun., Sin., KPK, Bal., AJK, GB, FATA	Provinces of Pakistan
6	Attacker motive	Rel., Pol., Gen., VIP, law enforcement	The motive of the attack
7	Target selection	Army, civil, Shia, FC	Target audience
8	Modus operandi	Fadai, van hit, multiple targets, multiple attackers	Chosen steps to execute the attack
9	Attack type	Suicide, gunfire, explosion, mixed	Type of attack conducted
10	Space selection	Open, closed	Indoor/outdoor environment
11	Weather selection	Numeric value	Temperature, humidity, etc.
12	Attacker attire	Police uniform	Dress to approach the target
13	Attacker gender	Male, female, both	Female or male or both
14	Attacker age group	Young (15–30), middle-age (30–40), other	Which age group conducted the attack
15	Killed	Numeric value	Number of people killed in an attack
16	Injured	Numeric value	Number of people injured in an attack

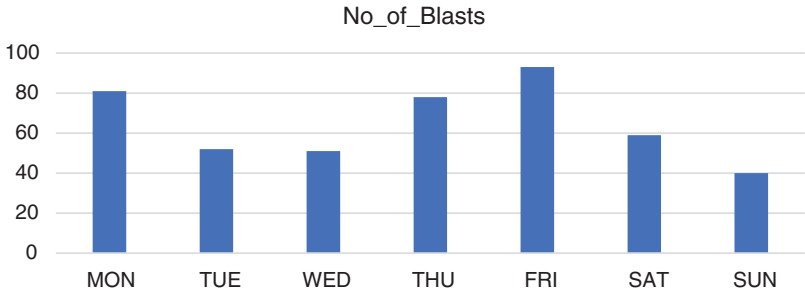
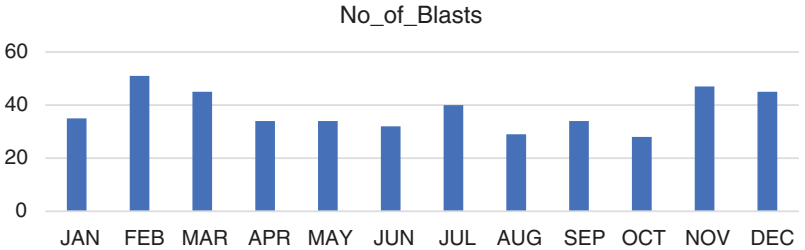
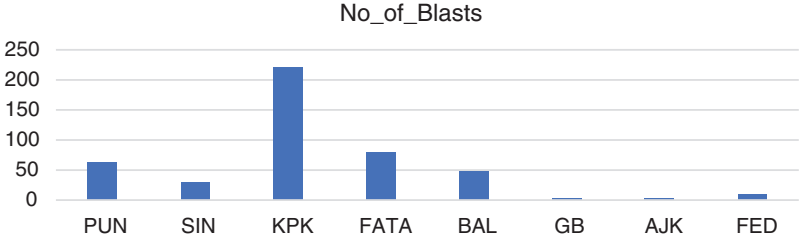
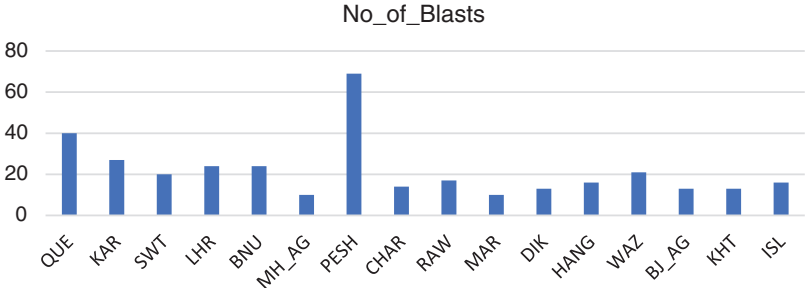
3.3 Obtained Dataset

The data obtained via the SATP recorded 479 suicide bomb blasts. There were, however, missing values and duplication in the available data; thus, the obtained dataset includes 454 instances. The dataset's most important attributes include the month, day, state/province, district/city, and blast type. [Tab. 4](#) presents the preliminary analysis and complete detail of the obtained dataset.

[Tab. 4](#) demonstrates some of the fundamental and most notable patterns from the dataset. The most critical days of the week in Pakistan are Friday, with weight 93, followed by Monday and Thursday with weights 81 and 78. November, December, February, and March are the most critical months, with weights of 47, 45, 51, and 45, respectively. KPK (Khyber Pakhtunkhwa), the third most populated province of Pakistan, is the most affected province, with 224 incidences recorded. This is followed by FATA (the territory that has now been merged with KPK), with 79 incidences recorded, and Punjab (the most populated province of Pakistan), with 63 incidences reported. The most affected city is Peshawar (capital of the province KPK) with 69 blasts,

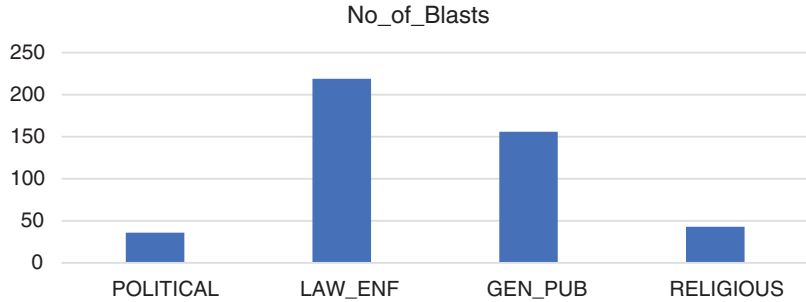
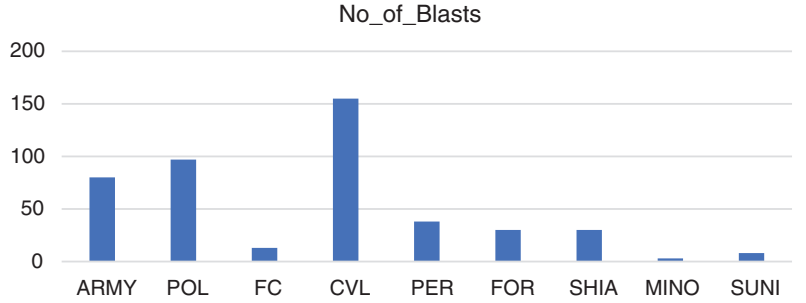
followed by Quetta (capital of the province Baluchistan) with 40 blasts, and Karachi (the most populated city and economic hub of Pakistan with a seaport; the capital of the province Sindh) with 27 blasts. Variable blast type represents the main agenda or motive (for instance, political or religious) of the attack. The most recorded blast type is ‘Law_Enf’, the forces that maintain law and order in the country. These forces were the prime targets of the suicide bomb blasts, with a count of 219. These blasts also claimed a further 165 and 43 lives by targeting the general public and religious ceremonies.

Table 4: Dataset patterns

S. No.	Attribute	Detail
1	Day	 <p>No_of_Blasts</p>
2	Month	 <p>No_of_Blasts</p>
3	State/Province	 <p>No_of_Blasts</p>
4	City/District	 <p>No_of_Blasts</p>

(Continued)

Table 4: Continued

S. No.	Attribute	Detail																				
5.	Blast type	 <table border="1"> <caption>Data for Figure 5: Blast type</caption> <thead> <tr> <th>Blast Type</th> <th>No_of_Blasts</th> </tr> </thead> <tbody> <tr> <td>POLITICAL</td> <td>35</td> </tr> <tr> <td>LAW_ENF</td> <td>220</td> </tr> <tr> <td>GEN_PUB</td> <td>155</td> </tr> <tr> <td>RELIGIOUS</td> <td>40</td> </tr> </tbody> </table>	Blast Type	No_of_Blasts	POLITICAL	35	LAW_ENF	220	GEN_PUB	155	RELIGIOUS	40										
Blast Type	No_of_Blasts																					
POLITICAL	35																					
LAW_ENF	220																					
GEN_PUB	155																					
RELIGIOUS	40																					
6.	Target audience	 <table border="1"> <caption>Data for Figure 6: Target audience</caption> <thead> <tr> <th>Target Audience</th> <th>No_of_Blasts</th> </tr> </thead> <tbody> <tr> <td>ARMY</td> <td>80</td> </tr> <tr> <td>POL</td> <td>90</td> </tr> <tr> <td>FC</td> <td>10</td> </tr> <tr> <td>CVL</td> <td>155</td> </tr> <tr> <td>PER</td> <td>35</td> </tr> <tr> <td>FOR</td> <td>25</td> </tr> <tr> <td>SHIA</td> <td>25</td> </tr> <tr> <td>MINO</td> <td>5</td> </tr> <tr> <td>SUNI</td> <td>10</td> </tr> </tbody> </table>	Target Audience	No_of_Blasts	ARMY	80	POL	90	FC	10	CVL	155	PER	35	FOR	25	SHIA	25	MINO	5	SUNI	10
Target Audience	No_of_Blasts																					
ARMY	80																					
POL	90																					
FC	10																					
CVL	155																					
PER	35																					
FOR	25																					
SHIA	25																					
MINO	5																					
SUNI	10																					

The dataset contains nine labelled classes. Each class represents a particular target audience of a suicide blast in which an instance belongs. These nine labelled classes are ARMY (Pakistan Armed Forces), POL (police), FC (Frontier Core, which is the core of the Pakistan Army), CVL (civilians), PER (personalities), FOR (armed forces, such as Khasadar Force and the Military Police, which is appointed to protect Pakistan's borders and primarily tribal areas), SHIA (the second biggest religious sect in Pakistan), MINO (minorities, such as Christians, Sikhs, etc.) and SUNI (the most prominent religious sect in Pakistan). The most recorded class is civilians with a count of 155, followed by police and the Pakistan Armed Forces with the counts of 90 and 87, respectively. For this study, the dataset has been divided into 75% of the training set (339 instances) and 25% of the testing set (115 instances).

3.4 Supervised Learning

Supervised and unsupervised learning are the two main branches of ML. SL is still the most esteemed branch of pattern recognition in the ML field [19]. Supervised machine learning, also called the classification learning approach, is used for analysing training or labelled data to map unseen instances of data for future classification. Features extracted from recognition units train a classifier that learns to differentiate between different pattern classes [20]. SL techniques require a sufficient amount of labelled training data for classification or to label unseen test data [21].

Conversely, DL, which recently emerged from ANNs, requires minimal engineering by hand and can thus take advantage of an increase in the amount of available computation and data compared with classical ML techniques and external neural networks [22]. The results generated by the SL approach are discussed in the next section.

4 Results

WEKA™ (the Waikato Environment for Knowledge Analysis) version 3.8.4, considered one of the most efficient ML and data mining tools, was used to determine the overall efficiency of the proposed method and dataset. Different algorithms are available for SL, including Bayes, Function, Lazy, Meta, Rules, Tree, etc. In this study, the proposed technique and dataset's efficiency was demonstrated using different, widely used algorithms; nevertheless, results could also be generated using other SL algorithms.

4.1 Training Results

The Bayesian network consists of a structural model and a set of conditional probabilities. Bayesian-based algorithms are often used for classification problems in which learning is done by constructing a classifier from a set of training instances with labelled classes [23,24]. Bayes algorithms are simple, supervised probabilistic classifier algorithms [21] used for binary or multiclass classification based on Bayes' theorem.

The NB algorithm is a simple supervised probabilistic classifier algorithm used for binary or multiclass classification. This algorithm is highly scalable.

In Bayes' theorem, the probability of a hypothesis (h) given data (D) can be expressed as:

$$P(h|D) = P(h) \cdot P(D|h) \quad (1)$$

$$P(D|h) = P(D) \cdot P(h|D) \quad (2)$$

From Eqs. (1) and (2), we get:

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)} \quad (3)$$

Bayes' theorem can be defined as:

$$h_{MAP} = \text{Arg}_{h \in H} \max P(h|D)$$

$$h_{MAP} = \text{Arg}_{h \in H} \max \frac{P(D|h) \cdot P(h)}{P(D)}$$

$$h_{MAP} = \text{Arg}_{h \in H} \max P(D|h) P(h) \quad (4)$$

Using Eq. (3),

$$P(h|D) = \frac{P(D|h) \cdot P(h)}{P(D)}$$

$$P(Y|X') \propto P(X'|Y) \cdot P(Y)$$

$$\text{NB} = P(X_1 \cdot X_2 \cdot X_3 \dots X_n | Y) \cdot P(Y) \quad (5)$$

$$= P(X_1 | Y) \cdot P(X_2 \cdot X_1 | Y) \cdot P(X_3 \cdot X_2 | Y) \dots P(X_n | X_1 \dots X_{n-1} Y) \cdot P(Y)$$

$$Y^{new} = P(X_1 | Y) \cdot P(X_2 | Y) \cdot P(X_3 | Y) \dots P(X_n | Y) \cdot P(Y)$$

$$\text{NB} = \text{Arg}_{h \in H} \max \sum_{i=1}^n P(X_i | Y = Y_h) \quad \text{Where } P(Y = Y_k) \quad (6)$$

Table 5: Obtained training results generated by different SL algorithms

Supervised learning algorithms		Training results parameters							
		No. of instances	Correctly classified instances	Incorrectly classified instances	Kappa statistic	Mean absolute error	Root Mean squared error	Relative absolute error (%)	Root relative squared error (%)
Bayes	Bayes net	339	274 (80.8%)	65 (19.2%)	0.754	0.063	0.1735	36.2	58.6
	Naïve Bayes	339	267 (78.7%)	72 (21.2%)	0.726	0.074	0.179	40.0	60.5
Function	Logistics	339	311 (91.7%)	28 (8.3%)	0.895	0.027	0.1161	15.9	39.24
	Multilayer Perceptron	339	322 (94.9%)	17 (5.1%)	0.931	0.014	0.087	8.41	29.45
Lazy	SMO	339	257 (75.8%)	82 (24.1%)	0.687	0.080	0.191	45.8	64.86
	IBK	339	290 (85.5%)	49 (14.5%)	0.813	0.173	0.283	99.0	95.74
	K-Star	339	331 (97.6%)	8 (2.4%)	0.97	0.024	0.070	13.98	23.85
Meta	Lazy-LWL	339	222 (65%)	117 (35%)	0.52	0.093	0.222	56.6	74.61
	Bagging	339	281 (82.8%)	58 (17.1%)	0.77	0.051	0.161	33.10	54.42
	Classification via regression	339	286 (84.3%)	53 (15.6%)	0.79	0.063	0.160	36.17	54.62
	Iterative classifier optimiser	339	244 (71.9%)	95 (28.0%)	0.63	0.089	0.207	51.22	70.13
	Logit boost	339	280 (82.5%)	59 (17.4%)	0.77	0.067	0.169	38.58	57.28
Rules	Multi class classifier	339	303 (89.3%)	36 (10.6%)	0.86	0.034	0.126	19.75	42.79
	Randomise filtered classifier	339	331 (97.6%)	8 (2.35%)	0.97	0.010	0.053	5.94	18.20
	Part	339	237 (69.9%)	102 (30.1%)	0.606	0.086	0.208	49.4	70.44
	Hoeffding tree	339	267 (78.7%)	72 (21.2%)	0.726	0.070	0.179	40.08	60.55
Tree	LMT	339	257 (75.8%)	82 (24.1%)	0.682	0.080	0.191	45.85	64.84
	Random forest	339	331 (97.6%)	8 (2.3%)	0.97	0.041	0.094	23.6	31.93
	Random tree	339	331 (97.6%)	8 (2.3%)	0.97	0.005	0.053	3.23	18.01

NB algorithms, alongside other SL algorithms, generated the training results, as shown in [Tab. 5](#). This study's primary objective is to train the system to classify different alphabets and characters and their different writing styles into accurate classes. It is suitable to use information retrieval measures to evaluate the overall accuracy of the proposed system. The two main information retrieval measures are break-even points and the F-score. The F-score measures the effect of a system's performance on a particular class, whereas the break-even point is a value where precision and recall become equal: it is not, therefore, a good indicator of classification performance. However, accuracy is a valid parameter to measure the system's overall performance [19]. The measures of F-score and accuracy are illustrated below:

True positive (Tp) = No. of characters correctly classified to the class

True negative (Tn) = No. of characters correctly rejected to the class

False positive (Fp) = No. of characters incorrectly rejected to the class

False negative (Fn) = No. of characters incorrectly classified to the class

By using the terms mentioned above, we can generate measurement factors as:

$$\text{Accuracy} = \frac{\mathbf{Tp} + \mathbf{Tn}}{\mathbf{Tp} + \mathbf{Tn} + \mathbf{Fp} + \mathbf{Fn}} \quad (7)$$

$$\text{Precision} = \frac{\mathbf{Tp}}{\mathbf{Tp} + \mathbf{Fp}} \quad (8)$$

$$\text{Recall} = \frac{\mathbf{Tp}}{\mathbf{Tp} + \mathbf{Fn}} \quad (9)$$

$$\text{F - Score} = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

Table 6: Test results generated by different SL algorithms

Supervised learning algorithms		Training results parameters							
		No. of instances	Correctly classified instances	Incorrectly classified instances	Kappa statistic	Mean absolute error	Root mean squared error	Relative absolute error (%)	Root relative squared error (%)
Bayes	Bayes net	115	82 (71.30%)	33 (28.6%)	0.637	0.0886	0.2233	50.12	74.97
	Naïve Bayes	115	83 (72.17%)	32 (27.8%)	0.642	0.091	0.2201	51.72	73.92
Function	Logistics	115	63 (54.78%)	52 (45.21%)	0.445	0.101	0.3010	57.59	99.29
	Multi-layer Perceptron	115	82 (71.30%)	33 (28.69%)	0.638	0.069	0.232	39.497	78.016
Lazy	SMO	115	81 (70.43%)	34 (29.56%)	0.624	0.176	0.288	99.03	96.77
	IBK	115	68 (59.13%)	47 (40.86%)	0.479	0.097	0.263	55.11	88.38
	K-Star	115	74 (64.34%)	41 (35.65%)	0.545	0.112	0.239	63.43	80.45
Meta	Lazy-LWL	115	67 (58.26%)	48 (41.73%)	0.438	0.106	0.227	59.955	6.42
	Bagging	115	76 (66.08%)	39 (33.91%)	0.569	0.085	0.226	48.53	76.11
	Classification via regression	115	81 (70.43%)	34 (29.56%)	0.625	0.086	0.219	48.86	73.55
	Iterative classifier optimizer	115	81 (70.43%)	34 (29.56%)	0.620	0.094	0.210	53.58	70.70
Rules	Logit boost	115	79 (68.69%)	36 (31.30%)	0.605	0.087	0.214	49.23	72.06
	Multiclass classifier	115	65 (56.52%)	50 (43.47%)	0.458	0.102	0.274	57.84	92.16
	Randomise filtered classifier	115	39 (33.91%)	76 (66.08%)	0.189	0.148	0.378	84.12	127.15
Tree	Part	115	77 (66.95%)	38 (33.04%)	0.575	0.092	0.211	52.26	70.90
	Hoeffding Tree	115	83 (72.17%)	32 (27.82%)	0.646	0.091	0.220	51.72	73.92
	LMT	115	82 (71.30%)	33 (28.69%)	0.635	0.089	0.209	50.85	70.21
	Random forest	115	75 (65.21%)	40 (34.78%)	0.556	0.107	0.231	60.92	77.59
	Random tree	115	48 (41.73%)	67 (58.26%)	0.248	0.125	0.307	71.21	103.0

Some classes are not shown in [Tabs. 6](#) and [7](#), such as F1-1 and F8-2, depicting that the instances belonging to these classes are part of the test set (i.e., the training set and test set are not overlapping). The precise accuracy by level using the NB algorithm is illustrated in [Tab. 6](#).

Table 7: Detailed class accuracy using NB algorithms

TP rate	FP rate	Precision	Recall	F-Measure	MCC	ROC area	PRC area	Class
0.400	0.078	0.588	0.400	0.476	0.374	0.824	0.535	Army
0.800	0.200	0.526	0.800	0.635	0.526	0.862	0.607	Police
0.333	0.018	0.333	0.333	0.333	0.315	0.899	0.254	FC
0.972	0.000	1.000	0.972	0.986	0.980	0.998	0.996	Civilians
0.889	0.000	1.000	0.889	0.941	0.938	0.990	0.942	Personality
0.000	0.009	0.000	0.000	0.000	-0.020	0.649	0.075	Force
1.000	0.028	0.750	1.000	0.857	0.854	0.990	0.852	Shia
0.000	0.000	0.000	0.000	0.000	0.854	0.889	0.113	Minority
0.000	0.009	0.000	0.000	0.000	-0.009	0.904	0.083	Sunni
0.722	0.064	0.000	0.722	0.000	0.000	0.909	0.713	Weighted Average

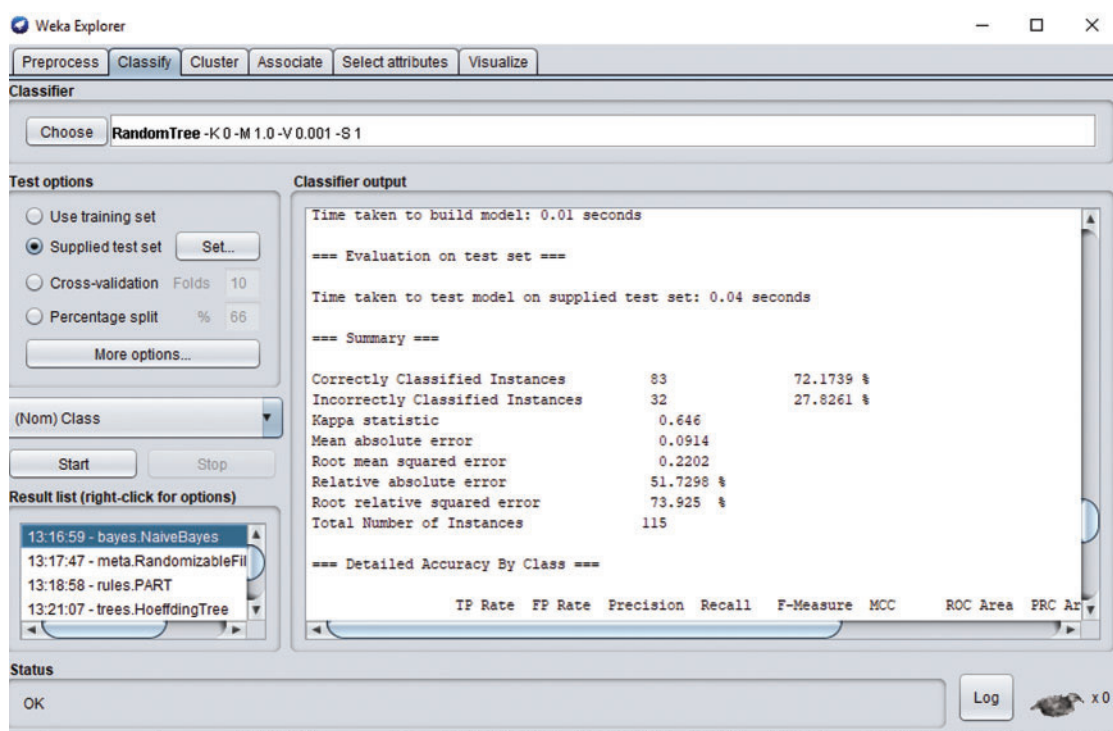


Figure 4: Training results using NB algorithm

4.2 Testing Results

The training results generated by the NB algorithm are shown in Fig. 4 as a sample. Tab. 6 presents the comparison of results obtained using the different supervised algorithms, and Tab. 7 demonstrates the detailed accuracy by class using the NB algorithm.

5 Conclusion

Different algorithms, such as NB, SMO, and the Hoeffding Tree, gave accuracies of 72.17%, 71.30%, and 72.17%, respectively. Other SL algorithms, such as LMT, Logit Boost and Iterative Classifier Optimiser, also show promising results. These results can help predict a specific class of people who may become victims of suicide bomb blasts. With the help of simple variables, including the day, month, province, and city, the target can be predicted, allowing for the prevention of these attacks altogether or, at least, minimising the casualty rates and damages. In the future, establishing more information about the impact of suicide bombing would be useful for developing proficient datasets and improving the exactness of results.

Furthermore, more variables can be introduced for accurate predictions. Results can also be generated using different SL classifiers and algorithms and JRip, OneR, WAODE, etc. Criminal and suicide attacker profiling could be accomplished if the data are available, and an accurate target audience of a suicide bomb blast can be achieved using SL classifiers. Indeed, SL classifiers (e.g., NB and SMO) are generating significant results. The present study helps predict the target class of the suicide blast, which would aim to prevent or reduce the impact of damages, including casualties and loss of properties.

6 Future Plan

The ontology will be developed to explain concepts of behavioural aspects in the terrorism domain for machines, law enforcement personnel and researchers. An ontology is an explicit specification of a conceptualisation [25]. The ontology of the terrorism domain will focus on concepts, relationships, and mapping by observing specific standards and principles. ML and data mining techniques will be used to predict and classify the incidents and groups' involvement on the basis of features extracted from the dataset using this domain ontology.

Acknowledgement: Thanks to our families and colleagues, who supported us morally.

Funding Statement: The authors received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

- [1] I. A. Abbasi, M. K. Khatwani and F. Y. Panhwar, "Social costs of war against terrorism in Pakistan 2002–2012," *Indian Journal of Science and Technology*, vol. 13, no. 2, pp. 127–140, 2020.
- [2] T. Mehmood, K. Rohail and K. Khan, "Cluster analysis of Pakistani terrorism events to support counterterrorism," *Societies MDPI*, vol. 8, no. 4, pp. 1–24, 2018.
- [3] I. H. Witten and E. Frank, "Data mining: Practical machine learning tools and techniques with Java implementations," *ACM Sigmod Record*, vol. 31, no. 1, pp. 76–77, 2020.
- [4] G. M. A. Soliman and T. H. M. A. E. Eniem, "Terrorism prediction using artificial neural network," *Revue d'Intelligence Artificielle*, vol. 33, no. 2, pp. 81–87, 2019.

- [5] Y. Gao, X. Wang, Q. Chen, Y. Guo, Q. Yang *et al.*, “Suspects prediction towards terrorist attacks based on machine learning,” in *5th Int. Conf. on Big Data and Information Analytics*, Kigali, Rwanda, IEEE, pp. 126–131, 2019.
- [6] A. Basuchoudhary and J. T. Bang, “Predicting terrorism with machine learning: Lessons from predicting terrorism: A machine learning approach,” *Peace Economics, Peace Science and Public Policy*, vol. 24, no. 4, pp. 1–8, 2018.
- [7] P. Agarwal, M. Sharma and S. Chandra, “Comparison of machine learning approaches in the prediction of terrorist attacks,” in *Twelfth Int. Conf. on Contemporary Computing (IC3)*, Noida, India, pp. 1–10, 2019.
- [8] E. Ahishakiye, O. E. Omulo, D. Terewana and I. Niyonzima, “Crime prediction using decision tree (J48) classification algorithm,” *International Journal of Computer and Information Technology*, vol. 6, no. 3, pp. 188–195, 2017.
- [9] H. Kang, “Prediction of crime occurrence from multimodal data using deep learning,” *PLOS ONE*, vol. 12, no. 4, pp. 1–19, 2017.
- [10] S. Rasheed and N. Khalid, “A study of assorted data on suicide bombings in Pakistan,” *Special Issue: Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, vol. 16, no. 3, pp. 1–10, 2016.
- [11] S. S. R. Rizvi, S. Abbass, A. Khan and M. Asad, “Optical character recognition system for Nastalique Urdu-like script languages using supervised learning,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 10, pp. 1–32, 2019.
- [12] T. T. Nguyen, A. Hatua and A. H. Sung, “Building a learning machine classifier with inadequate data for crime prediction,” *Journal of Advances in Information Technology*, vol. 8, no. 2, pp. 141–147, 2017.
- [13] L. G. A. Alves, H. V. Ribeiro and F. A. Rodrigue, “Crime prediction through urban metrics and statistical learning,” in *Physica A*. Amsterdam, Netherlands: Elsevier, pp. 435–433, 2017.
- [14] S. A. Azizan and I. A. Aziz, “Terrorism detection based on sentiment analysis using machine learning,” *Journal of Engineering and Applied Sciences*, vol. 12, no. 3, pp. 691–698, 2017.
- [15] F. L. Greitzer, M. Imran, J. Purl, E. T. Axelrad, Y. Mang *et al.*, “Developing an ontology for individual and organizational sociotechnical indicators of insider threat risk,” in *11th Int. Conf. on Semantic Technology for Intelligence, Defense, and Security*, Fairfax, VA, USA, pp. 1–6, 2016.
- [16] K. Singh, A. S. Chaudhary and P. Kaur, “A machine learning approach for enhancing defence against global terrorism,” in *Twelfth Int. Conf. on Contemporary Computing (IC3)*, Noida, India, pp. 1–5, 2019.
- [17] M. S. Gerber, “Predicting crime using Twitter and kernel density estimation,” *Decision Support Systems*, vol. 61, pp. 115–125, 2014.
- [18] M. Lim, A. Abdullah, N. Z. Jhanjh and M. Supramaniam, “Hidden link prediction in criminal networks using the deep reinforcement learning technique,” *Computers MDPI*, vol. 8, no. 8, pp. 1–13, 2019.
- [19] F. Schwenker and E. Trentin, “Pattern classification and clustering: A review of partially supervised learning approaches,” *Pattern Recognition Letters*, vol. 37, no. 4, pp. 4–14, 2014.
- [20] S. Naz, A. I. Umar, R. Ahmad, S. B. Ahmed, S. H. Shirazi *et al.*, “Offline cursive Urdu–Nastaliq script recognition using multidimensional recurrent neural networks,” *Neurocomputing*, vol. 177, no. 10, pp. 228–241, 2016.
- [21] I. Ahmed, R. Ali, D. Guan, Y. K. Lee, S. Lee *et al.*, “Semi-supervised learning using frequent itemset and ensemble learning for SMS classification,” *Expert Systems with Applications*, vol. 10, pp. 45–52, 2014.
- [22] J. Leng, Q. Chen, N. Mao and P. Jiang, “Combining granular computing technique with deep learning for service planning under social manufacturing context,” *Knowledge-Based Systems*, vol. 143, pp. 1–36, 2017.

- [23] L. Jiang, C. Li, S. Wang and L. Zhang, "Deep feature weighting for naïve Bayes and its application to text classification," *Engineering Application of Artificial Intelligence*, vol. 52, no. 7, pp. 26–39, 2016.
- [24] K. Bedjou, A. Faiçal and A. Abdelouhab, "Detection of terrorist threats on Twitter using SVM," in *3rd Int. Conf. on Future Networks and Distributed Systems*, Paris, France, pp. 1–5, 2019.
- [25] T. R. Gruber, "A translation approach to portable ontology specification," *Knowledge Acquisition*, vol. 5, no. 2, pp. 199–220, 1993.