SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 5, NO. 3, DECEMBER 2022
DOI: 10.35377/saucis.05.03.879855

Research Article

Prediction of Unknown Terrorist Group Names Responsible for Attacks in Turkey

Ibrahim A. Fadel[1], Cemil Öz[2]
[1]Corresponding Author; Dept. of Computer Engineering, Sakarya University; ibrahim.fadel@ogr.sakarya.edu.tr
[2]Dept. of Computer Engineering, Sakarya University; coz@sakarya.edu.tr

Received 13 February 2021; Revised 9 June 2022; Accepted 23 August 2022; Published online 31 December 2022

Abstract

In this paper, the dataset of real incidents that occurred in Turkey between 2013 and 2017 and are regarded as acts of terrorism without any doubt, according to Global Terrorism Database (GTD) is used to predict the group names responsible for unknown attacks. Principal Component Analysis (PCA) technique was used for feature selection. A novel voting method between five classification algorithms such as Random Forests, Logistic Regression, AdaBoost, Neural Network, and Support Vector Machine was used to predict the names. The results clearly demonstrate that the classification accuracy of all classifiers studied in this paper improved when PCA was used to select features as compared to selecting features without using PCA. The prediction of terrorist group names with PCA based feature reduction and the original features is carried out and the results are compared.

Keywords: prediction, classification, GTD dataset, PCA

1. Introduction

Since the September 11 terror attack, terrorism has become a global phenomenon and terrorist attacks are leading issues today and have become a focal point of concentration for different communities in the world. The term ‘terrorism’ is defined by the Central Intelligence Agency, U.S. state department and Department of defense as “premeditated, politically motivated violence perpetrated against noncombatant targets by subnational groups or clandestine agents, usually intended to influence an audience”[1].

This phenomenon is attracting the attention of various researches belonging to different organizations such as the National Union for the Study of Terrorism and Terrorism Responses (START). START is a division of the Center for Homeland Security of Excellence at the University of Maryland [2],[3] which monitors terrorist operations in the world and puts them in an open source database called Global Terrorism Database (GTD).

GTD is the most comprehensive base of operational information on terrorism in the world. The base contains information on terrorist events around the world from the year 1970 with annual updates. Based on this database, the Institute for Economics and Peace (IEP) publishes its annual Global Terrorism Index (GTI) on terrorism that assigns ranks to nations in the world according to the impact of terrorism.

According to the 2018 global terrorism index, The Middle East, to which Turkey belongs is reported to be the most affected region by terrorism. Turkey is ranked the 12th most affected country by terrorism in the world [4].

GTD dataset is used as source of the entire information related with the terrorist attacks examined in this work. From this dataset, terrorist attacks which occurred in Turkiye between 2013 and 2017 are referenced. The most active terrorist groups that are identified based on the dataset are Kurdistan Workers' Party (Partiya Karkerên Kurdistanê PKK), Islamic State of Iraq and the Levant (ISIL), Kurdistan Freedom Hawks (Teyrêbazên Azadiya Kurdistan TAK), Revolutionary People's Liberation Party/Front (Devrimci Halk Kurtuluş Partisi-Cephesi DHKP/C), Peace at Home Council (PHC), etc. However, there are a significant number of attacks which are not claimed by any of these known terrorist groups.

The rest of the paper is outlined as follows. Section 2 presents related works that have been done in the area. Our proposed methodology is presented in Section 3, Section 4 presents experiments, and the results obtained. Section 5 presents the results and discussion. The conclusion and future works are presented in Section 6.

2. Related Work

Prediction of terrorist groups after an attack is one of the most important steps for counter terrorism. As soon as we are able to find the involved group name, we will be able to make strategies to catch the culprits.

There is no international consensus on what counts as terrorism and what is not. However, Global Terrorism Index (GTI) defines terrorism as "the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation." Terrorism can be claimed by known terrorist groups, affiliated groups or individual terrorists. Terrorist attacks can have an enormous impact on wide sections of society [5].

The risks posed by a particular terror incident signify the magnitude of the act. In addition to major risks to people such as death, injuries and abduction, terrorism has a great effect on the economy too. Considering these and other minor disorders introduced due to the act of terrorists, responsible bodies such as law-enforcement agencies, police department and homeland security in general have to act accordingly. These acts which are generally known as counter terrorism may include community-based prevention and military operations. Counterterrorism requires understanding of the dynamic nature of terrorism, identification of patterns and determination of the magnitude of an attack, and consequently prioritizing the resources. However, the availability of large volumes of data related to terror makes manual analysis unimaginable. Traditional terrorist group name prediction techniques include email tracking, telephone signal information, and social network analytics [6]. These methods rely on manual analysis and are not efficient any more mainly because of the dynamic nature of terrorist groups and their actions. As important as it is, prediction requires more intelligent techniques which are reliable and can cope with the complexities associated with each terrorist acts. As is being effective in other prediction tasks, pattern recognition and machine learning techniques can be considered as a potential solution for terrorist group name prediction too. More importantly, prediction requires more intelligent techniques which are reliable and can cope with the complexities associated with each terrorist act. Communities from various fields have participated in one or more ways to provide tools that facilitate counter terrorism. Among these tools are crime category prediction [7], perpetrator prediction [6], [8], [9], [10] , geographical and socioeconomic features [11],[12], future trend prediction and risk magnitude determination [11] tools. Machine learning techniques such as classification and clustering are the core for solutions provided by computer scientists and statisticians purposely for pattern discovery and therefore determine how far threatening a given terrorist group is.

There are a limited number of prior works relying on machine learning. However, these works report low accuracy and efficiency which can be attributed to feature redundancy and non-descriptiveness of datasets. Talreja et.al [6] proposed Factor Analysis of Mixed Data on the dataset to reduce the dimension of attributes and include only the twelve most prominent features to predict the perpetrators. Tolan and Soliman [9]studied five classification algorithms for terrorism prediction in Egypt. Their paper proposed mode-imputation and Litwise deletion approach to handle missing data and only six features were used based on manual, feature selection. Gohar et al [10] proposed a classification based approach for terrorist group prediction. In their work, it is stated that only seven attributes are selected for the classification task. However, there is no clear information on how these features are determined to be the most descriptive. Sachan and Roy [14] proposed a clustering based terrorist group prediction model which takes six attributes of an incident into account. Redundant features are removed and missing values are either deleted or edited based on other information sources. In their paper, there is no particular feature selection method, rather weight is assigned to each attribute representing its importance. The most

important features and their weights are determined through trial and error.Fatih et.al [15] proposed a crime prediction model that identifies and clusters incidents based on the similarity of attacks and attributes. Selection of the most important features involves the intervention of a domain expert. Python et al. [11] developed a model to predict terrorist attacks by training various machine learning models using data from GTD terrorist attacks committed between 2002 and 2016. They focused on geographic and socio-economic features to predict attacks on separate spatial time periods. Their results were impressive but inaccurate due to the wide geographical division because each region may contain different terrorist groups with different goals. In the study by Buffa et al. [12], the study target area was divided into hexagonal-grid cells of 25 square kilometers. They used five machine learning models and four spatial statistics to assess the validity of the results and improve inferences for the spatial processes between terrorist attacks. This analysis resulted in a Random Forest model that achieves an accuracy of 0.99 in predicting the presence or absence of terrorism, with a spatial accuracy of about 5 km. The results were validated by strong F1 and mean accuracy scores of 0.96 and 0.97, respectively. Inspired by the effectiveness of feature selection and dimension reduction, in this work, the redundant and non-descriptive features are identified and removed from the original dataset before training a given supervised learning algorithm. We proposed the application of PCA for feature selection and dimension reduction. The classification accuracies obtained from the resulting feature sets are better than state- of the- art methodologies which rely on manual feature selection.

In GTD dataset, there are 1281 total incidents that occurred in Turkey between 2013 and 2017. In some cases, there is still an ambiguity to describe such incidents as terrorist attacks. In this work, we consider incidents where there is essentially no doubt as to whether they are acts of terrorism. Such incidents total to 890. From this total, 718 of them are identified by the entity that implemented them, which from now onwards are referred to as Known attacks. For example, an assailant opened fire on civilians celebrating the 2017 New Year outside Reina restaurant in Istanbul, At least 39 were killed and 69 injured including foreign tourists [16]. The assailant was arrested later and ISIL claimed responsibility of the attack.

Despite the terrorist groups' claim of their terrorist operations and their actions, there are still many terrorist attacks whose perpetrators have remained unknown. The remaining 172 instances (20%) of the 890 terrorist attacks in Turkey between 2013 and 2017 are not claimed by any group name, and as a result are called Unknown attacks. For example, the assailants opened fire on Ufuk Cafe in Istanbul in 02/01/2016, two people lost their lives and five others sustained injuries from the attack. No group claimed responsibility for the incident [17].

In this paper, we use classification algorithms to predict the names of the groups that might be responsible for such attacks. In order to predict the names of unknown attacks, we use the existing description of the known attacks. The GTD dataset describes each attack with 132 features, including the date and location of the incident, the weapons used, the nature of the target, and the number of casualties, etc. Some of the features in the dataset are redundant, consequently, they have no role in improving the prediction accuracy. At the same time, there are features which are completely irrelevant for the classification to be done. Such features need to be identified and removed before proceeding to the next stage. The process of determining and removing redundant, irrelevant features is termed as feature dimension reduction.

There are two principal algorithms for dimensionality reduction: Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The basic difference between these two is that LDA uses information of classes to find new features in order to maximize its separability while PCA uses the variance of each feature to do the same [18]. The idea behind PCA is simply to find a low-dimension set of axes that summarize data. [19] found that PCA can improve the predictive performance of machine learning algorithms in the classification of high dimensional data. In this paper, we use PCA to reduce the dimension of these features. We also evaluated the prediction accuracy to assess how representative the remaining features are.

The main contributions of this research are as follows:

1. Presents a tool that can assist law enforcement personnel to take advantage of the historical information of terrorist operations to identify potential terrorist organizations that may be the perpetrators of new attacks through the characteristics of their previous attacks. This will help the authorities to gain time in order to take the necessary measures to arrest the perpetrators quickly.
2. Predict the names of the terrorist groups that carried out a number of terrorist attacks in Turkey, which were classified as unknown.
3. Demonstrates the application of PCA to reduce data, limit dimensions and the utility of improving the performance of a variety of automated learning methods. In previous works, a certain number of attributes are selected and a focus is placed on the one-way learning method.
4. Describes a method of rating the classification of materials based on the difference between the total probability of the class assigned in all the algorithms compared to the total probability of other items, which means selection based on the collection of several different algorithms.

3. Proposed Technique

The proposed framework consists of several phases. These phases (shown in Figure 1) include: preprocessing to clean the dataset, splitting the dataset in two (known dataset: containing attacks with known responsible group name and the unknown dataset: containing attacks with unknown responsible group name). PCA is applied on a known dataset to select the most important features describing a given class. These features are used later by classification algorithms to build the prediction model. The resulting features from PCA from the known dataset is used for unknown dataset too. Then the predictor model was employed to predict the names.

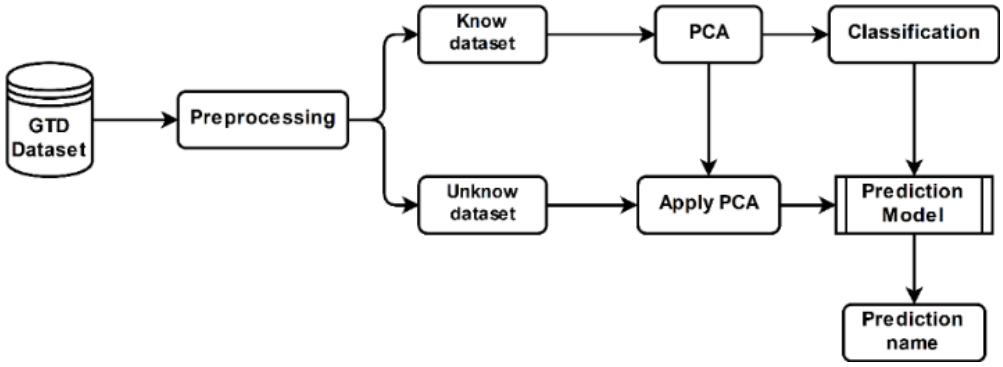

Figure 1 Proposed technique framework

3.1 Dataset Pre-processing

Pre-processing is an essential step performed on the data set to make data more suitable for mining. In the examined dataset in this work, each terrorist incident was described by 134 attributes. Dataset preprocessing was carried out manually. Some attributes have been removed and some attribute values are grouped based on some specified conditions. The result of this procedure is a dataset without any missing values.

The criteria employed to remove the attributes is given as follows.

- The dataset contains a lot of missing values, so the attributes whose fields contain more than 50% missing values are deleted. 69 attributes were deleted accordingly.
- The attributes that contain additional relevant details about the attack such as: *summary* (a brief narrative summary of the incident), *addnotes* (more information that was not contained in any of the dataset fields), *location* (information specifying the location where the incident took place), *latitude*, *longitude*, ... etc.
- The attributes representing data collection methods about an attack such as the *Dbsource* attribute (identify the original data collection effort), *scite1*, *scite2*, and *scite3* (cites the various sources from which the incident information was compiled).
- Attributes which we found that the data fields are fixed for all fields. Because the dataset we selected includes only events that occurred in Turkey and were classified as terrorist acts. These attributes include: *country* (Country name), *region* (the regions which the country is located in) *crit1*, *crit2*, *crit3* (variables that show that the inclusion criteria are met), and *individual* (indicates person(s) who carried out the attack but do not belong to a known terrorist group or organization).
- Attributes containing subcategories of major classifications like *weapsubtype1_txt* which shows the type of weapon used to carry out the attack.
- The features containing the number of casualties such as: *nkill* (number of killed), *nkillus* (The number of U.S. citizens killed), *nkillter* (number of perpetrators killed), *nwound* (number of wounded), *nwoundus* (The number of U.S. citizens wounded), and *nwoundte* (number of perpetrators wounded). These attributes were excluded, which is a result of the terrorist operation and has no direct impact on the name of the entity that carried out the operation.

The values of some attributes (types of instances) have been grouped as follows:

- *Day* and *month* attributes were integrated into one attribute called season and the values were distributed according to the four seasons. By closely analyzing the dataset, we note that the intensity of terrorist operations increases in the summer and the beginning of the fall (in the months of July, August and September) and decrease in winter and early spring (December until March).
- The attribute *provstate*, describe name of a place. The value of this field contains 43 city names and these cities are merged according to the region to which they belong. Turkey's provinces are distributed into 7 major regions.
- Attributes *targtype1_txt* (captures the general type of target), *attacktype1_txt* (captures the general method of attack type), and *gname* (name of the group that carried out the attack) were partially merged. Table 1 describes the merged values of each attribute.
- Attribute *natlty1_txt* which includes nationality of the target that was attacked, it merged to 3 categories: Turkey, Foreign, and international.

3.2 Principal Component Analysis (PCA)

PCA is a technique used for data compression and feature extraction. Its main purpose is to analyze data to identify and find patterns in order to reduce the dimensions of the dataset into fewer dimensions which act as summaries of features with minimal loss of information [20].

Assume that $X = x_1, x_2, \dots, x_n$ is a dataset consisting of n dimensional data vectors, our goal is to scale down this n-dimensional dataset to a k-dimensional subspace (where $k < n$).

Table 1 Merged attribute values

Attribute	Category	#	New Category
targtype1_txt	Government (General)	59	Government
	Government (Diplomatic)	4	
	Military	59	Military and Police
	Police	275	
	Private Citizens & Property	203	Private Property
	Business	70	
	Educational Institution	46	Institutions
	Journalists & Media	17	
	NGO	2	
	Tourists	2	
	Religious Figures / Institutions	9	
	Utilities	18	
	Transportation	18	Transport
	Airports & Aircraft	2	
	Unknown	66	Other
	Other	4	
attacktype1_txt	Hostage Taking (Kidnapping)	60	Hostage
	Hostage Taking (Barricade Incident)	2	
	Hijacking	2	
	Unarmed Assault	2	Other
	Unknown	29	
gname	Turkish Communist Party/Marxist (TKP-ML)	2	Other
	Fetullah Terrorist Organization	1	
	Peoples' United Revolutionary Movement (HBDH)	1	
	The Independent Military Wing of the Syrian Revolution Abroad	1	
	Maoist Communist Party (MKP)	1	
	Free Syrian Army	1	
	People's Defense Unit (Turkey)	1	

The n-dimensional mean vector μ is

$$\boldsymbol{\mu} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{x_i} \tag{1}$$

The covariance matrix (Cov) of the dataset is:

$$\boldsymbol{Cov} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^{\boldsymbol{T}} \tag{2}$$

The eigenvalues and eigenvectors of the covariance matrix are calculated using

$$\boldsymbol{Cov_{vi}} = \boldsymbol{\lambda_i v_i} \tag{3}$$

Where λ = Eigenvalue, ν =Eigenvector.

Let $\Lambda = diag[\lambda_1, \lambda_2, \dots, \lambda_k]$. Λ is a diagonal matrix of eigenvalues. The matrix V contains the eigenvectors $V = [v_1, v_2, \dots v_k]$ and the is orthonormal $V^T V = I_k$. Where, I_n is the $k \times k$ identity matrix.

Sort the eigenvectors by decreasing eigenvalues ($\lambda_i \geq \lambda_{i+1}$)

Let the matrix W = $[v_1, v_2, \dots v_k]$, contain the first k eigenvectors

The low dimensional feature vector of a new input data is determined by

$$\boldsymbol{y} = \mathbf{W}^{\boldsymbol{T}}.\boldsymbol{x} \tag{4}$$

When we applied the above equations to our dataset, from 51 principal components (PC), we found 31 components that are most contributing principal components. This contributes 95.9% of eigenvalues (Figure 2).

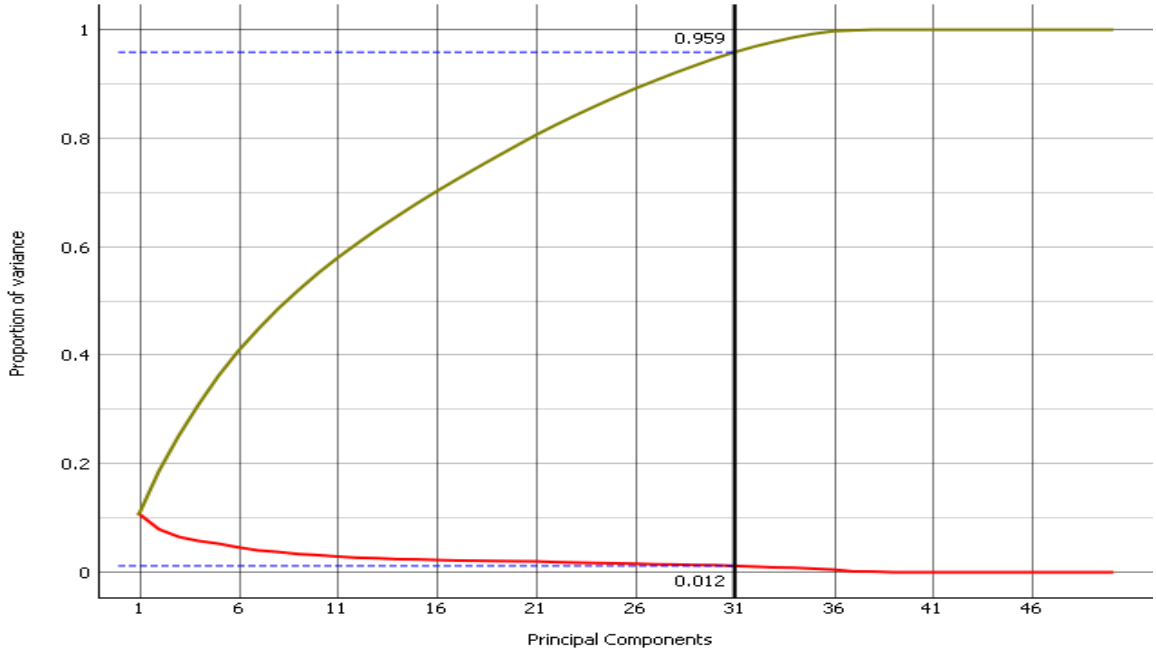

Figure 2 Principal components variance

Principle components also make distinction between classes much clearer. (Figure 3) shows the plot of the data points using only PC1 and PC7 principal components. It might be proper to use class tags to visualize the dataset as well. Looking at the graph, we can see that only two feature variables can be used to visualize the whole dataset properly as compared to having thirteen feature variables. Although the classes are not well differentiated but it helps to reduce the feature set without the loss of much information.

4. Classification Algorithms

Classification and prediction are the prominent approaches for data mining in various fields. They are predictive models that predict the future trends based on some training datasets.

In the classification phase, a method of voting between 5 different machine learning algorithms was used. These algorithms are Random Forests, Logistic Regression, Adaptive Boosting (AdaBoost), Neural Network, and Support Vector Machine (SVM).

- **SVM** is a widely used machine learning algorithm. It can also be employed for both classification and regression purposes. The main idea of SVM is to construct maximum-margin hyperplane between any class data point within the training set, this can give a greater chance of new data being classified correctly [21].

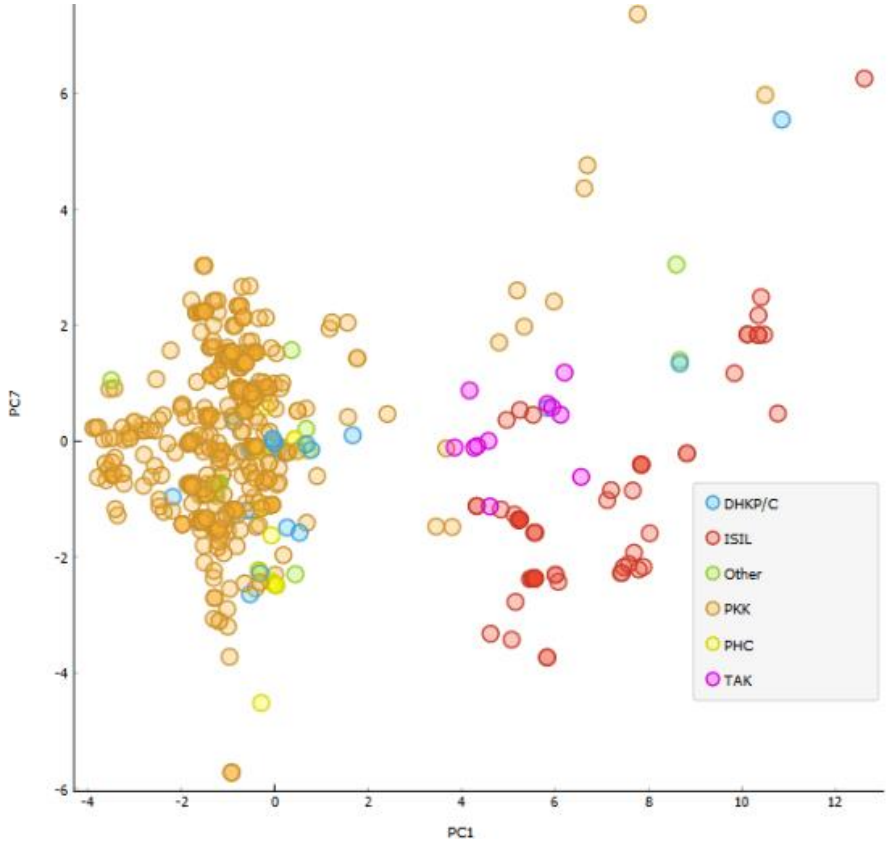

Figure 3 Clustering based on PC1-PC7

- **Random Forests** is an algorithm that is used for both classification and regression tasks by creating a forest and making it random. The created "forest", is an ensemble of Decision Trees. The forest is trained using "bagging" method in most cases. Bagging is a combination of learning models to improve the overall result. Bagging can also be used for both classification and regression problems [22].
- **Logistic regression** is a probabilistic, statistical method for classifying data into discrete outcomes. It is named as 'Logistic Regression' because its underlying technique is quite the same as Linear Regression. But the biggest difference lies in what they are used for. Linear regression algorithms are used to predict/forecast values, but logistic regression is used for classification tasks [23].
- **AdaBoost** is an ensemble classifier that attempts to create a strong classifier from a number of weak classifiers. When used in conjunction with various types of learning algorithms, performance is improved. A weighted sum of the outputs of the 'weak learners' is used to represent the final output of the boosted classifier [24].
- **Neural Network** is a combination of units called neurons that are arranged in layers. These neurons convert an input vector into some output. Each neuron takes an input, applies some function (often nonlinear) on it and then passes its output to the next layer. In more general terms, neural networks are applied in a feed-forward fashion, i.e. Each layer forwards all its output to the next layer without feedback to the previous layer(s). A weighting mechanism is applied to the signals passing from one neuron to another. During the training phase, the weights are tuned so that the neural network adapts to a particular problem at hand [21].

All these algorithms were assembled in a voting algorithm to select the appropriate class. The voting is carried out according to the following formula

Assume we have k number of classes represented by C,

$$\boldsymbol{c_j} \in \boldsymbol{C}, \qquad \boldsymbol{j} = \{\boldsymbol{1}, \dots, \boldsymbol{k}\} \tag{5}$$

and n classification algorithms represented by A

$$\boldsymbol{a_i} \in \boldsymbol{A}, \qquad \boldsymbol{i} = \{\boldsymbol{1}, \dots, \boldsymbol{n}\} \tag{6}$$

the probability p of the class c_j in the Voting algorithm V is the average of the probabilities of these classes in the classification algorithms

$$\boldsymbol{p(V, c_j)} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{p(a_i, c_j)} \tag{7}$$

The predicted class "$Pred_c$" or the voted class "V_c" in the voting algorithm is the class with the max probability.

$$\boldsymbol{V_c} = \boldsymbol{max}\,\{\boldsymbol{p(V, c_1)}, \boldsymbol{p(V, c_2)}, \dots, \boldsymbol{p(V, c_k)}\} \tag{8}$$

Let us consider 2 algorithms (a_1 and a_2) and 3 classes (c_1,c_2, and c_3) as given in the Table 2.

Table 2 Voting algorithms

Classification Algorithm 1				**Classification Algorithm 2**				**Voting Algorithm**			
$p(a_1, c_1)$	$p(a_1, c_2)$	$p(a_1, c_3)$	$Pred_c$	$p(a_2, c_1)$	$p(a_2, c_2)$	$p(a_2, c_3)$	$Pred_c$	$p(V, c_1)$	$p(V, c_2)$	$p(V, c_3)$	V_c
0.53	0.07	0.4	c_1	0.45	0.03	0.62	c_3	0.44	0.05	0.51	c_3
0.05	0.15	0.8	c_3	0.01	0.10	0.89	c_3	0.03	0.125	0.845	c_3
0.58	0.32	0.1	c_1	0.18	0.77	0.05	c_2	0.38	0.545	0.075	c_2

4.1 Performance measurement

Accuracy, Precision, Recall and F1-score metrics generally are used to evaluate the performance of different classification algorithms [25].

Since the dataset is multiple class, averaging the evaluation measures can give a view of the general results. There are two names to refer to averaged results: micro-averaged and macro-averaged results.

- In the Micro-average method, a sum of the individual true positives, false positives, and false negatives of the system for different sets is obtained to get the statistics about them.
- In Macro-average, the average of the precision and recall of the system on different sets is taken.

Since there is no balance between the number samples for each class in the dataset, (PKK represents 82.3% of the class while ISIL = 9.9, DHKP/C = 2.8, PHC=2.7 TAK = 1.5% and Other = 1.1), Micro-average is preferable if there is a class imbalance problem [26].

5. Results and Discussion

Each classifier was employed on known datasets both before and after feature dimension through PCA. The accuracy was estimated with 10-fold cross validation. The performance comparison of the classification learners on the two datasets is shown in Table 3. The results clearly show that PCA based feature dimension reduction led to an improved accuracy in all algorithms.

The comparative analysis based on results obtained using the proposed approach to that of other literature using GTD dataset is shown in Table 4.

Table 3 Accuracy results of selecting featurs usnig PCA (⊕) and without using PCA (⊖)

Algorithm	**Accuracy**		**F1 Micro-average**	
	⊕	⊖	⊕	⊖
AdaBoost	92.5%	90.8%	96.1%	95.2%
Logistic regression	90.7%	88.2%	95.1%	93.7%
Neural Network	91.9%	90.4%	95.8%	95.0%
Random Forests	92.2%	90.5%	95.9%	95.0%
SVM	88.3%	86.9%	93.8%	93.0%
Voting	92.9%	91.2%	96.3%	95.4%

Table 4 Comparative results

		Current	**Talreja et al.**	**Mohammed and Karabatak**	**Tolan and Soliman**	**Gohar et al**
Dataset	Country	Turkey	India	Turkey	Egypt	world
	Period	2013-2017	1970-2015	2016	1970-2013	1970-2012
	Features	17	12	6	6	7
Classification algorithm	AdaBoost	92.5				
	Bayes Net			61.41		
	C4.5/J48		60	64.13	56.56	
	decision stump (DS)					84.97
	ID3				26.01	91.30
	KNN			51.1	73.03	83.43
	Logistic regression	90.7				
	NB			58.15	69.03	92.75
	Neural network	91.9				
	Random forest	92.2	58.5			
	SVM	88.3	73.2	59.78	75.42	
	Voting	92.9				93.40

In terms of predictions of names of terrorist groups in the unknown dataset. Table 5 shows the number of predicted names for each terrorist group using the algorithms in both cases. Note the difference in the number of operations per group depending on the algorithm and the case. However, all the algorithms and cases agreed to predict 0 times to the terrorist group PHC. This is logical because all their operations occurred between 15/July/2016 and 16/July/2016 when it announced its name and made a coup attempt to overthrow the government in Turkey.

Table 5 Predicted terrorist organizations' names and number of attacks using PCA (⊕) and without (⊖)

Algorithm	**PKK**		**ISIL**		**DHKP/C**		**TAK**		**PHC**		**Other**	
	⊕	⊖	⊕	⊖	⊕	⊖	⊕	⊖	⊕	⊖	⊕	⊖
AdaBoost	144	132	22	25	3	6	2	5	0	0	1	4
Logistic Regression	122	139	24	19	16	10	7	3	0	0	3	1
Neural Network	127	123	25	29	11	10	6	7	0	0	3	3
Random Forest	146	144	20	21	1	3	2	4	0	0	3	0
SVM	110	103	34	33	13	18	8	9	0	0	7	9
Voting	147	142	18	20	3	3	3	5	0	0	1	2

6. Conclusion

This paper proposes a novel method for predicting responsible terrorist groups for unknown classified incidents. The proposed method consists of PCA technique for selecting features and a voting method between 5 best classification algorithms. The experiments are conducted using the GTD data set for the last 5 years terrorist incidents that occurred in Turkey. The proposed method using PCA obtained high results based on performance accuracy as compared to a method without using PCA. Our accuracy results show a significant improvement when compared to the results obtained by (Mohammed and Karabatak 2018) which do not exceed 64.13%. Moreover, an impressive result was obtained from the voting algorithm which is dependent on the sum of the probabilities resulting from all algorithms.

We suggest that future work can focus on the folowing:

- Improving the quality of terrorism-related data, by integrating GTD data with data from local authorities and the media.
- Linking the geographical factor of the terrorist operation with the areas of activity of the terrorist organization.
- Deeper studies of the relationship between the increase in terrorist operations at certain times and its relationship to political issues in the country, or to regional or international issues.

References

[1] C.C. Aggarwal, *Data Classification Algorithms and Applications*. London, England, CRC Press Taylor & Francis Group, 2015.

[2] A. Babakura, M. N. Sulaiman and M. A. Yusut, "Improved method of classification algorithms for crime prediction". *Proc. - 2014 Int. Symposium on Biometrics and Security Tech., ISBAST 2014*, Kuala Lumpur, Malaysia, 26-27, August 2014.

[3] BBC News, "Istanbul new year Reina nightclub attack leaves 39 dead," 2018. [Online]. Available: https://www.bbc.com/news/world-europe-38481521 [Accessed: 09-Dec-2018].

[4] E. S. Chris, *The Psychology of Terrorism: Theoretical understandings and perspectives*. Volume III, Lonon, England, Praeger Publishers, 2002.

[5] Institute for Economics & Peace, *Global terrorism index 2018. Measuring the impact of terrorism*. Sydney, Australia,2018.

[6] J. Feng, H. Xu, S. Mannor and S. Yan, "Robust Logistic Regression and Classification," *Proc. - 27th Inter. Conf. on Neural Info. Processing Syst. (NIPS)*, Montréal, Canada, 08-13 December 2014.

[7] F. Gohar, W. Haider and U. Qamar, "Terrorist Group Prediction Using Data Classification," *Proc. – Inter. Conf. on Artificial Intelligence and Pattern Recognition*, Kuala Lumpur, Malaysia, 17-19 November 2014.

[8] GTD, "About the GTD," 2019 [Online]. Available: https://www.start.umd.edu/gtd/about/ [Accessed: 09-Jan-2019].

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. 2nd ed. New York, USA, Springer, 2008.

[10] T. Howley, M. G. Madden, M. L. O'Connell, and A. G. Ryder, "The effect of principal component analysis on machine learning accuracy with high-dimensional spectral data," *Knowledge-Based Syst.*, 19(5), 363–370, 2006.

[11] A. Python, A. Bender, A. K. Nandi, P. A. Hancock, R. Arambepola, J. Brandsch, T. C. D. Lucas, "Predicting non-state terrorism worldwide,". *Science Advances*, 7(21), 1-13, 2021.

[12] C. Buffa, V. Sagan, G. Brunner and Z. Phillips. "Predicting Terrorism in Europe with Remote Sensing, Spatial Statistics, and Machine Learning,". *ISPRS International Journal of Geo-Information*, 11(4), 1-12, 2022.

[13] Hurriyet, "Kahvehaneye 58 hain kurşun - Son Dakika Haberler," 2018. [Online]. Available: http://www.hurriyet.com.tr/gundem/kahvehaneye-58-hain-kursun-40048592 [Accessed: 09-Dec-2018].

[14] I. T. Jolliffe, *Principal Component Analysis*. 2nd ed. New York, USA, Springer, 2002.

[15] T. Kim, D. Park, D. woo, T. Jeong and S. Min, "Multi-class Classifier-Based Adaboost Algorithm," *Proc. - The Secd. Sino-foreign-interchange conf. on Intelligent Science and Intelligent Data Engineering*, Xi'an, China, 23 October 2011.

[16] Z. C. Lipton, C. Elkan and B. Narayanaswamy, "Optimal Thresholding of Classifiers to Maximize F1 Measure," *Proc. - European Conf. on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2014)*, Nancy, France, 15-19 September 2014.

[17] A. M. Martinez and A. C. Kak, "PCA versus LDA,". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), 228–233, 2001.

[18] T. Feakin, "Terror, Security, and Money: Balancing the Risks, Benefits, and Costs of Homeland Security," *The RUSI Journal*, 157(4) 99, 2012.

[19] F. Ozgul, Z. Erdem and C. Bowerman, "Prediction of past unsolved terrorist attacks," *Proc. - 2009 IEEE Inter. Conf. on Intelligence and Security Informatics (ISI 2009)*, Dallas/TX, USA, 26 June 2009.

[20] D. M. W. Powers, *Evaluatıon: From Precısıon, Recall And F-Measure To Roc, Informedness*. Markedness & Correlatıon. Journal of Machine Learning Technologies, 2(1), 37–63. 2011.

[21] A. Sachan and D. Roy, "TGPM: Terrorist Group Prediction Model for Counter Terrorism," *Inter. Journal of Comp. Appl.*, 44(10), 49–52. 2012.

[22] M. Shermila, A. B. Bellarmine and N. Santiago, "Identity using Machine Learning Approach," *Proc. - 2nd Inter. Conf. on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 11-12,May 2018.

[23] START, "Global terrorism database codebook: Inclusion criteria and variables 2018," 2018. [Online]. Available: https://www.start.umd.edu/gtd/downloads/Codebook.pdf. [Accessed: 02-Nov-2018].

[24] D. Talreja, J. Nagaraj, N. J. Varsha and K. Mahesh, "Terrorism analytics: Learning to predict the perpetrator," *Proc. - 2017 Inter. Conf. on Advances in Computing, Communications and Informatics (ICACCI 2017)*, Mangalore India, 13, September 2017

[25] Institute for Economics & Peace. *Global terrorism index 2019. Measuring the impact of terrorism*. Sydney, Australia, 2019.

[26] G. M. Tolan and O. S. Soliman, "An Experimental Study of Classification Algorithms for Terrorism Prediction," *Inter. Journal of Knowledge Engineering-IACSIT*, 1(2), 107–112. 2015.