Prediction of Terrorism Activities Using Supervised Machine Learning Techniques

Adithi.M[1], Bhavana G V[2], Chaitra K[3], Shravya G[4]
[1, 2, 3, 4] Dept of Computer Science
[1, 2, 3, 4] Sai Vidya Institute of Technology, Bangalore,Karnataka.

Abstract- *The purpose of this work is to predict the area and country of a terrorist attack using approaches in machine learning. The research was carried out on the Global Terrorism Database (GTD), an open database which contains a list of terrorist activities. Eight machine learning algorithms have been applied on some selected set of features from the data set to achieve a maximum accuracy of 93%. This project provides an approach to analysing terrorism region and country with the machine learning techniques and terrorism specific knowledge to fetch conclusions about terrorist behaviour patterns. Through analysis of events using GTD, eight supervised machine learning models (Gaussian Naïve Bayes, Linear Discriminant Analysis, k-Nearest Neighbours, Linear Regression, Support Vector Machines, Decision Tree, Random Forest and Logistic Regression) were built and evaluated on their performances.*

Keywords- Terrorism; prediction; machine learning; accuracy

I. INTRODUCTION

Terrorist attacks are spreading on a great pace across the world. According to United Nations definition of Terrorism," Any action with a political goal that is intended to cause death or serious bodily harm to civilians"[1]. In the last year, around 22 thousand events occurred globally, causing over 18 thousand casualties [2]. The factors leading to terrorism change over time since they are dependent upon multiple political and social reasons.

Apart from predicting reason behind the attack, identification of the responsible agencies is also difficult. There has been a dearth of the information regarding patterns of widespread terrorist behaviour. The existing analyses are either case studies or use of quantitative methods such as regression analysis.

The former of these is specific to certain events, while the latter approach is restricted to interviews of civilians impacted by the attack. Most of these analyses depend on factors such as weapons used for the attacks and number of people harmed. Other type of analysis includes investigation of unusual patterns in individual behaviors or questioning detainees to acquire data pertaining to the attacks.

The current research is focused on finding out the correlation between terrorism and its causal factors. Existing efforts have not been good enough for prediction. Machine learning approaches help predict the likelihood of a terrorist attack, given the required data. The results of this work can help the security agencies and policy makers to eradicate terrorism by taking relevant and effective measures.

This paper provides an approach to analyzing terrorism region and country with the machine learning techniques and terrorism specific knowledge to fetch conclusions about terrorist behavior patterns. Through analysis of events using GTD, eight supervised machine learning models (Gaussian Naïve Bayes, Linear Discriminant Analysis, k-Nearest Neighbours, Linear Regression, Support Vector Machines, Decision Tree, Random Forest and Logistic Regression) were built and evaluated on their performances.

II. LITERATURE SURVEY

The large number of events makes it difficult to predict terrorist group responsible for some terrorist activity.

The work in [1] has tested machine learning approaches for classifying and analyzing global terrorist activity. The authors have explored supervised machine learning approaches to study terrorist activity, and then developed a model to classify historical events in Global Terrorism Database.

In reference to [2], more advanced information technologies have been developed to counter terrorism domain to enhance the performance of early warning system. Machine learning based data mining can be applied to predict terrorist event hidden in terrorist attack events and by which the experts expect to get a clear picture of what the terrorists.
In reference to [3], Global terrorism means the use of intentionally indiscriminate and illegal force and violence for creating terror among masses in order to acquire some political, monetary, religious or legal goals. Identification of

these ideologies and prediction of future attacks has proven to be of the greatest importance.

This paper focuses on analyzing the historical dataset of Global Terrorism Database and predicting the factors that might give blow to rise of terrorism.

III. PROPOSED METHODOLOGY

The objective of our study is to predict the region and country of terrorist attacks.

Dataset and Features The Global Terrorism Database (GTD) is a database which is open source and includes information on terrorist events for the years 1970-2017. It includes wholesome data regarding domestic incidents, transnational and international terrorist incidents which took place in this duration.

The number of cases included is 180,000 {bombings (88,000), assassinations (19000) and kidnappings (11000)}. The parameters include date of incident, month of attack, location of incident, and country of incident, region of incident the weapons used in the incident, nature of the target, type of attack, the number of casualties, the group or individual responsible for the incident.

The data source for GTD has been a variety of open media sources, more than 4,000,000 news articles and 25,000 news sources. It has been considered the most comprehensive unclassified database on terrorist attacks in the world.

B. Exploratory Data Analysis Before building the model and to gain high level understanding of dataset features we performed some exploratory data analysis.

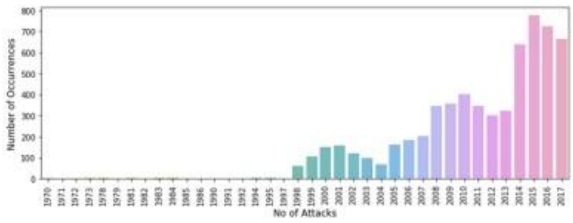

Fig. 1. Number of Yearly Terrorist Attacks

Fig. 1 depicts a significant increase in number of terrorist attacks from 2008 and reaches the peak in 2015.

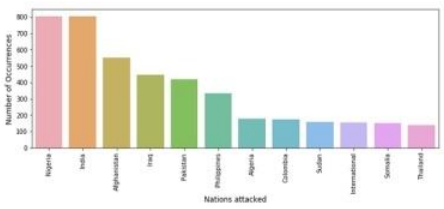

Fig. 2. Number of Nation-wise Attacks

The exploration results that Nigeria has faced maximum number of terrorist attacks across the world while India holds the second position of terrorist attacks faced (Refer Fig. 2).

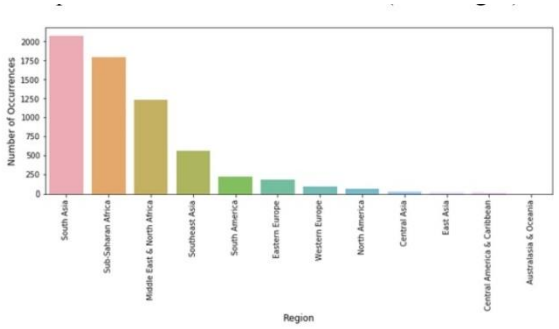

Fig. 3. Regions Attacked

The graph in Fig. 3 gives the overview of regions that are targeted by terrorists; South Asia being the top targeted region and Africa being the second across other regions.

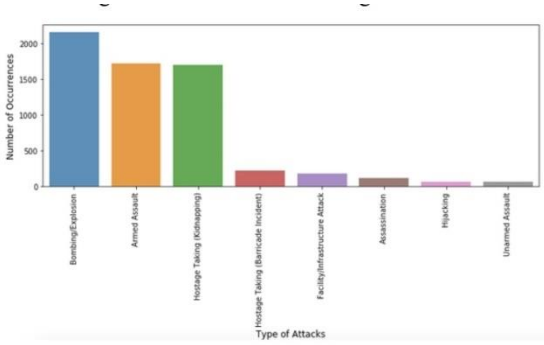

Fig. 4. Type of Attacks

The type of attacks mostly happened from 1970 to 2017 are widely conducted by Bombing/explosions (Fig. 4). It is interesting to see that armed assault and Kidnapping are the type of attacks that widely used by terrorists after Bombing and explosion.

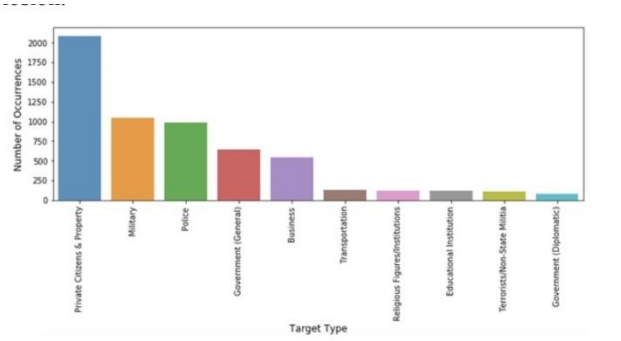

Fig. 5. Target Type

Citizens are targets among the targets and military/armed forces are the second most favorable target of the terrorists' attacks from 1970 to 2017, as depicted by Fig. 5.

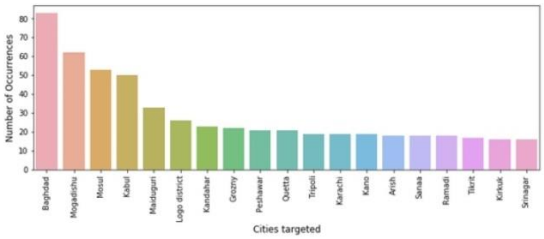

Fig. 6. Target Cities

Baghdad has received the most terrorist attacks in the world and Srinagar comes on 20th number of cities that have received terrorist attacks, as may be observed from Fig. 6.

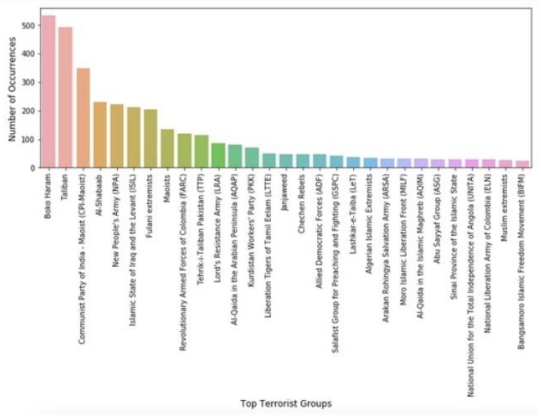

Fig. 7. Terrorist Groups

Boko Haram is the terrorist group which has conducted maximum number of terrorist attacks, Taliban being the second and Maoists in India are on the third number of top terrorist groups (Fig. 7).

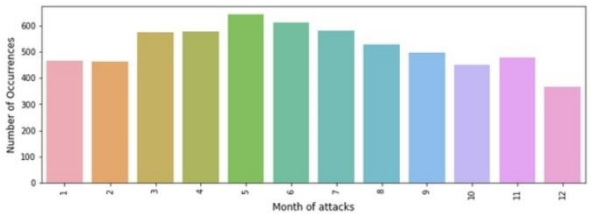

Fig. 8. Month of Attacks

May and June are the months that faced terrorist attacks most frequently from 1970 to 2017, as can be seen in Fig. 8. It can be observed from Fig. 9 that the most common date of terrorist attacks is 16th while 13th and 28th are the 2nd and 3rd most frequent dates of terrorist attacks.

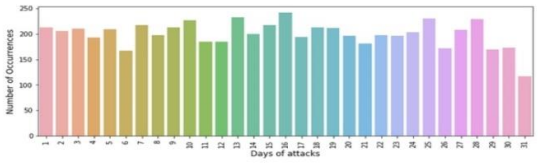

Fig. 9. Day of Month

IV. MACHINE LEARNING

After exploratory data analysis on GTD (Global Terrorism Database), we further applied six supervised machine learning models (Gaussian Naïve Bayes, Linear Discriminant Analysis, k-Nearest Neighbors, Support Vector Machines, Decision Tree, Linear Regression ,Random Forest and Logistic Regression) for prediction and subsequently evaluated their performances. We have considered month of attack, Target Type1 and attack type1 to predict region and country.

Gaussian Naive Bayes: Naive Bayes classifier has been considered as one of the simplest supervised approaches. In this, Bayes theorem provides a way to calculate probability of hypothesis (given prior information), hence the presence of one feature does not affect the presence of another feature. The advantage of NB is it can be easily trained with small and large and the execution time is relatively fast.

Linear Discriminant Analysis: LDA is also based on Bayes' Theorem. But instead of directly calculating posterior probability, it estimates multivariate distribution of its distribution. If we see its mathematical aspect, the algorithm does training by first setting the linear combination of predictors (features) that is helpful in separating different classes. The predicted class is classified by detecting the

training samples which falls into linear decision boundaries. The advantage of LDA is it always produces an explicit solution and is feasible due to its , but suffers from the assumption that linear separability is achievable in all classifications.

K-Nearest Neighbors: k-NN is another algorithm commonly used for supervised classification problems. First introduced in 1951, the algorithm aims to identify homogeneous subgroups such that observations in the same group (clusters) are more similar to each other than others. Each data points' k-closest neighbors are found by calculating Euclidean or Hamming distance and grouped into clusters. The k-closest data points are then analyzed to determine which class label is the most common among the set. The most common class is then classified to the data point being tested. For k-NN classification, an input is classified by a majority vote of its neighbors. That is, the algorithm obtains the classification of its k neighbors and outputs the class that represents a majority of the k neighbors.

Support Vector Machines: In machine learning, they basically under the category of supervised learning which analyzes the data used for classification and regression analysis. SVM model is a representation of points in space, mapped properly so that the categories get divided by a wide gap. If new examples are mapped, then they fall accordingly into the right side of the gap.

Logistic Regression: Logistic regression is the machine learning approach used as regression analysis when the dependent variable is binary. It is also a type of predictive analysis. It basically describes the relationship between one dependent binary variable and one or more independent variables (which can be ordinal, nominal, interval or).

Decision Trees: Decision trees classifiers applies questions and conditions in a tree structure. This approach applies decision rules inferred from the data features to predict the value of target variable and create model accordingly. The condition for categorization is included in the root and internal nodes. Inputs are entered at the top and tree is traversed down, following the branches. Once the input node reaches the terminal node, a class is assigned. The advantage of decision trees is that they can be easily handled continuous and discrete data. When the training set is small in comparison with the number of classes, it also leads to higher classification error rate, hence causing outfitting.

Random forest: An RF is simply a collection of decision trees. The random forest starts with training many different decision trees and combining them into an ensemble , the "forest". Then ,when classifying an unknown data point, each decision tree will test the observation and vote on which class it believes the observation to be. By majority vote, the random forest will output the most likely classification.

Linear Regression: The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.The linear equation inserts a scale factor to each input column or value, called a coefficient and represented by the ca Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.For example, in a simple regression problem (a single x and a single y), the form of the model would be: $y = B0 + B1*x$.

Logistic Regression: It's a classification algorithm, that is used where the response variable is *categorical*. The approach of Logistic Regression is to find a relationship between features and probability of particular outcome. E.g. When we have to predict whether a student will pass the exam or not , when the number of hours spent studying is given as a feature, the response variable has two values, pass and fail. This type of a problem is know as Binomial Logistic Regression, where the response variable 0 and 1 have two values 3or pass and fail or true and false. Multinomial Logistic Regression deals with situations where the response variable can have three or more possible values.

V. RESULTS

Results of test sets are listed in the following tables, the results shows that Random forest has a higher training accuracy and SVC gives least training accuracy.

	Accuracy	Recall	Precision
Random Forest	93.00	0.983185	0.934986
K-Nearest Neighbors Algorithm	81.86	0.851764	0.933632
Linear Regression	86.88	0.868800	0.868800
Linear Discriminant Analysis	86.89	0.868480	0.868480
Decision Tree Classifier	90.00	0.886069	0.886069
Naive Bayes	86.90	0.868865	0.868865
Support Vector Classifier	84.67	0.846667	0.846667
Logistic Regression	87.00	0.868865	0.868865

VI. CONCLUSION

After training our models on the month variables, Traget type, Attack type and many more independent features

to predict the success of Attack. It is estimated that Random forest offer 93 percent accuracy in on predicting success of Attack. The findings of the presented project can be used in the coming times to improve security against terrorist attacks.

REFERENCE

[1] S. . Naïve Bayesian, from Predicting the Future.[Online],

[2] E. and J.. Discriminatory analysis: discrimination: Consistency properties. PsycEXTRA , (1951)

[3] https://cogsci.yale.edu/sites/default/files/files/Thesis2018 Peng.pdf

[4] D. Y., & , M. (2018, March). Terrorist attacks in Turkey: of terrorist acts that occurred in 2016. In 2018 6th International Symposium on Digital Forensic and Security (ISDFS) (pp. 1-3).

[5] Bang, J., , A., David, J., & , A. (2018). Predicting terrorism: a machine learning .

[6] Mathews, T., & Sanders, S. (2019). Strategic and experimental analyses of conflict and terrorism. Public Choice, 179(3-4), 169-174.

[7] I. (2019). Terrorism, religion and self-control: An unexpected connection between conservative religious commitment and terrorist efficacy. Terrorism and Political Violence, 1-16

[8] Khalifa, N. E. M., Taha, M. H. N., Taha, S. H. N., & Hassanien, A.E. (2019, March). Statistical Insights and Association Mining for Terrorist Attacks in Egypt. In International Conference on Advanced Machine Learning Technologies and Applications (pp. 291-300). Springer, Cham

[9] Wheatley, W., Robbins, J., Hunter, L. Y., & Ginn, M. H. (2019). Terrorism's effect on Europe's centre-and far-right parties. European Political Science, 1-22.

[10] Klenka, M. (2019). Major incidents that shaped aviation security. Journal of Transportation Security, 1-18.

[11] Guo, W., Gleditsch, K., & Wilson, A. (2018). Retool AI to forecast and limit wars.

[12] Hao, M., Jiang, D., Ding, F., Fu, J., & Chen, S. (2019). Simulating Spatio-Temporal Patterns of Terrorism Incidents on the Indochina Peninsula with GIS and the Random Forest Method. ISPRS International Journal of Geo-Information, 8(3), 133.