Fuzzy Systems and Data Mining VI
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA200701

Machine Learning and Statistical Analysis Techniques on Terrorism

Rajesh P[a,1], Babitha D[a], Mansoor Alam[b], Mansour Tahernezhadi[b], Monika A[c]
[a]*K L University, India*
[b]*Northern Illinois University, United States*
[c]*Sanjivani College of Engineering, Kopargaon*

Abstract. Terrorism is a major issue facing the world today. It has negative impact on the economy of the nation suffering terrorist attacks from which it takes years to recover. Many developing countries are facing threats from terrorist groups and organizations. This paper examines various terrorist factors using data mining from the historical data to predict the terrorist groups most likely to attack a nation. In this paper we focus on sampled data primarily from India for the past two decades and also consider International database. To create meaningful insights, data mining, machine learning techniques and algorithms such as Decision Tree, Naïve Bayes, Support Vector Machine, Ensemble methods, Random Forest Classification are implemented to analyze comparative based classification results. Patterns and predictions are represented in the form of visualizations with the help of Python and Jupyter Notebook. This analysis will help to take appropriate preventive measures to stop Terrorism attacks and to increase investments, to grow the economy and tourism.

Keywords. Data Mining, Classification, Global Terrorism Database (GTD), Machine Learning.

1. Introduction

Terrorism has become a relatively influential factor in the International Politics. A major terrorist cause is an aggrieved organization that resorts to violence for various factors such as political, cultural, religious, etc. This research focuses on the global spread of terrorism as well as in India in the last two decades. From the literature review, it is identified that there are several statistical analyses performed on terrorism database.

Due to terrorism, common people are getting fear, nervousness, and ambiguity on maximum scale of community rather than single individual. According to Statistics on GTD database,2019 contains 1411 various terrorist attacks have occurred, causing 6362 fatalities, deficiently disturbing quality of life of human individuals in society. It is similarly important to understand statics, analyzing terrorist events data, to create awareness, to assist kind of people, take preventive measures not occurring those events in future.

Terrorism attacks has been considered major impact among all nations for decades to identify factors reason to perform of terrorism or to carry out counterterrorism, social

[1] Corresponding Author: RAJESH P, K L University, India. Email: rajesh.pleti@gmail.com

and fiscal effects of terrorism. Due to terrorism complexity of problem in nature, it is difficult to find efficient solution to protect lives of individuals. Classification of terrorist philosophy, forecast of future terrorist attacks have been demonstrated to be of immense importance, time consuming process. We tried to identify interesting insights using machine learning algorithms on GTD database.

This paper is set out as follows. Section 2 illustrates literature review. The pre-processing performed on the Global Terrorism Database, feature selection, analytical tool used for the data mining and the software used for the study was described in detail in Section 3. Section 4 provides a detailed description of results and experimental analysis exploits machine learning algorithms such as Decision Tree, Naïve Bayes, Support Vector Machine and Random Forest Classification to build a classification model. Finally, Section 5 summarizes the insight research findings and future works.

Terrorist attacks got mainly significant imperative issue to all humankind in the world. These kind attacks are more in developing countries like India, Middle East, North Africa, and South Asia. This paper describes insight hidden patterns using machine learning algorithms in GTD Database and how it has grown over time of decades, social impacts, growth of economy and impacts on tourism, and to take proper preventive measures to eradicate the terrorism. It also predicts the region, country, number of terrorist attacks by machine learning approaches.

The Objectives of this paper can be described as follows.

- To perform statistical data analysis on the GTD data to obtain hidden pattern and insights.
- To identify the most terror-prone regions and the weapons mostly used for attack in both India and around the world.
- To identify the most targeted sector and the topmost victim regions in India and the major radical groups behind those attacks in the last two decades.
- To perform several classification techniques for predicting the success or failure of the terrorist attack based on its type and the weapon used for attack.

From the above objectives we can identify a goal of recognizing the patterns and success rate of occurrence of terrorist incidents. These kinds of insight patterns are useful in providing the following benefits.

- To take proper preventive measures to eradicate the terrorism.
- To increase the foreign investments by making the country safe and secure.
- It promotes the growth of economy and develops the tourism.

2. Literature Review

M. Khalifa et. al, presents "Terrorist attacks got mainly significant imperative issue to all humankind in the world". It also employed statistical techniques and association mining algorithms applied to terrorism attacks to identify frequent invisible patterns in database [1].

J. V. Pagán described three classification algorithms (K-Nearest Neighbor, Discriminant Analysis, and Recursive Partitioning) of terrorist attacks database and reduces classification error rate [2].

S. Ray provides Artificial Intelligence has major attention in digital area and described machine learning algorithms merits, demerits of application perception to make decision in selecting suitable learning algorithm to meet explicit requirement of application [3].

K. P. Shroff and H. H. Maheta incorporated comparative study on machine learning classification algorithms performance based on feature selection in high dimensional data [4].

Tarik A. Rashid et. al, presented attack user behaviors of the wicked web mining systems and employed possible solutions for securing web users, organizations, society to prevent attacks. Naïve Bayes approach (NB) and K-Nearest Neighbor (K-NN) algorithms are incorporated to detect terrorist threats in web mining-based approaches [5].

Tarik A. Rashid and Salwa Mohamad described the detection of wicked internet user behaviors in internet forums, Wi-Fi, websites, email accounts, Facebook, etc, using machine learning techniques Random Forest (RF) and Support Vector Machines (SVM) methods [6].

Enrique Lee Huamaní and Alva Mantari, incorporated Terrorist attacks influence confidence, security of citizens, destruction of order and as increase of social networks, terrorist attacks in global are also ongoing. Author employed Artificial Intelligence techniques and classification models to visualize and predict possible terrorist attacks [7].

M. Irfan Uddin and Nazir Zada, address Terrorist Activities and their consequences suppressed physically, emotionally. Five different models on deep neural network (DNN) are incorporated to understand behavior of terrorist activities and presented analytical based results [8].

V. Kumar, M. Mazzara, A. Messina, focused on data mining classification methods and the part of United Nations counterterrorism. It analyzes performance of classifiers (Multilayer Perceptron, Lazy Tree, Naïve Bayes, Multiclass) for detection of trends in terrorist attacks world GTD database [9].

N. Ouassini and A. Verma illustrated association between social, economic, and demographic indices and left-wing intolerance in state of Jharkhand in India [10].

R. Alhamdani, M. Abdullah, exploits deep learning techniques to identify the terrorist attacks behavior and distribution online misinformation using different forms of social media by employing global terrorist database [11].

3. Dataset

The Global Terrorism Database (GTD) [12] is an open-source database from 1970 to 2017, with information on terrorist events around the world. This includes about 1,81,691 terrorist attacks for each scenario, with 135 categories, making it the most detailed unclassified data based on terrorist events in the world. For each terrorist event, the information about the date, location, attack, weapon, target/victim, and perpetrator (the group which carried out the attack) etc. are provided. In this study, we performed analysis on the GTD data during the time 2000 to 2017 [13],[14].

The GTD data was collected from different resources which would cause data inconsistency. Another problem is missing data values which if unhandled causes distortion in the analysis. Therefore, the pre-processing techniques like feature selection, treating the missing values and null values, normalizing the values are

performed to accomplish good analysis. Out of 135 attributes in the original dataset, by performing the Data reduction [15],[16], we selected the features relevant to our study which are described in the Table 1.

Table 1. Selected Attributes for Analysis

Attribute	Description
eventid	ID allotted to the incident occurred.
iyear, imonth, iday	The year, month, and day the incident occurred.
country_txt, provstate, city	Country name, state, city where the incident took place.
latitude, longitude	Location details of the place where the incident occurred.
attacktype1_txt	Type of the attack occurred.
success	Whether the incident was successful.
weaptype1_txt, weapsubtype1_txt	Type and subtype of the weapons used in the attack.
weapdetail	Information about the type of weapon used in the attack.
targtype1_txt, targsubtype1_txt	Type and subtype of the target/victim of the attack.
gname	Name of the group that carried out the attack.
nkill, nwound	Number of confirmed fatalities and non-fatal injuries of the incident.

The missing values and the instances with value as unknown of the attribute 'gname' have been removed as it is of no use to the analysis if the radical group responsible to the attack is not known. Similarly, the null values in the attributes 'nkill', 'nwound' have been replaced by the value 0 since we must perform aggregation function on the attribute values. The data is now preprocessed and ready to use for data mining [17],[18].

GTD data extracted for was checked and numerically coded using the Python programming language. The analysis was done purely python based in the Jupyter Notebook. Jupyter notebook is open-source software to perform programming and it is an interactive computational environment. The entire data is split into two sets to construct a classifier model: training data with 80% and testing data with 20% of the dataset [19],[20].

4. Experimental Methodology

In this paper, we conduct the experiments to test the performance of different machine learning models to identify the most suitable algorithm for the GTD database and to identify the interesting insights of the results. Classification models are built using such algorithms as Decision Tree, Naïve Bayes, Support Vector Machine. Naive Bayes belongs to the Bayes family as a probabilistic classifier. Decision Tree induction is an algorithm for the top-down recursive induction of tree [21],[22]. Support Vector Machine transforms the original training data into a higher dimension using a non-linear mapping to figure out the best separating hyper plane.

The classifier's accuracy can be further improved with the use of ensemble approaches such as the Random Forest Classifier. Ensemble Approach is a machine learning method to improve accuracy by studying a series of individual (base) classifier models and integrating them. The performance of the ensemble methods is compared with the base classifier models[23],[24]. To estimate the performance of the classifiers, evaluation metrics like accuracy score, f-measure, Jaccard similarity index, ROC curve are applied on the models developed to find the model with good accuracy [25].

In this section, first we discuss about the visualization of several statistical insights derived and then about the different classifier models developed and their accuracy results.

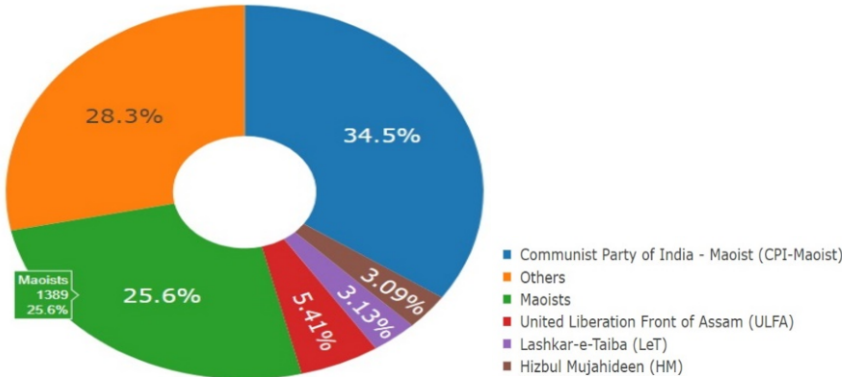

Figure 1. Donut chart showing major radical groups in India.

The donut chart in Figure 1 shows the major radical groups who contributed highest frequency attacks in India. It is evident that highest number of terrorist attacks are done by CPI-Maoist group. In India, most of the terrorist attacks are done by the domestic radical groups compared to the International Terrorist Organizations.

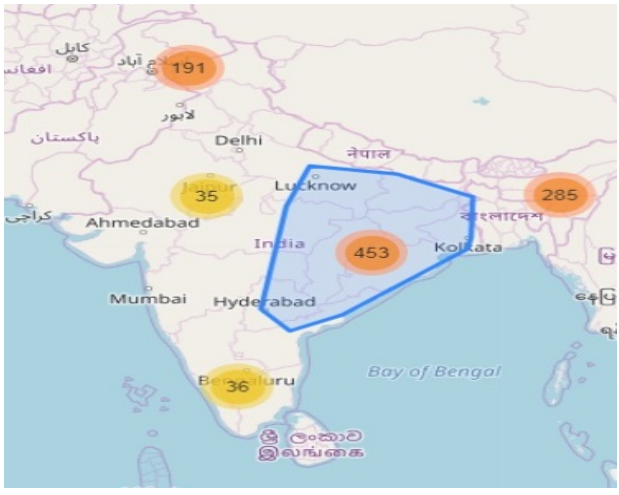

Figure 2. Map view of terrorist attacks in India.

The map in Figure 2 shows that based on several features like the type of attack and the number of fatalities, the data points can be partitioned into several clusters.

Each cluster can be further divided into several sub-clusters. The location details and the number of attacks on each city can be identified; thereby counter measures can be taken to concentrate on the most targeted areas. From the analysis, it is evident that the Middle East region of India is highly prone to terrorist attacks.

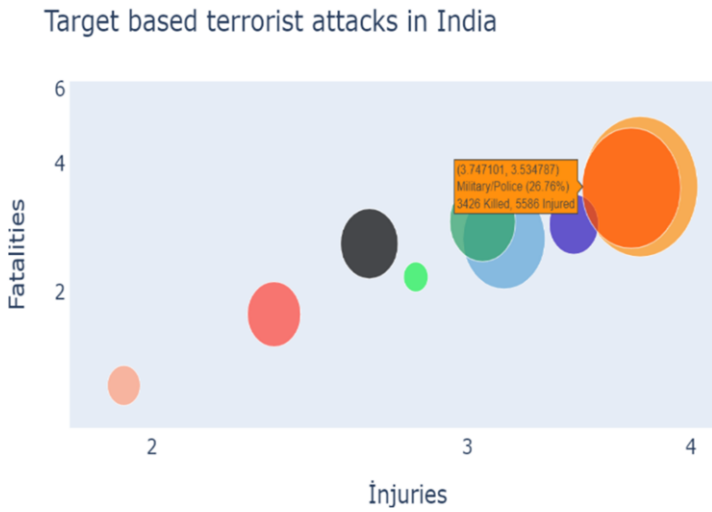

Figure 3. Scatter plot showing target-based attacks in India.

The scatter plot generated as in Figure 3 reveals the hidden information that Military is targeted the most in India with about 27% of all the attacks i.e., 3426 people were killed, and 5586 people were injured. Both the fatalities and injuries are comparatively higher in Military because of the terrorist attacks. Even the Education sector is relatively affected by the terrorism.

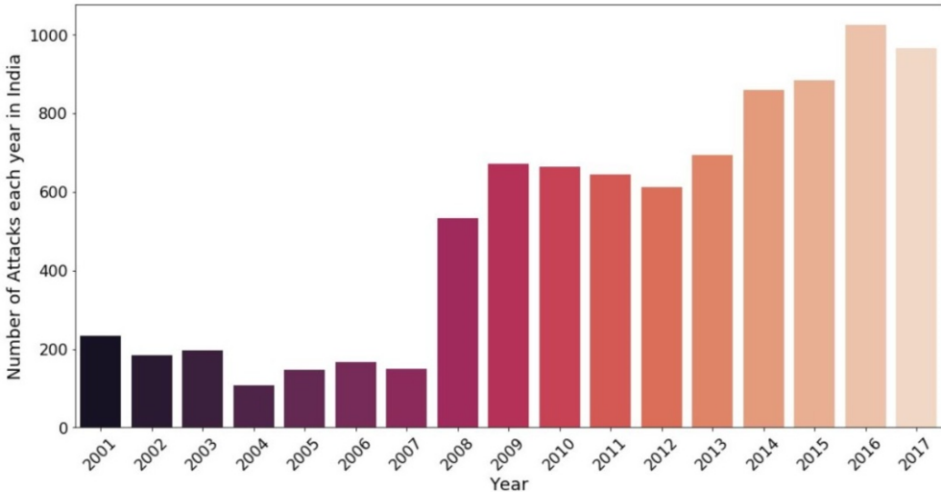

Figure 4. Histogram showing the number of attacks in India.

The histogram plotted in Figure 4 indicates how many terrorist attacks have been committed in India in the last two decades. The rate of attacks gradually increased from the year 2012. However, the attacks occurrence may either be successful or failure.

The Tree map in Figure 5 depicts the top 10 victim regions in India. Out of all the regions in India, terrorist attacks are more frequent in Imphal (36.59%) and Srinagar (24.54%) contributing a major share of greater than 50% out of all the attacks in India.

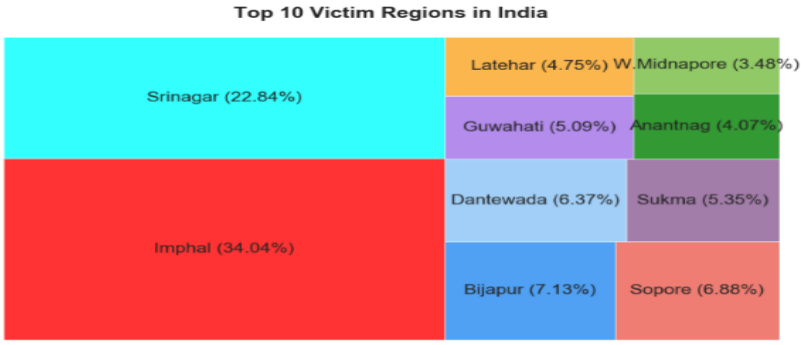

Figure 5. Tree Map showing the Top 10 victim regions in India.

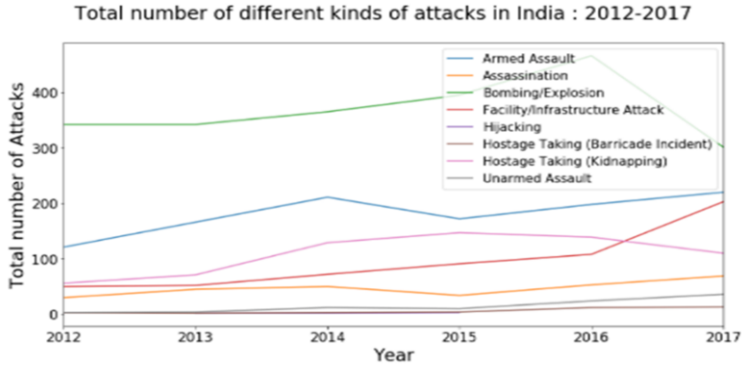

Figure 6. Time-Series graph showing different kinds of attacks in India.

The time-series graph in Figure 6 shows the total number of different types of attacks in India from 2012 to 2017. It shows that Bombing/Explosion is the most frequent method of attack implemented by terrorists in all the years.

The word cloud in Figure 7 shows the states which are highly prone to terrorist attacks. Further protection must be established for Jammu and Kashmir out of all states to prevent terrorist attacks.

Figure 7. Word Cloud showing the states which are mostly attacked.

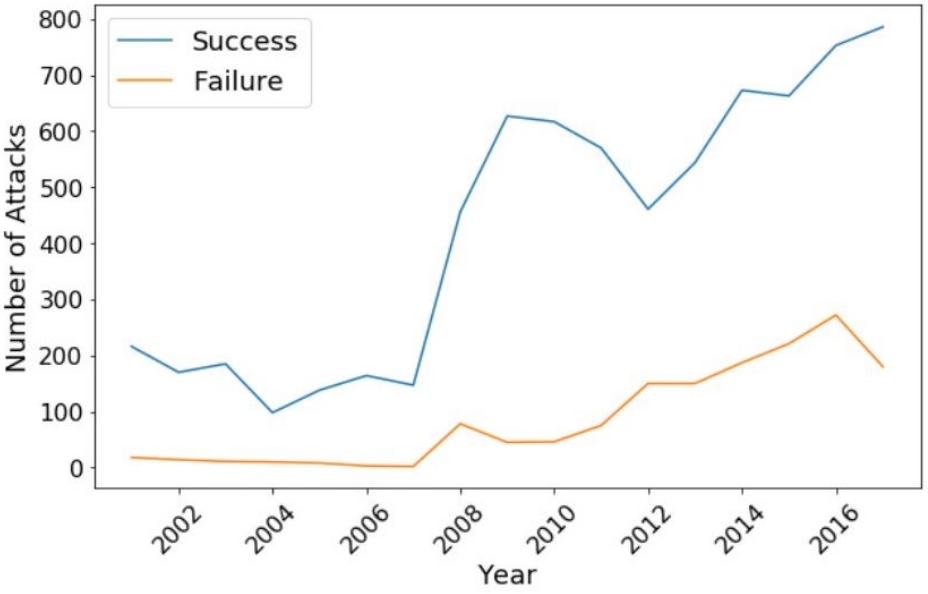

Figure 8. Time-Series graph showing the number of successful and failure events.

Attacks planned by terrorists may either become successful or failure. The time-series graph in Figure 8 describes the summary of the attack's success vs failure during the period 2000 to 2017. This graph depicts that the number of successful attacks increased gradually from the year 2012. Terrorism has seen a drastic increase in the year 2009 when compared to the number of attacks in the year 2007.

The word cloud generated in Figure 9 highlights the most frequent word vocabulary of weapons used for attack in India. Out of all the weapons, explosives, and firearms (portable guns) are used the most in India.

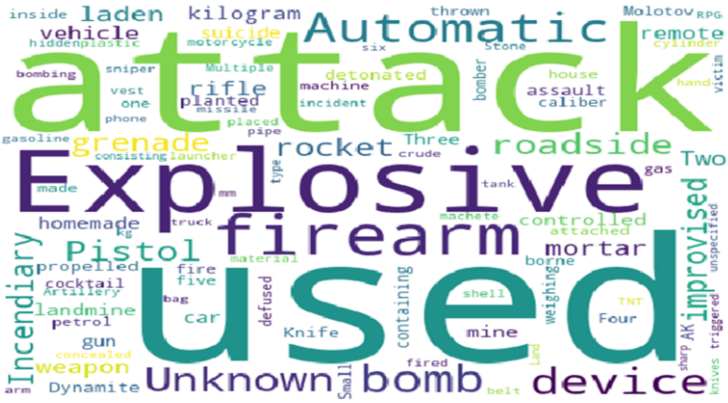

Figure 9. Word Cloud showing the Weapons used for attack.

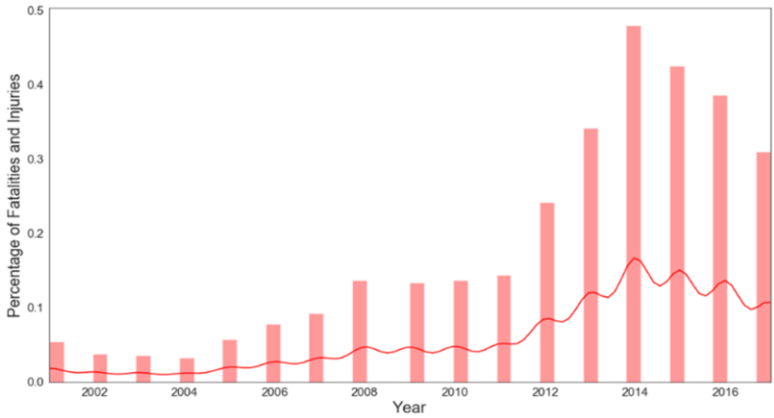

Figure 10. Dist. Plot showing the percentage of fatalities and injuries.

The seaborn dist. plot in Figure 10 shows the density distribution of the total percentage of fatalities and injuries during the time period 2000-2017.It is evident that in the year 2014, the fatalities and injuries are high compared to other years.

Table 2. Classifier Models and Evaluation results

Algorithm	Accuracy Score	F1 Score	Jaccard Similarity
Naive Bayes	0.410755	0.45617	0.410755
Random Forest	0.837529	0.91080	0.837529
Decision tree	0.836957	0.91085	0.836957
SVC	0.834668	0.90988	0.834668

The dataset is divided into training set and test set in the 80:20 split ratio to implement the classification algorithm. The GTD incident is graded into either success or failure based on the type of attack that occurred, and the type of weapon used for the attack. The Table 2 shows the detailed results of the classification algorithms applied. Out of all the individual classifier models developed, Decision Tree Classifier gives the good accuracy. To improve the classification accuracy, ensemble methods like Random Forests classifier is build.

In addition to the standard evaluation metrics applied, ROC curve serves as the metric which can determine the good classifier. Of all the classifiers applied, Random Forest Classifier is the better one as shown in the Figure 11.

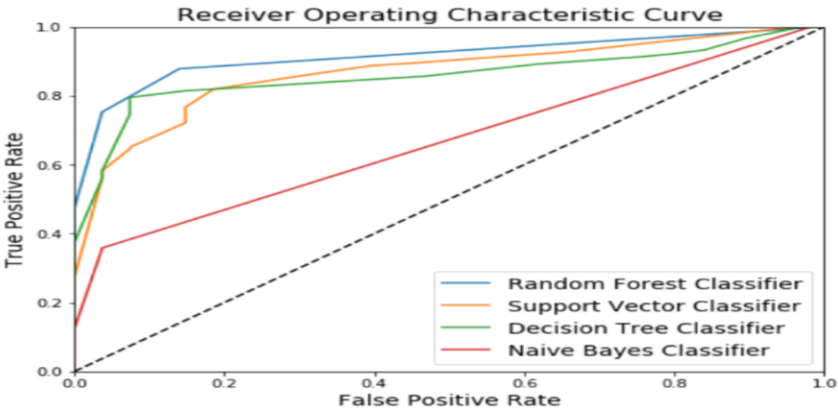

Figure 11. ROC curve showing the accuracy of Classification models.

Several metrics can be calculated from the confusion matrix like the accuracy, sensitivity, and precision etc. The Figure 12 shows the confusion matrix plotted for the Random Forest Classifier.

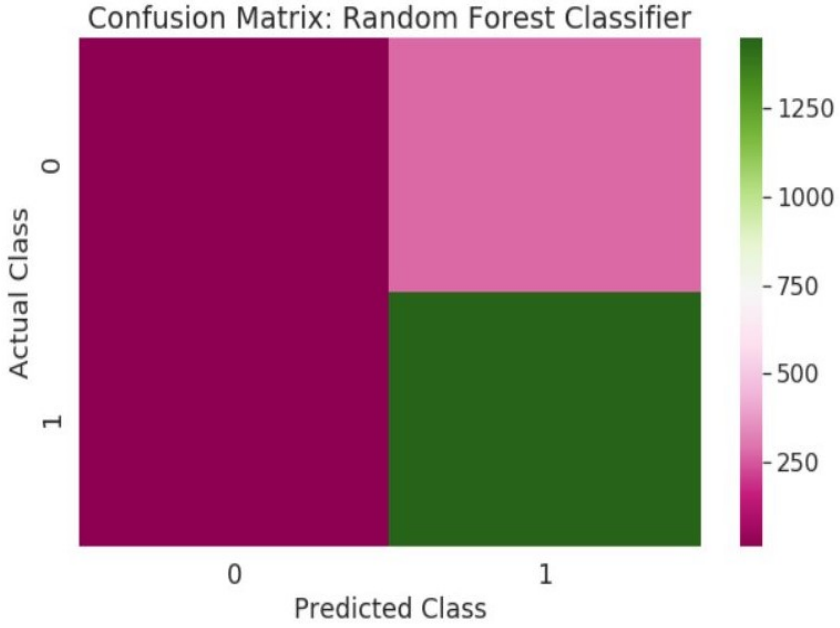

Figure 12. Confusion Matrix of Random Forest Classifier.

The time-series graph in Figure 13 plots the total number of different types of attacks in the world from 2011 to 2017. It shows that Bombing/Explosion is the most frequent method of attack implemented by terrorists in all the years.

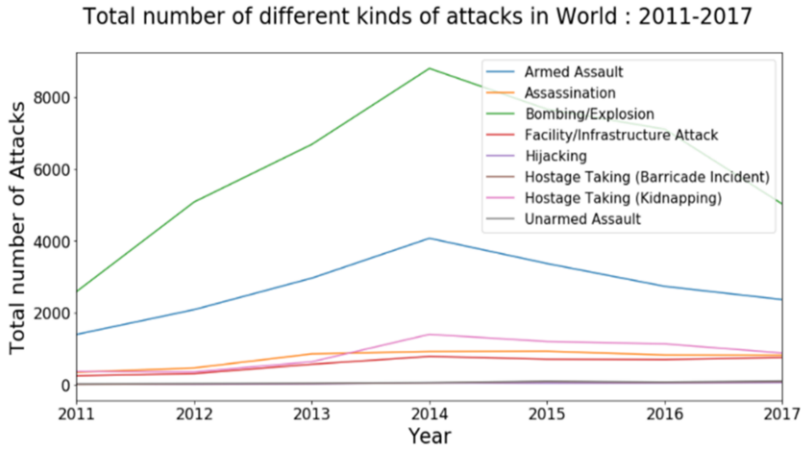

Figure 13. Time-Series graph showing different kinds of attacks in the world.

The map in Figure 14 shows that based on several features like the type of attack and the number of fatalities, the data points can be partitioned into several clusters. Each cluster can be further divided into several sub-clusters. The location details and the number of attacks on each city can be identified; thereby counter measures can be taken to concentrate on the most targeted areas. From the analysis, it is evident that the South-west-Asian region is highly prone to terrorist attacks.

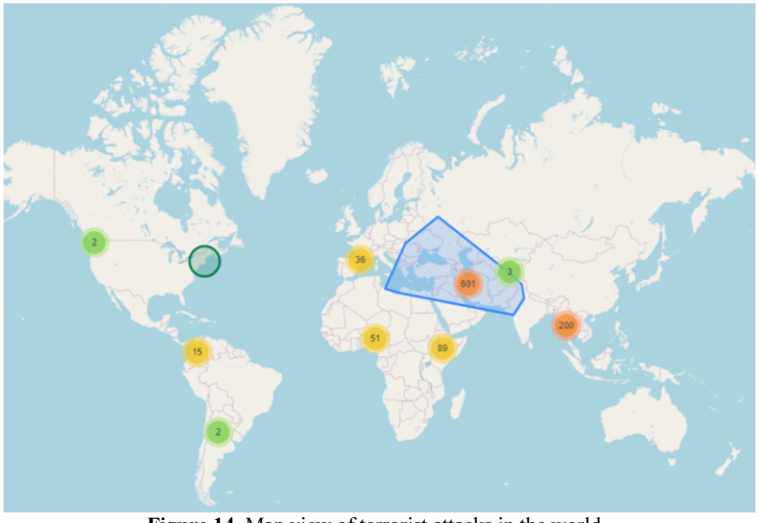

Figure 14. Map view of terrorist attacks in the world.

The scatter plot in Figure 15 reveals the hidden information that Military is targeted the most in the world. Both the fatalities and injuries are comparatively higher in Military because of the terrorist attacks. Even the Education sector is relatively affected by the terrorism.

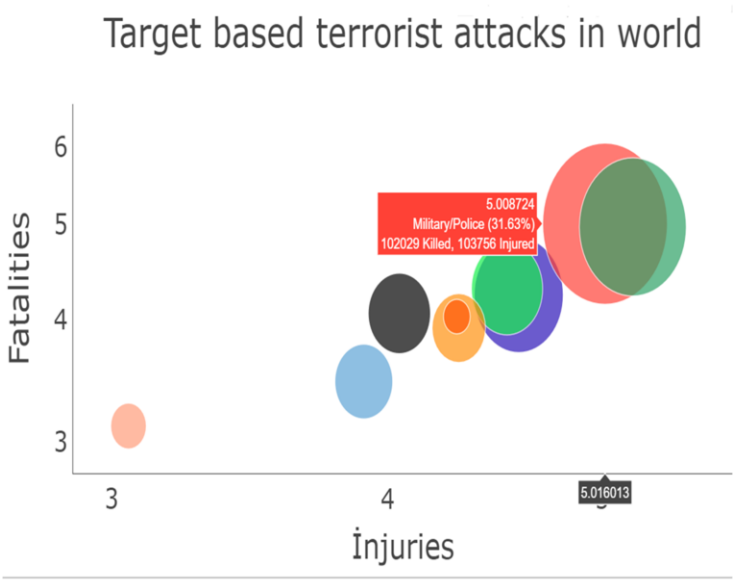

Figure 15. Scatter plot showing target-based attacks in the World.

5. Conclusion and Future Scope

This article has addressed both the worldwide and India's statistical perspectives on terrorism. To develop a model best suited for data analysis, many machine learning algorithms were applied to the GTD data. Classification models such as Decision Tree, Naive Bayes, Random Forest, Support Vector Machine are created. Model evaluation metrics like accuracy score, f-measure, Jaccard similarity index, ROC curve was applied to find the model with good accuracy. The results slow that, among all the models, Random Forest Classification algorithm had the highest possible accuracy.

The study disclosed the top 10 victim regions of terrorism in India and predicted the count of fatalities and injuries based on target. The study also revealed the major radical groups responsible for the most frequent attacks in India and identified which states are mostly likely to be attacked. The experimental results suggest that the spread of terrorism in India is mostly due to extreme domestic groups compared with external threats.

An additional direction, innovative based AI solutions can be incorporated by utilizing the comparative based analytical results for progressing future enhancements for decision making on different data sets by considering other aspects like injuries, GPS data, spatial-temporal , video surveillance performance extraction and trajectories data.

References

[1] Khalifa NEM, Taha MHN, Taha SHN, Hassanien AE. Statistical Insights and Association Mining for Terrorist Attacks in Egypt. The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) Advances in Intelligent Systems and Computing. 2020, 921: 291-300.

[2] Pagán JV. Improving the classification of terrorist attacks a study on data pre-processing for mining the Global Terrorism Database. In 2nd International Conference on Software Technology and Engineering, 2010 San Juan,110:104.

[3] Ray S. A Quick Review of Machine Learning Algorithms. In International Conference on Machine Learning, Big Data Cloud and Parallel Computing, 2019 Faridabad, p. 35- 39.

[4] Shroff KP, Maheta HH. A comparative study of various feature selection techniques in highdimensional data set to improve classification accuracy. In International Conference on Computer Communication and Informatics (ICCCI), 2015 Coimbatore, p. 1-6.

[5] Rashid TA, Rashad DD, Gaznai HM, Shamsaldin AS. A Systematic Web Mining Based Approach for Forecasting Terrorism. In Communications in Computer and Information Science. 2017, Singapore, 752 .

[6] Rashid TA., Mohamad SO. Enhancement of Detecting Wicked Website Through Intelligent Methods. In Security in Computing and Communications. SSCC 2016 Singapore, 62.

[7] Enrique Lee H, Alva Mantari A. Machine Learning Techniques to Visualize and Predict Terrorist Attacks Worldwide using the Global Terrorism Database. International Journal of Advanced Computer Science and Applications , 2020 11(4) .

[8] Irfan Uddin M, Nazir Zada. Prediction of Future Terrorist Activities Using Deep Neural Networks. Hindawi Complexity 2020, 20.

[9] Kumar V, Mazzara M, Messina A, Lee J. A conjoint application of data mining techniques for analysis of global terrorist attacks prevention and prediction for combating terrorism. Advances in Intelligent Systems and Computing, 2019, Berlin, Germany. p.146-158.

[10] Ouassini N, Verma A. Socio economic inequality or demographic conditions a micro-level analysis of terrorism in Jharkhand.In Journal of Victimology and Victim Justice, 2018, 1(1), p. 63–84.

[11] Alhamdani R, Abdullah M, Sattar I. Recommender system for global terrorist database based on deep learning. In International Journal of Machine Learning and Computing, 2018, 8, p. 571–576.

[12] Global Terrorism Database (GTD) https://www.kaggle.com/START-UMD/gtd, https://ourworldindata.org/terrorism.

[13] Ben Meskina S. On the effect of data reduction on classification accuracy. In 3rd International Conference on Information Technology and e-Services (ICITeS), 2013 p.1-7.

[14] Tripathi A, Yadav S, Rajan R. Naive Bayes Classification Model for the Student Performance Prediction. In 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) 2019 Kerala, India,p. 1548-1553.

[15] Rajesh P, Sai Prasanna. A Forensic Approach to Perform Android Device Analysis. In International Women Conference on Technological Innovations. 2019 (7).

[16] Diao R. Decision Tree-Based Online Voltage Security Assessment Using PMU Measurements. In IEEE Transactions on Power Systems 2009 24(2), p.832-839.

[17] Suriya Prakash J, Annamalai Vignesh K, Ashok C, Adithyan R. Multi class Support Vector Machines classifier for machine vision application. In International Conference on Machine Vision and Image Processing (MVIP) 2012 Taipei,p. 197-199.

[18] Rajesh P, Narsimha G. Cerebration Of Privacy Preserving Data Mining Algorithms. International conference on machine learning and data analysis ICMLDA 2014 (2) USA,P:813-817.

[19] Hasan MK, Alam MA, Das D, Hossain E, Hasan M.Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers. In IEEE Access 2020 (8), p. 76516-7653.

[20] Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. In IEEE Transactions on Knowledge and Data Engineering 2005 17(3), p.299-310.

[21] Mrunal k, srinivasu N. support value based fusion matching using iris and sclera features for person authentication in unconstrained environment. In Journal of Engineering Science and Technology 2020 15(4):p. 2595 - 2609.

[22] Ismail B, Rajesh P. A Machine Learning Classification Technique for Predicting Prostate Cancer.In 20th Annual IEEE International Conference on Electro Information Technology Eit2020 USA.

[23] Rajesh P, Alam M. A Data Science Approach to Football Team Player Selection. In 20th Annual IEEE International Conference on Electro Information Technology EIT 2020 USA.

[24] Rajesh P, Narsimha G. Cerebration Of Privacy Preserving Data Mining Algorithms. In International conference on machine learning and data analysis ICMLDA_ 2014 USA.

[25] Ranjeeth M, Srinivasu N. DES secured k-NN query over secure data in clouds. In Journal of Theoretical and Applied Information Technology 2016. 91(2), p.384-389.