Analysis and Prediction of Terrorist Attacks using Supervised Machine Learning and Deep Learning Techniques

MSc Research Project
MSc in Data Analytics

Sinchana Jyothilinga
Student ID: x21128952

School of Computing
National College of Ireland

Supervisor: Dr. Cristina Hava Muntean

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Sinchana Jyothilinga
Student ID:	x21128952
Programme:	MSc in Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Cristina Hava Muntean
Submission Due Date:	01/02/2023
Project Title:	Analysis and Prediction of Terrorist Attacks using Supervised Machine Learning and Deep Learning Techniques
Word Count:	5752
Page Count:	26

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Sinchana Jyothilinga
Date:	31st January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Analysis and Prediction of Terrorist Attacks using Supervised Machine Learning and Deep Learning Techniques

Sinchana Jyothilinga
x21128952

Abstract

Terrorism has been a major concern worldwide for a very long time. There have been thousands of civilian casualties as a result of these assaults. If these assaults could be foreseen, then governments and their defense establishments can work together to devise a plan of action to prevent them. This is achievable with the use of machine learning. This study proposes to analyze the historical terrorist attack data available in the Global Terrorism Database (GTD) to forecast the future terrorist attacks and fatalities of the top three most attacked nations and also classify the attack weapon type and the terrorist group responsible for the attack. The dataset was preprocessed by converting relevant variables to categorical form, filling the missing values with Mean/Median imputation, and shortening lengthier categorical names. Weapons were classified using a kNN classifier, perpetrators were categorized using Decision Tree and MLPClassifiers, and fatality predictions were made using Random Forest and MLPRegressors. Time Series Analysis using FbProphet model was used to forecast terrorist strikes in the top three most attacked countries. The kNN classifier yielded an accuracy of 92.25%. For perpetrator classification, both the Decision Tree classifier and MLPClassifier performed equally well, providing an accuracy of 90%. The Random Forest regressor outperformed the MLPRegressor obtaining a lower MSE value. The forecasts of the terrorist attacks in 2021 seen in the plots obtained by time series analysis match with the actual attacks that happened during that year. ML offers a large potential for an investigation into terrorism to precisely anticipate future attacks.

1 Introduction

1.1 Background and Motivation

Terrorism is the employment of unlawful means to attain a political, religious, or ideological objective. There is no commonly acknowledged definition of terrorism; rather, the term may be interpreted in several ways. Assaults, hijackings, bombings, hostage-taking, and other violent crimes undertaken with the intent to cause fear or advance a political, social, economic, or religious agenda may all be categorized as terrorist activities. Terrorist assaults have resulted in the deaths of thousands of innocent individuals. The availability of more powerful weapons over the last several years has seemingly resulted in an increase in the severity of assaults. Only if the next attack and its target location can be predicted can a preventative countermeasure be devised. Consequently, terrorism research is an ongoing undertaking with room for improvement.

To make progress in the study of terrorism and to forecast future acts of terrorism, it is essential to have access to historical data for the purposes of analysis. In order to accomplish this goal, a group of investigators contacted the Pinkerton Global Intelligence Service (PGIS) and requested information on acts of terrorism. This data was then included into the Global Terrorism Database (GTD).[1] Data collection efforts, the pros and cons of open-source data in general and the GTD in particular, and descriptive statistics on the GTD's contents and how, despite well-known constraints, the data offers a broad range of analysis are all covered in this discussion (LaFree and Dugan; 2007) by the director of the National Center for the Study of Terrorism and Responses to Terrorism (START). The author (Lafree; 2010) also offers a full discussion of the GTD's accomplishments as well as its challenges.

Using a broad variety of machine learning and data mining techniques, a large number of scholars have investigated the GTD in order to forecast terrorist activities. Nevertheless, there is always room for improvement. Because of developments in machine learning and the increased computational capacity of computers, it is now possible to make more accurate predictions of future terrorist acts. Such predictions could be helpful in the development of a countermeasure that would help prevent such attacks from occurring.

1.2 Research Question

How accurately can future terrorist attacks be forecasted using time series analysis and how well can the weapon and the terrorist group responsible for the attack be classified? Also, which learning technique is better at predicting the fatalities of an attack, machine learning or deep learning?

1.3 Objectives

The purpose of this study is to analyze the past terrorist attacks data available in GTD in order to determine the patterns of these attacks, the top three terrorist-prone countries, the weapons used for the attack, and the perpetrator group responsible for the attack. GTD contains information from 1970 to 2020. There are no accessible statistics post-pandemic. The major contribution of this research is forecasting future attacks using time series analysis and examining the seasonality and trend components to draw insights from it. In tandem with this, the classification of the weapon and perpetrator will be conducted to offer a clearer picture of the attack type. Regression analysis is used to predict the number of casualties in an attack. This may aid the governments of these nations in formulating a defensive plan to prevent these assaults from occurring and to safeguard the lives of their citizens.

1.4 Structure of the Paper

This segment will offer the general layout of the research paper, which will be discussed in the following sections.

[1] https://www.start.umd.edu/gtd/

- Section 2: This section provides a detailed review of the work related to employing machine learning to forecast terrorist acts. According to the mentioned criteria, the section has been divided into numerous pertinent subsections.
- Section 3: This section elaborates on the methodology used to acquire the findings. It has a number of subsections matching each stage of the overall approach, each of which describes and defends the method followed during that step.
- Section 4: This section explains the methods that would serve as the foundation of the proposed implementation and the related requirement descriptions.
- Section 5: This section concerns methodology implementation. It includes the tools and software used in the implementation and the final phases of model building, including outputs, transformed data, and questionnaires administered.
- Section 6: This section outlines the specifics of several experiments conducted and their outcomes. Additionally, it is separated into subsections for each experiment. It also includes a discussion section that describes each experiment's findings.
- Section 7: This section addresses the conclusion and the potential for further research on the topic.

2 Related Work

To this day, several studies and other forms of research work have been conducted on the subject of terrorism. Researchers have applied a number of machine learning techniques, ensemble approaches, and deep learning methods on the GTD as well as other open-source datasets dedicated to terrorism. A selection of these may be found in the works that are discussed below.

2.1 Review of research work done for the prediction of the perpetrator group

The authors (Tolan and Soliman; 2015) used data mining classification technique to compare five base classifiers: K-Nearest Neighbour (KNN), Nave Bayes (NB), Support Vector Machine (SVM), Iterative Dichotomiser (ID3), and Tree Induction (C4.5) in an attempt to predict the terrorist group responsible for a terrorist attack. The authors dealt with the missing data values using two distinct methods: Mode-Imputation and Litwise-Deletion. In the Mode-Imputation approach, SVM outperformed other classifiers, but KNN outperformed other classifiers in the Litwise-Deletion method. The authors have solely analyzed data relevant to Egypt, which could have been broadened to include the top five nations or regions.

In order to forecast the terrorist organization responsible for assaults in the Middle East and North Africa from 2009 to 2013, researchers (Khorshid et al.; 2015) evaluated standard, ensemble, hybrid, and hybrid ensemble machine learning algorithms such as K-nearest neighbors, Nave Bayes, Support Vector Machine, Decision Tree; Functional Tree, Hybrid Hoeffding Tree, Hybrid Nave Bayes with Decision Table, Classification through Clustering, Random Forests, and Stacking. Ensemble approaches outperformed hybrid ensemble methods. In the future work section, they advise using Neural Networks to predict a terrorist group.

The GTD has also been used by another research team (Talreja and Mahesh; 2017) to forecast the group of perpetrators using Support Vector Machine (SVM), and they achieved 75% accuracy. The writers zeroed in on the terrorist organizations responsible for the strikes in India. Support Vector Machines excel in solving classification issues, particularly when the dataset is clean and well-balanced. Moreover, SVM prevents the overfitting of data. Overfitting of data in such datasets would provide misleading conclusions, making this especially crucial for a dataset like the GTD.

Many criminal organizations in the GTD remain nameless. Text-based feature creation and logistic regression on the GTD were used by a group of researchers (Laite and Sankaranarayanan; 2019) in order to undertake terrorist group categorization of the terrorist acts that were carried out by these unidentified terrorist organizations. Using a train-test split of 75-25 and 10 folds of stratified cross-validation, the model was able to achieve a median accuracy of 88% in testing. The scope of this study may be broadened to determine, from the attack summary, which terrorist organization was responsible for the assault. Accuracy may also be enhanced by using more robust ensemble approaches.

2.2 Review of research work done for the prediction of the success of a terrorist attack

Researchers (Agarwal and Chandra; 2019) utilized the GTD to forecast the success of an assault, identify the group responsible for an attack, and identify the impact of external variables. To determine whether or not an assault was effective, the authors used Decision Tree and Random Forest classifiers. To choose the characteristics to input into the model, the authors relied on python's standard library. With an accuracy of 89%, the model performed well.

Three different machine learning methods were used by a group of researchers (Alsaedi and Alharbi; 2019): RF, KNN, and NB. Each algorithm was used to build two models, one for each job, with the data collected using the holdout method for the first model and cross-validation for the second. In comparison to other models, the authors found that the RF model was the most accurate (91.86%). Only three different algorithms have been investigated by the researchers, which restricts the scope of the study.

2.3 Review of research work done for the prediction of the region of attack

In order for a government body or intelligence agency to devise a countermeasure against the next terrorist assault, it is necessary to precisely foresee the attack. Researchers (Huamaní et al.; 2020) have made an effort in this direction by using two machine learning algorithms, namely random forest and decision trees. Both models produced the same range of probabilistic findings, which ranged from 75.49% to 90.414% of assertiveness. Although this study attempted to forecast assaults over broad geographical areas, it would have been more fruitful to narrow down on a specific nation or location.

Linear Discriminant Analysis (LDA), Support Vector Machines (SVM), Kernel Naive Bayes (GNB), Decision Trees, and Logistic Regression were the six different machine-learning techniques that were utilized by another group of researchers (Singh et al.; 2019) to forecast the assault region and country. The authors made the observation that LDA, NB, LR, and SVM generated a better accuracy of 82% in both scenarios when they were attempting to forecast the attack region as well as the country. The research may be advanced by making use of ensemble models and deep learning strategies, both of which are known to increase accuracy.

2.4 Review of research work done for the prediction of the attack weapon

In order for military agencies to be ready for the next assault, more information regarding the anticipated strike must be made accessible. A more effective defensive plan that can prevent the next assault from occurring and save the lives of innocent victims may be created if the weapon type to be utilized in the next attack is known in advance. Based on this, the research team (Verma et al.; 2018) used Random Forest to categorize the weapon and obtained an accuracy of 91%. The authors have confined the scope of their investigation to a single algorithm.

To further classify the lethal instrument, another team of researchers has used K-means cluster analysis (Li et al.; 2021). The authors have also used this clustering technique to categorize terrorists based on geographic location, method of attack, and victim demographics. This allowed the authors to identify the top five suspects in the 2015 terrorist strikes, for which the perpetrators had not yet been identified.

Reviewing and analyzing previous research on the prediction of the success of a terrorist attack, classification of weapon and perpetrator, and prediction of the next region or country of attack reveals that the researchers employed a variety of classical machine learning approaches, as well as ensemble learning approaches, in an effort to achieve high accuracy. In addition, the researchers attempted to predict which region or country would be the target of the next attack. In spite of the employment of a number of techniques for data preparation, feature selection, and dealing with missing values in the data, there is still room for development.

The originality of this study lies in the fact that it makes an attempt to forecast terrorist attacks in the top three countries with the highest risk of such attacks. It does so by collecting relevant data, cleaning that data, accounting for missing values, conducting exploratory data analysis, analyzing and selecting the most important features, employing time series analysis, and also attempting to put supervised machine learning models into practice for the classification of perpetrators and weapons.

3 Research Methodology

After doing an in-depth analysis of the research issue, the Knowledge Discovery in Databases (KDD) approach was selected as the methodology to be used throughout the research's execution. Figure 1 displays the research technique architectural diagram that is being used for the implementation. The KDD methodology, which is an iterative multi-stage technique, is being utilized in this study. The KDD methodology consists of the following six steps: data extraction, data pre-processing, feature selection, predictive analysis, interpretation and evaluation, and knowledge.

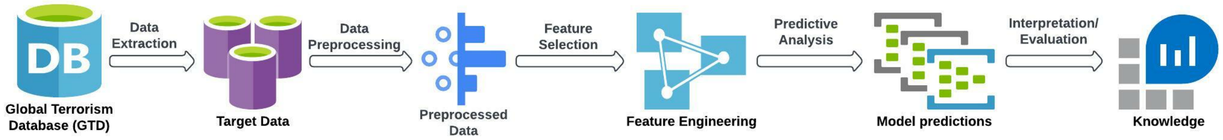
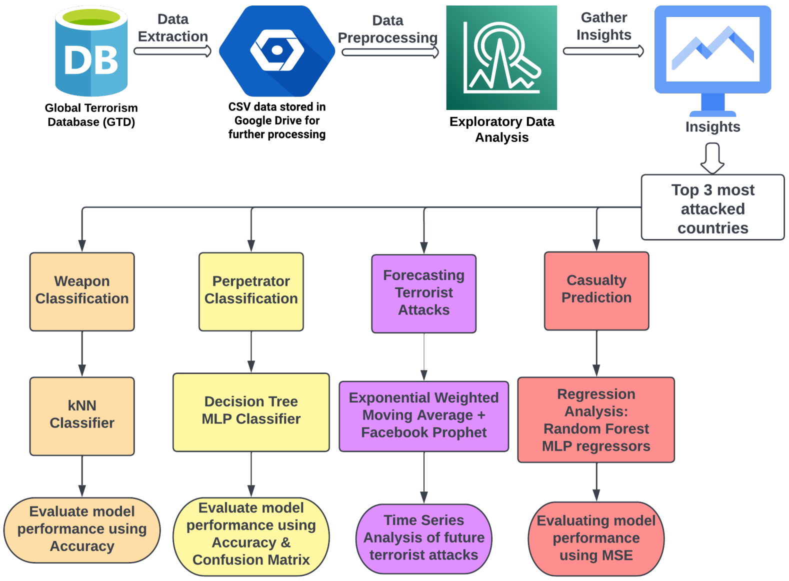


Figure 1: Knowledge Discovery in Databases

The system architecture diagram is shown in Figure 2. The GTD, which is an open-source database, will be queried for information pertinent to the proposed line of inquiry. The data that is needed for the study will be selected from this large database, and it will apply to terrorist acts that have occurred from 1995 to 2020.



Global Terrorism Database (GTD)
Data Extraction
CSV data stored in Google Drive for further processing
Data Preprocessing
Exploratory Data Analysis
Gather Insights
Insights
Top 3 most attacked countries
Weapon Classification
Perpetrator Classification
Forecasting Terrorist Attacks
Casualty Prediction
kNN Classifier
Decision Tree MLP Classifier
Exponential Weighted Moving Average + Facebook Prophet
Regression Analysis: Random Forest MLP regressors
Evaluate model performance using Accuracy
Evaluate model performance using Accuracy & Confusion Matrix
Time Series Analysis of future terrorist attacks
Evaluating model performance using MSE

Figure 2: Architecture Diagram

3.1 Data Acquisition

The information was gathered from the publicly accessible website known as the Global Terrorism Database (GTD)[2], which is administered by the University of Maryland. GTD is the most comprehensive unclassified database on terrorism ever compiled. The CSV file that was retrieved from GTD which includes information on terrorist acts from 1970 through 2020 has been uploaded to the cloud (Google Drive) for it to be processed further. Figure 3 shows the data points on the map where terrorist attacks have occurred. A glance of the attacks of past 9 years is shown using the wordclouds in Figure 4.

[2] https://www.start.umd.edu/gtd/

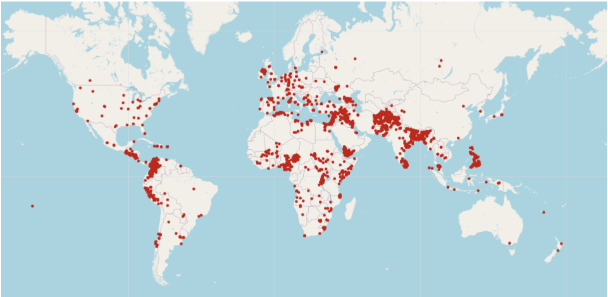
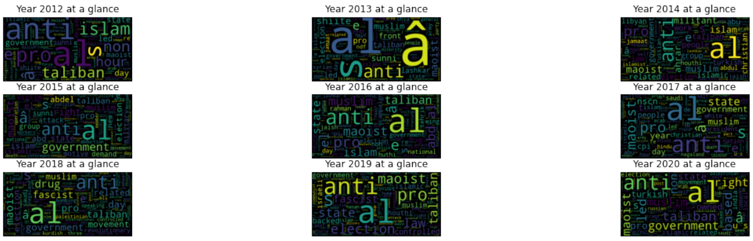


Figure 3: Geographical Attack Locations



Year 2012 at a glance
Year 2013 at a glance
Year 2014 at a glance
Year 2015 at a glance
Year 2016 at a glance
Year 2017 at a glance
Year 2018 at a glance
Year 2019 at a glance
Year 2020 at a glance

Figure 4: Past 9 years of attacks at a glance

3.2 Data Pre-processing

There are 135 columns and 2,09,706 rows in the data set. This data requires preprocessing and cleaning before it can be used for analysis. Numerous data points are missing from the collection. The following procedure is used to deal with these missing values: Classifying missing or unknown values in the code book is inconsistent. There were blank values, -9 and -99, in the original data. Numeric categorical characteristics are represented by -1, and text categorical attributes are represented by UNKNOWN. In the case of numeric characteristics, NAN is substituted for any coded missing values. 1, 0, and -1 are used as Yes, No, and Unknown values in many characteristics. As a means of facilitating exploratory data analysis, labels are substituted for these codes.

The Exploratory Data Analysis (EDA) is performed on the dataset after pre-processing the data. Figure 5 shows the graph of attacks by year from 1970 to 2020. It is evident from the graph that these attacks have been increasing. The highest number of attacks happened in the year 2014.

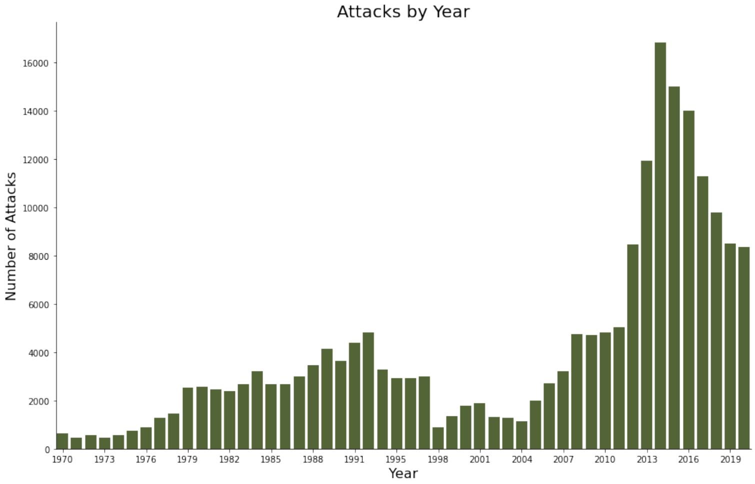


Figure 5: Number of attacks between 1970 - 2020

The EDA also revealed that the top 3 most attacked countries from the year 1995 to 2020 are Iraq, Afghanistan, and Pakistan as shown in Figure 6.

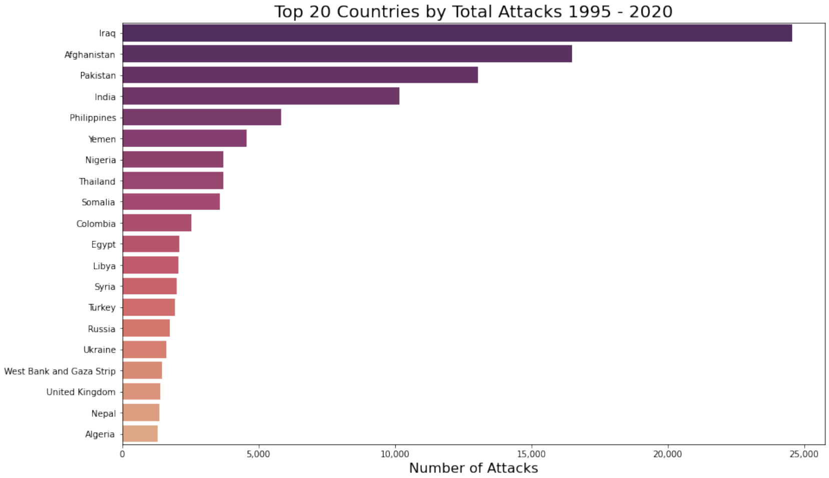


Figure 6: Top 20 countries by most attacks from 1995 - 2020

Figure 7 shows the favorite targets of terrorists. It is evident from the figure that Citizens and private properties are the most targetted for attacks. Hence, it is important to be able to be prepared for the attacks and prevent the fatalities.

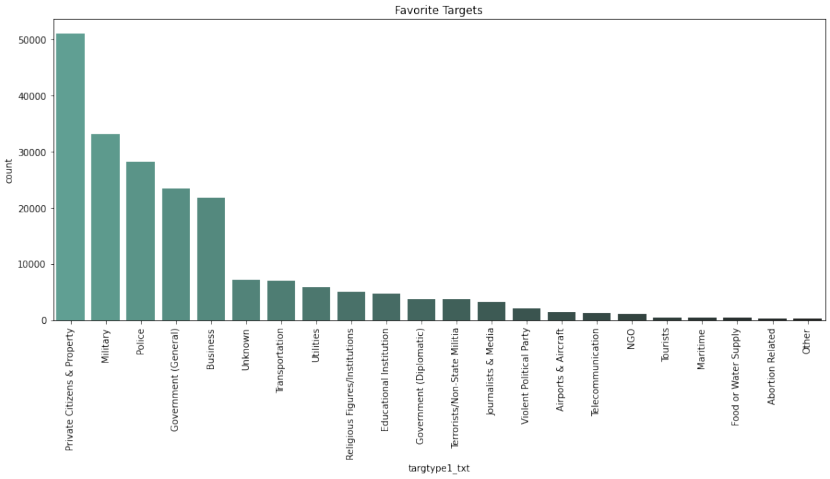


Figure 7: Favorite Targets of Terrorists

3.3 Feature Selection

This study endeavor is broken down into four tasks: classifying weapons, classifying perpetrators, projecting time series, and predicting casualties. A unique group of characteristics is chosen to represent each of these tasks depending on the kind of output or forecast that is anticipated. For the purpose of developing models and making predictions, the characteristics are pre-processed and cleaned.

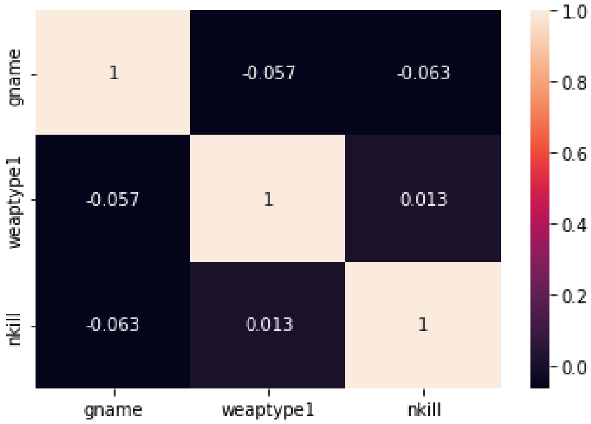


Figure 8: Correlation Matrix of Weapon type, Perpetrator group, and Casualty variables

Figure 8 shows the correlation matrix of the weapon type, perpetrator group, and casualty variables. As we can see, the variables that are positively correlated are weapontype1 and nkill. From this, we can say that as the type of weapon changes, the number of fatalities also changes. The type of weapon can significantly influence the number of casualties in an attack. For example, a biological or chemical weapon can result in a large number of casualties over a wide area, while a small knife may result in fewer casualties.

The perpetrator variable is negatively correlated with weapon type and casualty variables. However, the perpetrator's level of planning and target selection can also be important in determining the number of casualties. A perpetrator who has planned and prepared for an attack, and has selected a particularly vulnerable target, is likely to cause more casualties than a perpetrator who has acted impulsively and without planning.

Also, it is important to note how these factors interact and influence each other will also vary depending on cultural, social, political, and economic contexts.

3.4 Predictive Analysis

Once the features are selected, the predictive models are applied to the dataset. In this study, we have chosen the following models for the various prediction tasks:

- **Weapon Classification:** For predicting the type of weapon used in an attack such as Firearms, Explosives, chemicals, Biological, Nuclear, etc., the k-Nearest Neighbor (kNN) is being implemented. The kNN algorithm is used for the classification of weapon type since kNN performs well with a low number of features (curse of dimensionality). Only four predictor variables are being used for building the model (country_txt, region_txt, attacktype1_txt, nkill). If there were more features, it would have led to an overfitting problem. Figure 9 shows the boxplot of all the weapons used in attacks from 1995-2020.

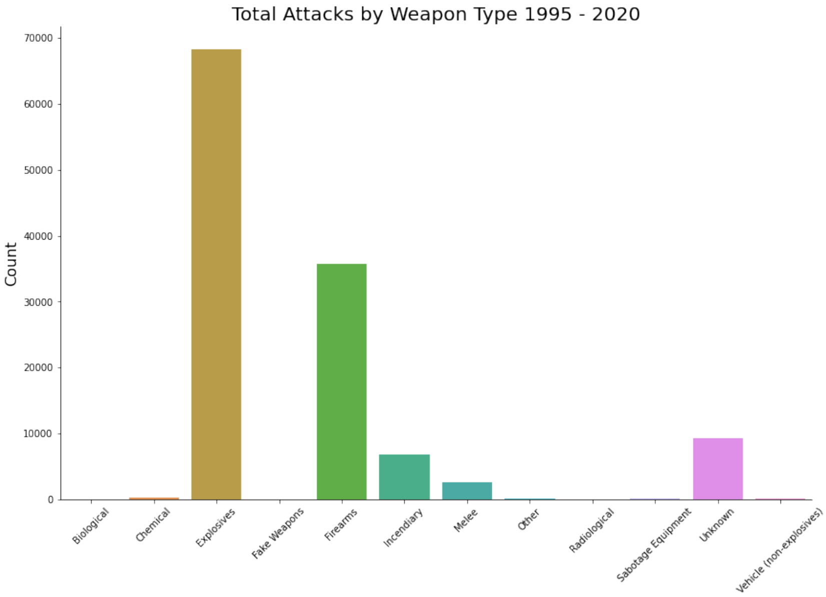


Figure 9: Attack Weapon Type

- **Perpetrator Classification:** For predicting the terrorist group responsible for an attack, two different algorithms are employed namely, Decision Trees, and Neural Networks. The accuracy of these two models is compared to choose the better-performing model. According to the literature survey done for the classification of perpetrator group, various researchers have already implemented machine learning and ensemble models and found that the accuracies of these models were not outstanding. One author suggested using Neural Networks in the future work section. Neural Networks can learn complex relationships between inputs and outputs. Hence, Neural Network was implemented for perpetrator classification Figure 10 shows the boxplots with perpetrator groups with highest terrorist attacks.

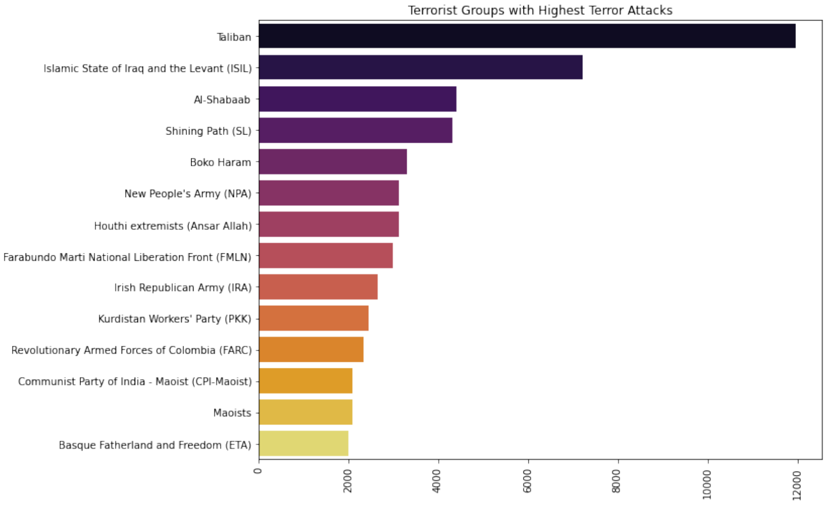


Figure 10: Perpetrator groups with highest terrorist attacks

- **Time Series Prediction:** The time series analysis is performed for the top 3 countries by considering the public holidays of these countries and using Facebook Prophet requirements to fit the model for predicting future attacks for 1 year.
- **Casualty Prediction:** Figure 11 shows the number of fatalities from 1970 to 2020. For predicting the number of casualties for the top 3 countries, two different regression algorithms are being used namely, Random Forest, and Multi-Layer Perceptron (MLP). The best performer is chosen by comparing the Mean Squared Error (MSE) values of these models.

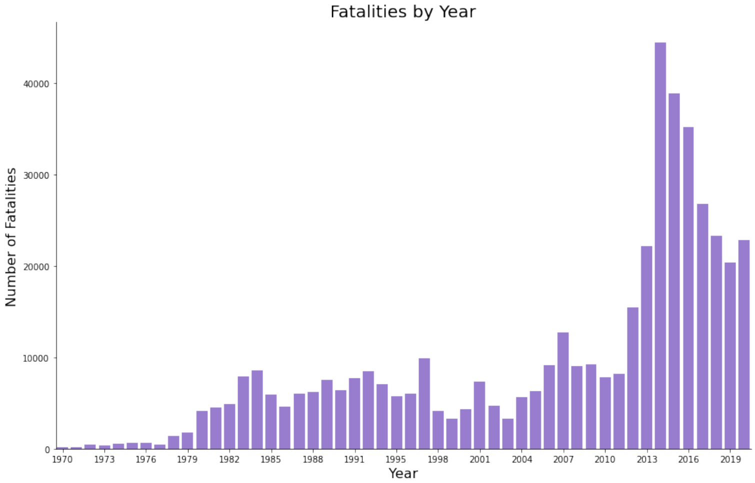


Figure 11: Fatalities By Year

3.5 Interpretation/Evaluation

It is vital to evaluate the predictions made by the models. There are various metrics available to evaluate the results. In this study, the following metrics are used for evaluations:

- **Accuracy:** The Accuracy score is determined by taking the total number of predictions and dividing it by the number of correct predictions obtained.

$$\text{Accuracy} \ = \frac{\text{Nuthe mber of correct predictions}}{\text{Total number of predictions}} \tag{1}$$

$$\text{Accuracy} \ = \frac{\text{TrueNegatives} \ + \ \text{TruePositive}}{\text{TruePositive} \ + \ \text{FalsePositive} \ + \ \text{TrueNegative} \ + \ \text{FalseNegative}} \tag{2}$$

- **Confusion Matrix:** A confusion matrix is a summary of the results of classification problem prediction. The amount of correct and inaccurate forecasts is summarized using count values and grouped by class.

- **Seasonality and Trend Components:** The seasonality and trend components are examined for the time series forecast of future terrorist attacks in Iraq, Afghanistan, and Pakistan.

- **Mean Squared Error:** Estimators are often judged by their Mean Squared Error (MSE), which is the average squared difference between the predicted and actual values. The value of this risk function is proportional to the square of the expected loss from errors.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 \tag{3}$$

  where, MSE = Mean Squared Error, n = number of data points, Y_i = observed values, and $\hat{Y}_i$ = predicted values

4 Design Specification

In the next subsections, the design approach for the study with respect to all prediction tasks is described in detail.

4.1 Weapon Classification

The kNN algorithm is being used to determine what kind of weapon was used in an assault, such as firearms, explosives, chemicals, biological, nuclear, etc. kNN functions by computing the distance between a query and each example in the data, selecting the k examples closest to the query, and then voting for the label with the highest frequency.

4.2 Perpetrator Classification

Decision trees and neural networks are the two major algorithms used to determine which terrorist organization is responsible for an attack. A mixture of algorithms is used by decision trees to decide when and how to split a node into two or more child nodes. The existing subnodes grow more uniform when new ones are generated. MLPClassifier is a Neural Network-based Multi-layer Perceptron classifier. MLPClassifier uses a Neural Network to classify, unlike SVMs or Naive Bayes.

4.3 Time Series Prediction

For forecasting future terrorist attacks in Iraq, Afghanistan, and Pakistan, the FbProphet method is being used. Prophet is a method for predicting time series data based on non-linear trends, annual, monthly, and daily seasonality, and holiday impacts. It works well with seasonal time series and historical data from several seasons.

4.4 Casualty Prediction

Four regression algorithms are employed for the prediction of casualties in the top 3 countries. Random Forest regressor is used to improve the predictive accuracy and avoid over-fitting. To prevent overfitting by penalizing excessively large weights, MLPRegressor makes use of a regularization (L2 regularization) term, the alpha parameter of which controls the degree of the penalty. The Figure 12 shows the working of an MLP with one hidden layer.[3] Ridge and Lasso regression are also implemented. These basic strategies minimize model complexity and avoid linear regression overfitting.

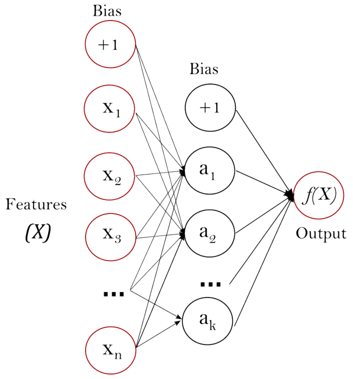


Figure 12: MLP with one hidden layer

5 Implementation

5.1 Tools and Technologies Used

The following requirements for the software and library resources were utilized in order to obtain the results.

- **System type:** 64-bit Windows operating system, x64-based processor
- **Processor:** Intel(R) Core(TM) i5-7200U CPU @ 2.50GHz 2.71 GHz
- **Programming Language:** Python
- **Database management:** Google Drive
- **Integrated Development Environment (IDE):** Google Colab Pro
- **Python Libraries/Modules:** Pandas, Numpy, Sklearn, Matplotlib, Seaborn, NLTK, Plotly, Folium

[3] https://scikit-learn.org/stable/modules/neural$_n etworks_s upervised.html$

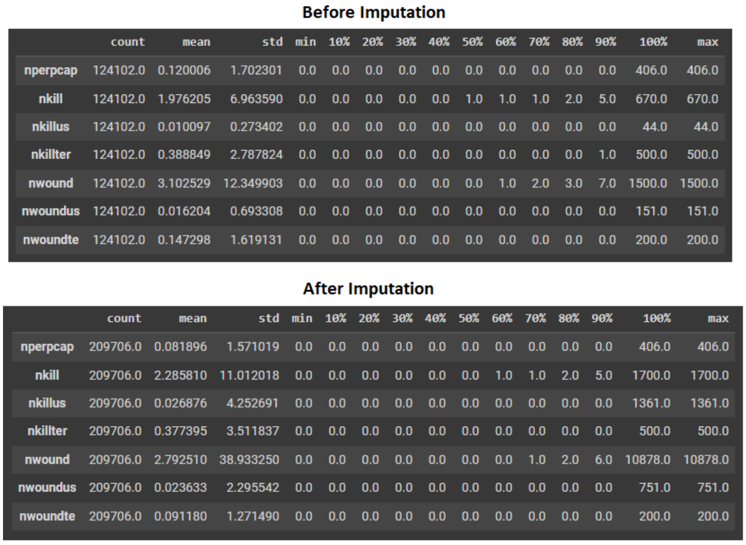
5.2 Data Preparation

Data preparation was the most important step to be carried out for this research since the dataset was inconsistent. The following methods were incorporated to clean and prepare the data.

- Lengthier categorical names are shortened.

- Converting some of the properties of the data frame to categorical form so that they are in line with the GTD code book.[4] By converting attributes to categorical, memory needs may be decreased, and other libraries will be notified to handle the property appropriately. As a consequence, the GTD data frame experiences a drop in percentage equivalent to 36.5

- Attributes with outliers are identified using a threshold of higher than three standard deviations. For imputation purposes, the median is preferred over the mean since the mean is vulnerable to extreme values. A function is used which imputes the median if an attribute contains outliers, otherwise, the mean is imputed. Figure 13 shows the summary statistics of the dataset before and after the imputation.



Before Imputation

	count	mean	std	min	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	max
nperpcap	124102.0	0.120006	1.702301	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	406.0	406.0
nkill	124102.0	1.976205	6.963590	0.0	0.0	0.0	0.0	0.0	1.0	1.0	1.0	2.0	5.0	670.0	670.0
nkillus	124102.0	0.010097	0.273402	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	44.0	44.0
nkillter	124102.0	0.388849	2.787824	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	500.0	500.0
nwound	124102.0	3.102529	12.349903	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	3.0	7.0	1500.0	1500.0
nwoundus	124102.0	0.016204	0.693308	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	151.0	151.0
nwoundte	124102.0	0.147298	1.619131	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	200.0	200.0

After Imputation

	count	mean	std	min	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	max
nperpcap	209706.0	0.081896	1.571019	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	406.0	406.0
nkill	209706.0	2.285810	11.012018	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	2.0	5.0	1700.0	1700.0
nkillus	209706.0	0.026876	4.252691	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1361.0	1361.0
nkillter	209706.0	0.377395	3.511837	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	500.0	500.0
nwound	209706.0	2.792510	38.933250	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	2.0	6.0	10878.0	10878.0
nwoundus	209706.0	0.023633	2.295542	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	751.0	751.0
nwoundte	209706.0	0.091180	1.271490	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	200.0	200.0

Figure 13: Summary Statistics before and after Mean and Median imputation

5.3 Model Building and Training

- **Weapon Classification:** The dataset containing data on terrorist attacks in Iraq, Afghanistan, and Pakistan is split into an 80:20 ratio of train and test sets. The kNN classifier is implemented and the k value is iterated from 1 to 12 to find the best value. The training and test scores are compared.

- **Perpetrator Classification:** The attacks that happened after the year 1995 is considered for this task. The dataset is split into 70% training and 30% test sets.

[4] https://www.start.umd.edu/gtd/downloads/Codebook.pdf

The variables are standardized using StandardScaler function for using the MLPClassifier. DecisionTreeClassifier is also implemented and the performance of both these classifiers are compared.

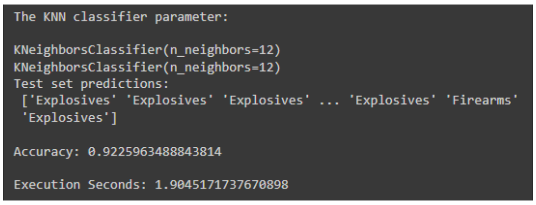
- **Time Series Prediction:** A subset of the dataset containing the past 10 years' data on terrorist attacks in Iraq, Afghanistan, and Pakistan is created for time series analysis. Data smoothing is done using exponential weighted moving average. The public holidays of the past 10 years are factored into the time series model. Attacks for the next 1 year are forecasted, and seasonality and trend components are analyzed.
- **Casualty Prediction:** For the purpose of estimating fatalities in Iraq, Afghanistan, and Pakistan, two supervised models are used: MLPRegressor and RandomForestRegressor. The GridSearchCV method is used to choose the optimal value for the tanh activation function in MLPRegressor's hidden layers, ranging from 1 to 18. The optimal settings for RandomForestRegressor are also determined using the same approach. The models are evaluated using Mean Squared Error.

6 Evaluation

The experiments listed below were conducted, and their outcomes were analyzed using the evaluation metrics outlined in subsection 3.5. Below is a comprehensive discussion of the outcomes of every experiment.

6.1 Weapon Classification using kNN

The kNN classifier is first implemented with k=12 neighbors. The accuracy obtained is 92.25% and the execution time is 1.904 seconds as shown in Figure 14.



```
The KNN classifier parameter:

KNeighborsClassifier(n_neighbors=12)
KNeighborsClassifier(n_neighbors=12)
Test set predictions:
 ['Explosives' 'Explosives' 'Explosives' ... 'Explosives' 'Firearms'
 'Explosives']

Accuracy: 0.9225963488843814

Execution Seconds: 1.9045171737670898
```

Figure 14: Accuracy of kNN model when k=12

The model is then iterated from 1 to 12 to find the optimal value of k. The training and test scores for each of these neighbors are plotted as shown in Figure 15. It is evident from the graph that the accuracy of the model increases as the k value increases.

Another kNN classifier is built by selecting the best k value of 11 neighbors from the data from the previous test. The accuracy of this model shows a slight improvement by 0.0081% and the execution time was faster by 0.072 seconds as shown in Figure 16.

6.2 Perpetrator Classification using Decision Trees and MLP

Two supervised models are implemented to classify the terrorist group responsible for the attacks. get_dummies function is used to convert textual data into numerical values so that the models can use it. GridSearchCV algorithm is used for hyperparameter

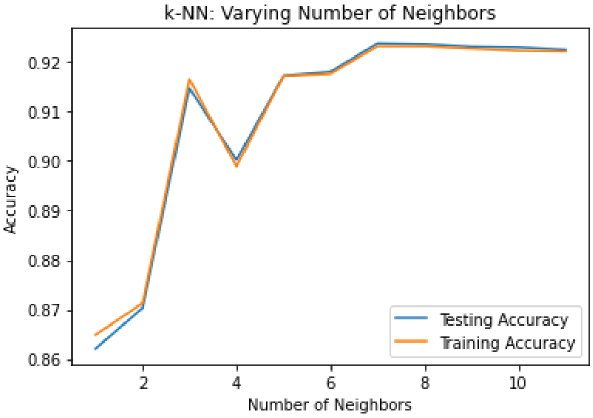
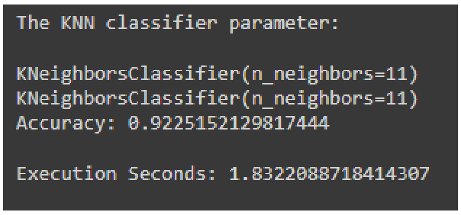


Figure 15: Training and Test scores plot with varying k values



```
The KNN classifier parameter:

KNeighborsClassifier(n_neighbors=11)
KNeighborsClassifier(n_neighbors=11)
Accuracy: 0.9225152129817444

Execution Seconds: 1.8322088718414307
```

Figure 16: Accuracy utilizing previous test's best K of 11 neighbors

tuning. The tuned parameters obtained for the DecisionTreeClassifier are as follows: max_depth=8 and min_samples_leaf=3. The tree generated by the model is shown in Figure 17.

The Confusion matrix of Decision Tree Classifier is shown in Figure 18. It is seen from the matrix that the model was able to correctly classify majority of the groups.

The variables were standardized before implementing the neural network model, i.e., MLPClassifier. The GridSearchCV algorithm is used for hyperparamater tuning with hidden layer sizes equal to (35,35,35). The Confusion matrix for MLPClassifier is shown in Figure 19.

From Figure 20 we can see that Decision Tree performed better than Multi-Layer Perceptron in terms of both Accuracy and Execution time.

6.3 Time Series Analysis using Exponential Weighted Moving Average

For the time series analysis, the text attributes from the dataset are dropped and the numeric attributes are standardized. The time series analysis is done separately on three countries with most terrorist attacks so far, i.e., Iraq, Afghanistan, and Pakistan as described below.

6.3.1 Terrorist Attacks Prediction for Iraq

The Iraq dataset is queried to subset data from the year 2013. The dataframe is reindexed to include all days from 2013 to 2020 and the added days are filled with zero. The data is resampled to plot the daily, weekly, and monthly attacks. The data is smoothened using exponential weighted moving average. The plots are shown in Figure 21.

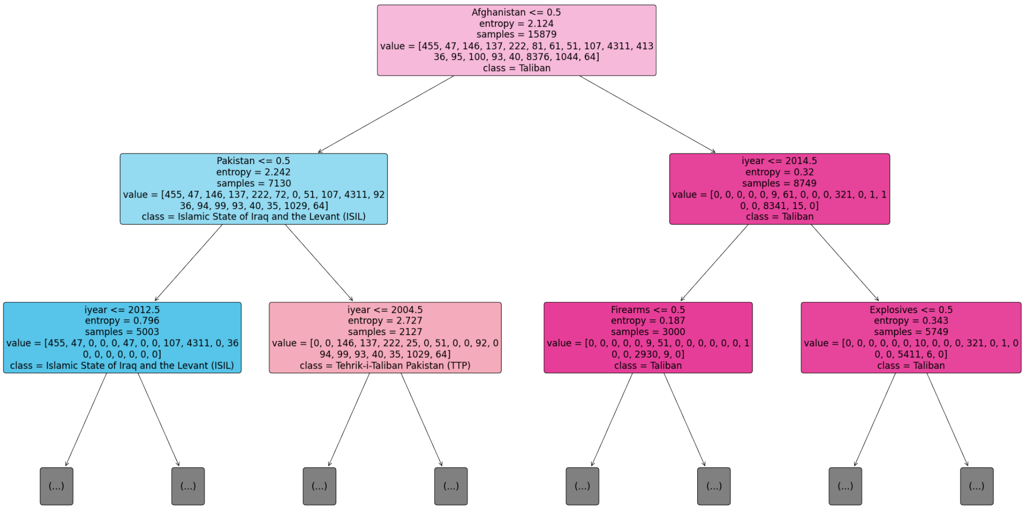


Figure 17: Decision Tree generated for Perpetrator Classification

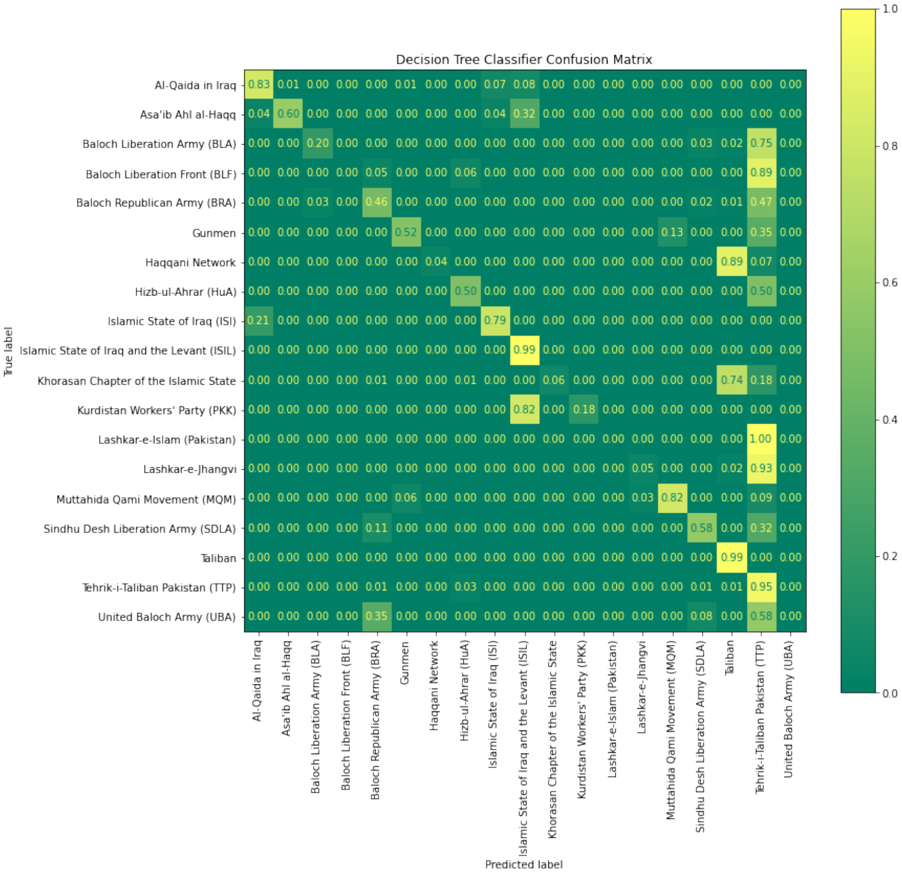


Figure 18: Confusion Matrix of Decision Tree Classifier

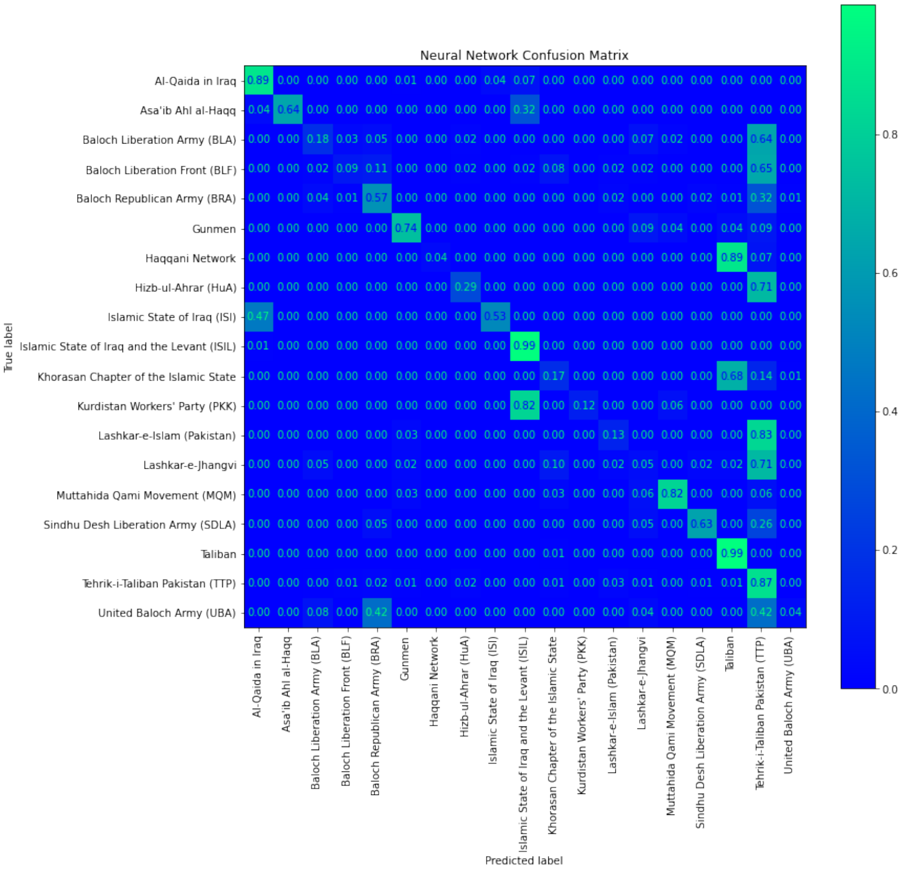
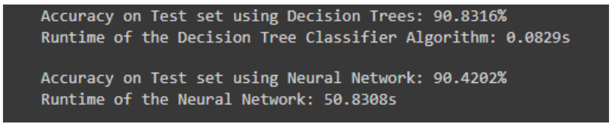


Figure 19: Confusion Matrix of MLP Classifier



```
Accuracy on Test set using Decision Trees: 90.8316%
Runtime of the Decision Tree Classifier Algorithm: 0.0829s

Accuracy on Test set using Neural Network: 90.4202%
Runtime of the Neural Network: 50.8308s
```

Figure 20: Accuracy and Execution time of DT and MLP Classifiers

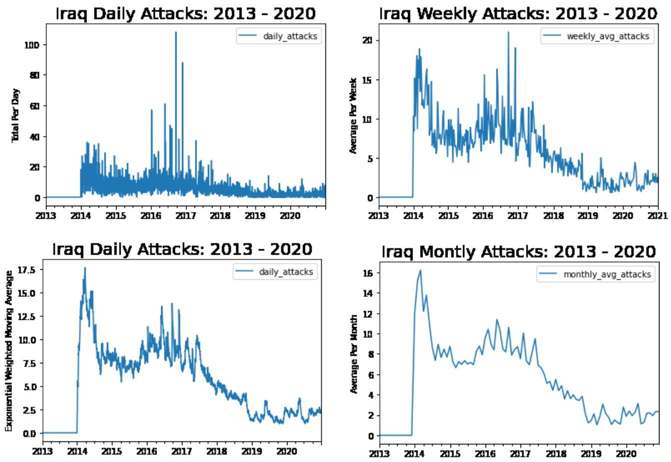


Figure 21: Daily, Weekly and Monthly plots of Terrorist Attacks in Iraq

Iraq holidays dataset is used to factor into the FbProphet model to forecast the attacks between 2021 to 2022 as shown in Figure 22.

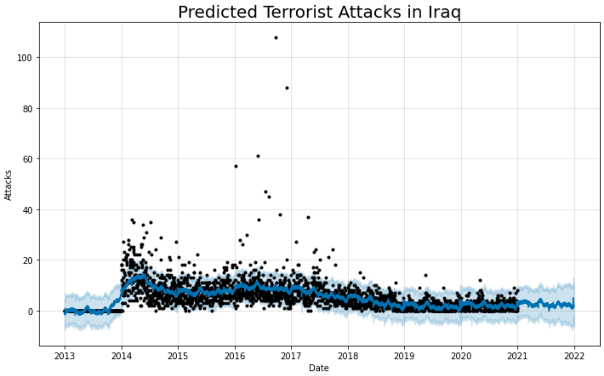


Figure 22: Predicted Terrorist Attacks in Iraq between 2021-2022

Looking at the Seasonality and Trend components in Figure 23, we can conclude that the surge in the month of June happens before Eid ul-Fitr (End of Ramadan), while December's following Mouloud (Birth of the Prophet). September's decline coincides with Eid al-Adha and the Islamic New Year. Friday assaults are minimal due to Friday prayer.

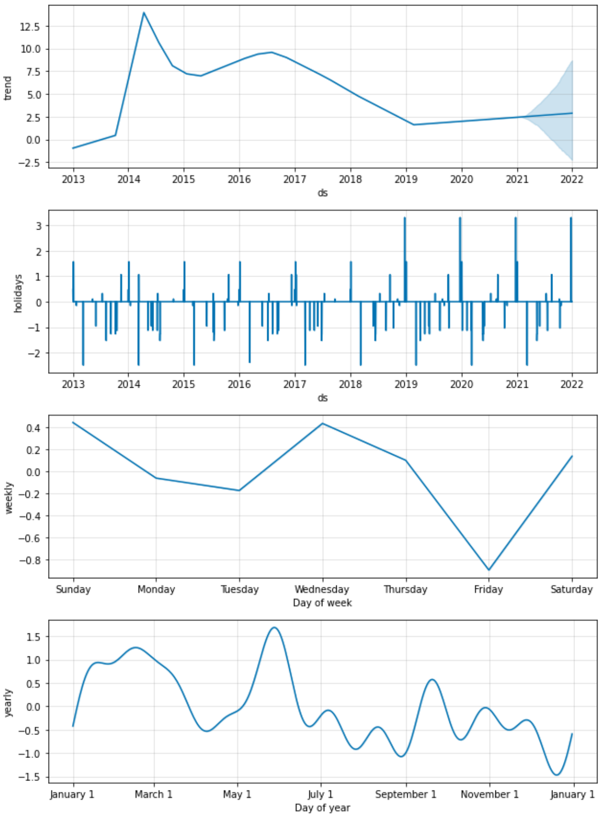


Figure 23: Seasonality and Trend Components of Iraq Terrorism

The pattern that has been predicted in Figure 24 appears to coincide with the annual seasonality that is observed in the component analysis.

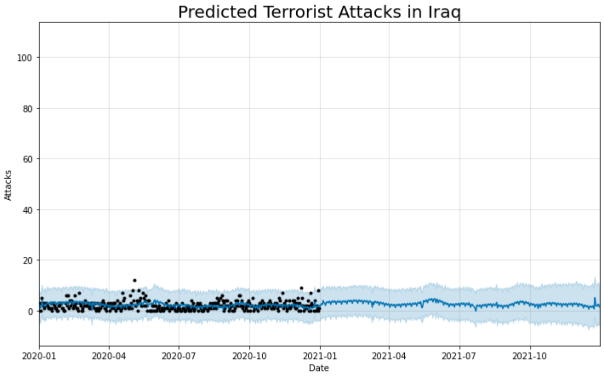


Figure 24: Predicted Terrorist Attacks in Iraq focused from 2020 - 2022

6.3.2 Terrorist Attacks Prediction for Afghanistan

A query is made to the Afghanistan dataset to subset information from 2014. The same process is followed for Afghanistan dataset as Iraq. The predicted terrorist assaults between 2021 and 2022 is shown in Figure 25.

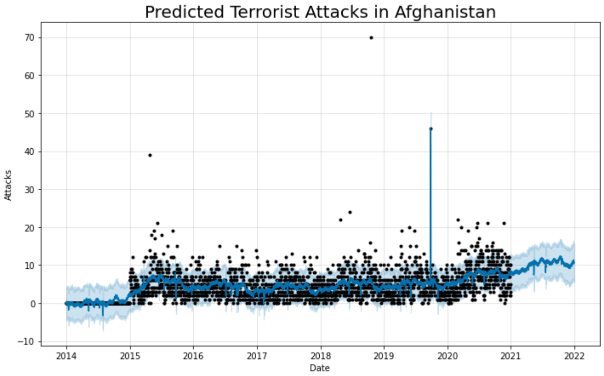


Figure 25: Predicted Terrorist Attacks in Afghanistan between 2021-2022

After taking a look at the Seasonality and Trend components in Figure 26, we can draw the conclusion that the surge that occurs in the month of June occurs prior to Eid al-Fitr (the holiday that marks the end of Ramadan), while the surge that occurs in October happens after Eid al-Qurban.

The projected pattern in Figure 27 seems to match the yearly seasonality shown in the component analysis.

6.3.3 Terrorist Attacks Prediction for Pakistan

Information from the Pakistan dataset is queried and filtered to include only records from 2010. The Pakistan dataset follows the same procedure as the Iraq dataset. Figure 28 depicts the frequency of anticipated terrorist attacks in the years 2021 and 2022.

From analyzing the Seasonality and Trend aspects shown in Figure 29, we learn that the January increase occurs around the time of Eid al-Adha (the Feast of Sacrifice), whereas the May increase is associated with Labour Day. The month of December sees a drop because of Quaid-e-Azam Day (Muhammad Ali Jinnah's birthday).

From Figure 30, we notice a pattern that appears to correspond to the annual seasonality shown by the component analysis.

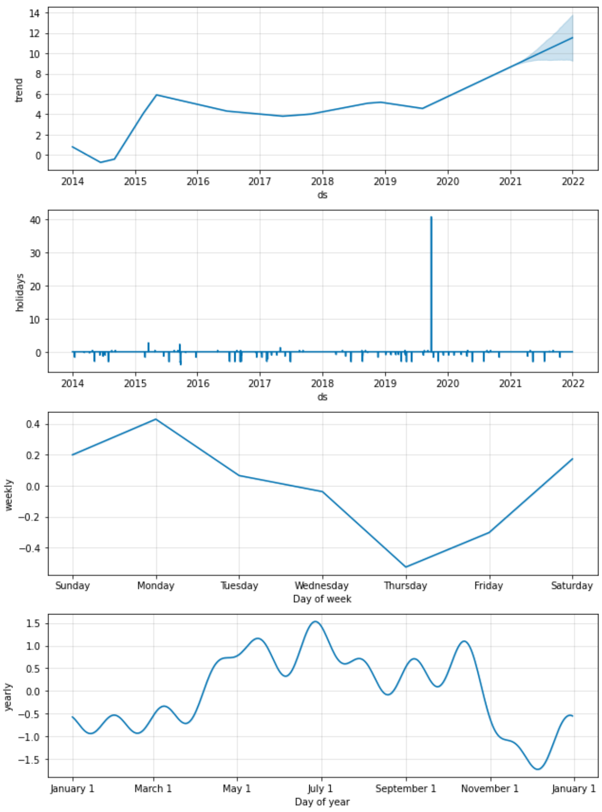


Figure 26: Seasonality and Trend Components of Afghanistan Terrorism

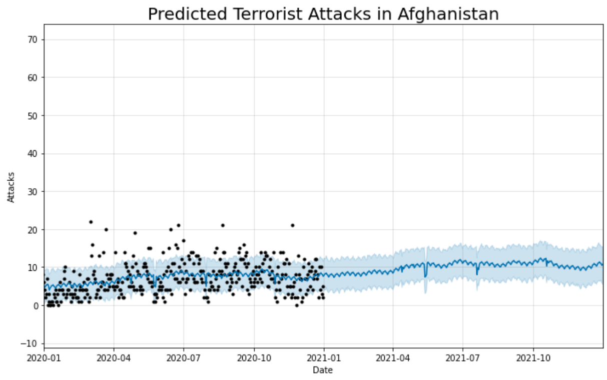


Figure 27: Predicted Terrorist Attacks in Afghanistan focused from 2020 - 2022

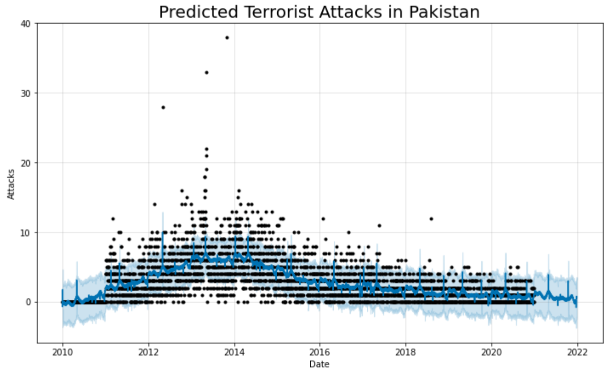


Figure 28: Predicted Terrorist Attacks in Pakistan between 2021-2022

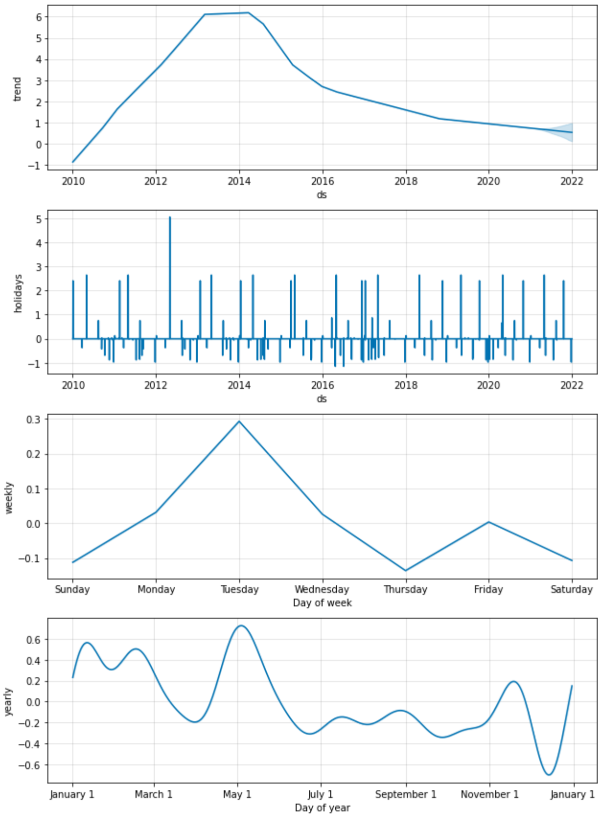


Figure 29: Seasonality and Trend Components of Pakistan Terrorism

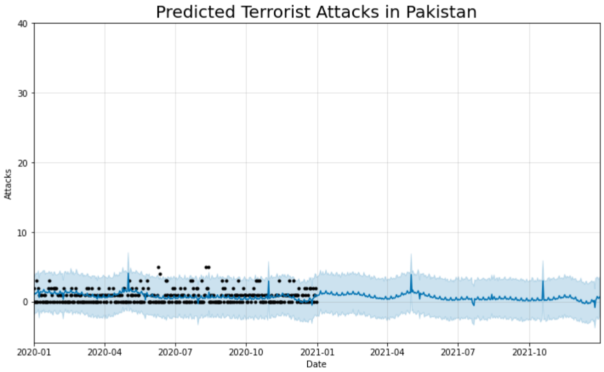


Figure 30: Predicted Terrorist Attacks in Pakistan focused from 2020 - 2022

6.4 Regression Analysis for Casualty Prediction

In order to predict the number of casualties in Iraq, Afghanistan, and Pakistan, one machine learning and one deep learning model are implemented. StandardScaler function is used for feature scaling. GridSearchCV algorithm is used for hyperparameter tuning. RandomForestRegressor and MLPRegressor models are implemented. The Random Forest model outperforms MLP model with a lower MSE value as shown in Figure 31.

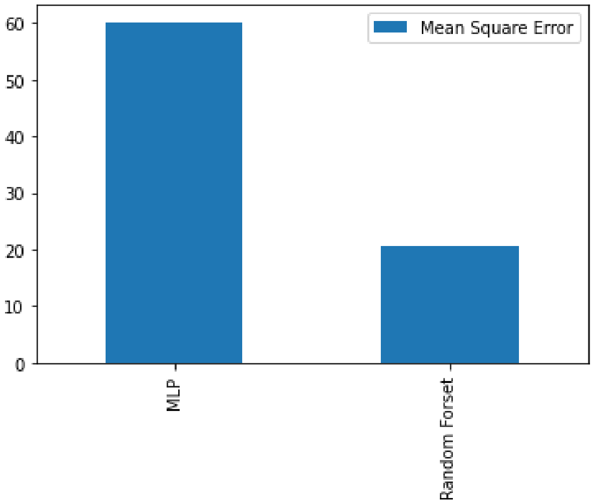


Figure 31: Boxplot of MSE values of RF and MLP Regressors

6.5 Discussion

The kNN classifier used for the weapon classification was able to predict the weapon used in an attack with an accuracy of 92.25% which is better than the accuracy obtained using Random Forest by (Verma et al.; 2018). The Decision Tree classifier utilized to classify terrorist organization behind an attack was able to predict the group name correctly with an accuracy of 90.83%, which outperforms the models implemented by (Talreja and Mahesh; 2017) and (Laite and Sankaranarayanan; 2019). The neural network classifier was also able to provide a better accuracy of 90.42% for predicting terrorist groups as suggested by (Khorshid et al.; 2015).

Due to the availability of data only till 2020, the forecasts using time series analysis were done for the year 2021. From the Figure 27 we can see that the terrorist attacks in Afghanistan was supposed to increase, which actually did happen according to the reports.[5] For predicting the number of fatalities in an attack, the Random Forest algorithm was chosen since it was able to perform really well for the prediction of the success of an attack with an accuracy of 91.82% as per (Alsaedi and Alharbi; 2019). It did perform well compared to the neural network with a lower error rate.

The accuracy of the predictions could be improved in the following ways:

- Collecting other sources of data that could be useful including financial transactions, transportation data, and weather patterns.
- Using a combination of different models to make predictions (ensemble methods), or reinforcement learning.

[5] https://en.wikipedia.org/wiki/List$_of_2021_Afghanistan_attacks$

- Gathering and analyzing information from law enforcement and intelligence agencies can provide important information about known terrorist groups and individuals, as well as ongoing investigations and surveillance.

- Analyzing the conversations and activities of individuals and groups on social media and other online platforms can provide insights into their ideologies, recruitment efforts, and potential plans for terrorist attacks.

- Transfer learning techniques could be leveraged for the prediction of terrorist attacks. Transfer learning is a technique that allows a model trained on one task to be used as a starting point for training a model on a related task. This can improve accuracy by leveraging knowledge learned from the first task to improve performance on the second task.

It is important to note that, accuracy always depends on the data, its quality, amount, and relevance.

7 Conclusion and Future Work

The aim of this study was to analyze the Global Terrorism Database (GTD) to extract valuable insights from it which can be used for the prediction of future terrorist attacks. After exploring the data, the focus of this study was narrowed down to the top 3 most attacked countries: Iraq, Afghanistan, and Pakistan. With the help of Time Series Analysis, the forecasts were made for the year 2021, since the data available in GTD is only till 2020. It is seen that the pattern forecasted was accurate for the year 2021 as the reports show that Afghanistan saw a 42% spike in terrorist attacks in that year. The kNN classifier, which was utilized for the purpose of weapon classification, was capable of predicting the weapon that was used in an assault with an accuracy of 92.25%. The Decision Tree and MLP classifiers performed equally well in predicting the perpetrator of an attack with an accuracy of 90%, but the execution time of DT was faster compared to MLP. The Random Forest regressor was able to predict the number of fatalities in an attack much better than a supervised deep-learning neural network.

Although this research work is comprehensive, there are a few limitations which are mentioned below:

- The research work was focused only on the top 3 most terrorist-prone countries: Iraq, Afghanistan, and Pakistan. The scope of the research can be broadened to include all the counties.

- There is often a lack of comprehensive and reliable data on terrorist attacks, which makes it difficult to train and validate machine learning models. Hence, more fine-grained and relevant data could be collected from various sources.

- Terrorist attacks are complex events that are influenced by a wide range of factors, including political, economic, and social conditions. It can be difficult to capture all of these factors in a machine-learning model.

- The correlation between factors and attack is very complex, it is difficult to understand the causality between them which makes it hard to interpret the results.

- It is very common to overfit the models with the limited data available which may cause models to perform well on the train set but not generalize well on new unseen data.

In conclusion, while machine learning and deep learning techniques have the potential to be useful tools for the analysis and prediction of terrorist attacks, several limitations need to be overcome to achieve accurate and reliable results. Hence, AI and ML play a vital role and can bring fruitful outcomes with extensive research.

Acknowledgement

I would like to express my deepest gratitude to *Dr. Cristina Hava Muntean* for the consistent assistance she has provided throughout each and every step of this research effort. Her guidance and advice have allowed me to gather extensive information, which has resulted in an abundance of ideas and perspectives for my research topic. I also want to express my gratitude to my parents and friends for their unfailing support, understanding, and encouragement.

References

Agarwal, P., S. M. and Chandra, S. (2019). *Comparison of machine learning approaches in the prediction of terrorist attacks. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-7). IEEE.*

Alsaedi, A.S., A. A. and Alharbi, S. (2019). *Mining the global terrorism dataset using machine learning algorithms. In 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA) (pp. 1-7). IEEE.*

Huamaní, E. L., Alicia, A. M. and Roman-Gonzalez, A. (2020). Machine learning techniques to visualize and predict terrorist attacks worldwide using the global terrorism database.
URL: *www.ijacsa.thesai.org*

Khorshid, M. M., Abou-El-Enien, T. H. and Soliman, G. M. (2015). Hybrid classification algorithms for terrorism prediction in middle east and north africa, *International Journal of Emerging Trends & Technology in Computer Science* **4**(3): 23–29.

Lafree, G. (2010). The global terrorism database (gtd): Accomplishments and challenges.

LaFree, G. and Dugan, L. (2007). Introducing the global terrorism database, *Terrorism and Political Violence* **19**: 181–204.

Laite, R., L. M. and Sankaranarayanan, K. (2019). *Terrorist group classification of historic terrorist attacks from the Global Terrorism Database. In 2019 14th International Conference on Computer Science Education (ICCSE) (pp. 237-242). IEEE.*

Li, Z., Li, X., Dong, C., Guo, F., Zhang, F. and Zhang, Q. (2021). Quantitative analysis of global terrorist attacks based on the global terrorism database, *Sustainability (Switzerland)* **13**.

Singh, K., Chaudhary, A. S. and Kaur, P. (2019). A machine learning approach for enhancing defence against global terrorism, *2019 Twelfth International Conference on Contemporary Computing (IC3)*, pp. 1–5.

Talreja, D., N. J. V. N. and Mahesh, K. (2017). *Terrorism analytics: Learning to predict the perpetrator. In 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI) (pp. 1723-1726). IEEE.*

Tolan, G. M. and Soliman, O. S. (2015). An experimental study of classification algorithms for terrorism prediction, *International Journal of Knowledge Engineering-IACSIT* **1**: 107–112.
URL: *http://www.ijke.org//show-36-53-1.html*

Verma, C., Malhotra, S., Verma, V. and Patel, S. V. (2018). Naïve bayes (nb), *Random International Journal of Pure and Applied Mathematics* **119**: 49–61.