MULTIDISCIPLINARY SCIENCE JOURNAL

RESEARCH PAPER

Published Online: July 14, 2023
https://doi.org/10.31893/multiscience.2024011

Artificial Intelligence Approach for Terror Attacks Prediction Through Machine Learning

Sagar Shinde[a] | **Suchitra Khoje**[b] | **Ankit Raj**[c] | **Lalitkumar Wadhwa**[d] | **Armaan Suhel Shaikh**[a]

[a]Department of Electronics & Telecommunication Engineering, Dr. D. Y. Patil Institute of Technology, India.
[b]CS2S Technology Pvt. Ltd., India.
[c]MIT World Peace University, India.
[d]Dr. D. Y. Patil Institute of Technology, India.

Abstract The rise of terrorism has become a global threat that affects every continent and country. Due to the complexity of the situation, and the technological developments that have occurred in the field, attacks and incidents have become more difficult to manage. To prevent these types of attacks, countries and organizations have started using various techniques such as machine learning and artificial intelligence.The goal of this paper is to create a framework that allows users to visualize the various characteristics of terrorist attacks by extracting data from the Global Terrorism database. The graph embedding process is performed by using two different methods.The coding approach was used for the non-coding model development. The data collected by the project were then fed to seven different machine learning models. These models included Random forest, KNN, adaboost, decision trees, and SVM. The classification accuracy of the two methods was estimated to be around 90%.

Keywords: graph database, graph embedding, GTD, prediction, node2vec

1. Introduction

Terrorism is regarded as one of the most common and significant threats to international security. It affects all countries and continents. The private and public sectors have been tasked with addressing the issue of terrorism and its effects.According to the Global Terrorism Index, which measures the impact of terrorism, the number of people who died due to attacks in 2020 was 22,847. This represents a significant increase from the previous decade when the average annual death toll was around 26,000. In 2023, the report noted that the number of attacks has increased by 26%. Analyzing past attacks and forming connections between them can help in predicting future patterns.

Due to the complexity of terrorist activities, the data collected about these attacks have been subjected to various variations.The importance of maintaining a comprehensive database regarding terrorist activities has been acknowledged. Over the years, various organizations have created specialized platforms that document the evolution and trajectory of attacks. The emergence of these tools has allowed researchers to utilize deep learning and machine learning to gain deeper insights from this information.The Global Terrorism Database (GTD), which is maintained by the University of Maryland, is a repository of information about terrorist activities spanning the globe since 1970. It has over 2,00,000 entries and features over 120 attributes. The database is a public record and contains details about every known incident since 1970.

Since the nature of terrorist activities is so sensitive, it is important that the database used for research maintains a high level of cleanliness and consistency. Unfortunately, public domain databases tend to suffer from inconsistency. By utilizing graph databases, researchers can expand the scope of attributes in the repository and improve its model and reliability.Deep learning and machine learning techniques have been utilized to identify patterns in the data collected by the GTD (Uddin et al 2020). They can also be used to establish connections between attacks. In a study, the researchers proposed using the Cypher Query language to convert a relational database into a graphical one.

Neo4j's Sandbox platform comes with a variety of plugins. Some of these include the APOC library and the GDS library, which can be used to extract important graph properties.The researchers utilized the GTD to implement the suggested method and extracted various features from it. The node2vec method is used to generate low-dimensional representations of complex graphs. It's based on a random walk method, which chooses a future node after examining the previous ones.The researchers used various machine learning models to categorize terror incidents. Some of these include RandomForest, Naive Bayes, KNN, AdaBoost, and SVM. They were able to create these models using the Python programming language and Orange data mining software.

 Received: March 25, 2023 | Accepted: June 26, 2023

2. Background and Related Work

As terrorism is a topic of worldwide concern, it has witnessed essential developments with an increase in the number of terrorist databases. Researchers and scholars have been actively engaged in enhancing the capabilities of predictive models and analytical tools to detect or identify patterns associated with terrorism. In 2021, Mohammed Abdalsalam et al extracted textual features from GTD using three techniques namely Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (Bow), and Word Embedding (W2vec). Textual features were converted to vectors and fed to classifiers of nine machine learning models for classifying the types of terrorist attacks. The highest accuracy of 95.3% was achieved by W2vec using Bag classifier (Abdalsalam et al 2021).

Uddin et al (2020) constructed five models based on neural networks (NN) and deep neural networks (DNN) which used logistic regression, SVM, and Naive Bayes as machine learning algorithms. The predictions of suicide attacks and success rate among the attacks were predicted using confusion matrices. The authors achieved 86% accuracy with NN and 92% accuracy with DNN. Olabanjo et al (2021) achieved an accuracy of 99% by employing an ensemble machine learning models that combined support vector machine and a K-nearest neighbour for predicting continents most susceptible to terrorist attacks. The ensemble models were applied to features selected using Chi-squared, Information Gain and a hybrid technique which was a combination of Chi-squared and Information Gain. The hybridized model performed with the highest accuracy of 97.81%.

Feng et al (2020) compared and examined other methods, but they focused on the XGBoost approach since it allowed them to attain an accuracy of 80% in predicting whether terror attacks will result in civilian casualties. The authors devised a method incorporating random forest (RF) and principal component analysis (PCA) for selecting feature set. Out of a total of 136 characteristics, Pan (2021) selected 36 for classification in a framework that consisted of five classifier prediction models, with the XG Boost classifier attaining the greatest accuracy of 98% in predicting terrorist organizations with the highest attack frequencies.

Random Forest and decision trees, two machine learning models utilized by Huamani et al (2020), yielded accuracy rates of 90.414 percent and 75.45 percent, respectively. Sarker et al (2020) attempted to apply KNN and SVM using two datasets. They were able to achieve an accuracy of 97.1 percent and 89.4 percent using KNN for datasets 1 and 2, respectively. The authors employed SVM for datasets 1 and 2 and were able to achieve accuracy rates of 97.7 percent and 92.4 percent, respectively. However, there is no information on the results of employing the KNN model; just the method for calculating the result is given. Five categorization models were put into use by Kumar et al (2020), and each one had an accuracy rating of 90% or above. They tested their model using a huge quantity of information, practically the whole global terrorist database.

The analysis and review of existing literature reveals that current systems predominantly employ established techniques for terrorism classification and mostly focus on textual data conversion for establishing feature sets (Shinde et al 2022). However, an extensive review of the literature did not yield any instances where classification has been conducted using graph databases and graph embeddings. This notable gap in the literature highlights the potential for novel approaches utilizing graph-based methodologies in the classification and prediction of terrorism. By leveraging the structural relationships and inherent connectivity within graph databases, such techniques offer promising avenues for enhancing the accuracy and effectiveness of terrorism classification systems.

3. Methodology

The goal of this paper is to create a framework that enables the creation of graphs and data cleansing, as well as testing the effectiveness of a machine learning algorithm in predicting terrorist attacks. The study utilizes two coding and non-coding techniques to address the challenges of data pre-processing. Graphical representation is then utilized to establish the relationships between the various attributes in the database.

3.1 System Block diagram

Figure 1 shows the proposed system's block diagram. The CSV file containing the dataset is then exported to Neo4j.The platform uses the query language known as the Cypher Query to generate a graph. The nodes' connections are then used to set up subgraphs, which are then extracted and calculated with the help of the graph features. Before the data is loaded into the model and pre-processing pipeline, the GTD and graph features are integrated. The data preparation phase involves scaling, data imputation, data standardization, data reduction, and more.

The next step is to train and evaluate the model using the collected data. This process involves splitting the data into test and training sets (Sardeshmukh et al 2023). There are numerous tools that can be used to create and manipulate graph-based datasets. This project uses the Neo4j Sandbox tool to build a database and the Query Language to modify the structure of the graph.

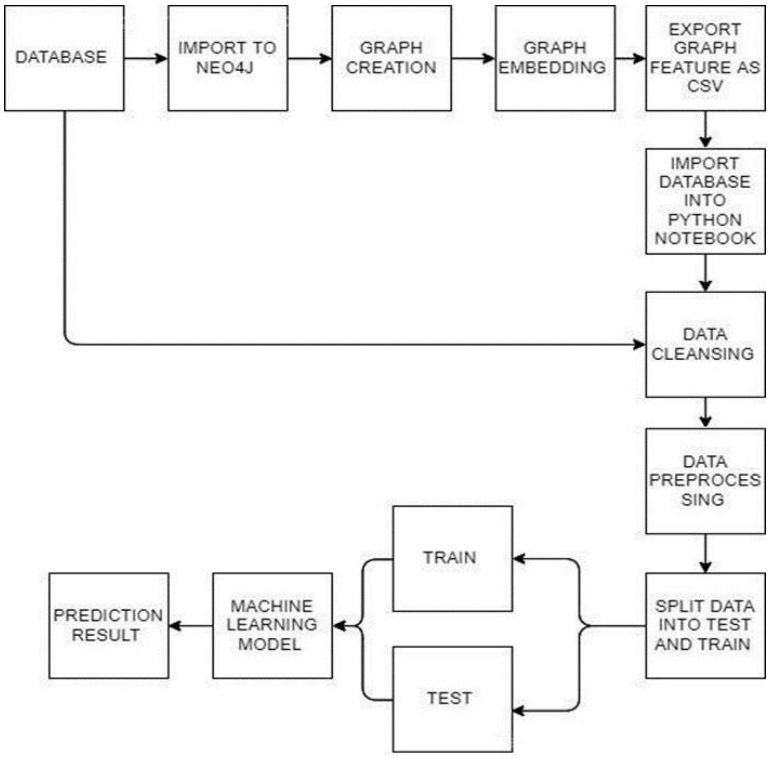

Figure 1 Block diagram of the proposed system.

3.2 Graph Embedding & Graph Creation

The LOAD CSV cypher command was used to create and expand the graphs in Neo4j. The information contained in the event data was organized into attack types, such as attack type 1, attack subtype 2, and attack sub 2. Event details also included the date, time, and location of the occurrence. The target type was composed of target types 1, 2, and 3. The weapon type was also categorized into weapon types 1, 2, and 3. The effects of certain elements, such as causality, were linked to the nodes.

The graph presented in Figure 2 was made and formatted using the Cypher Query Language. It has a limit of 100.The embedding procedure is performed with the help of Node2Vec. It takes into account the attributes and columns in a relational database. The rows and columns in the database are represented by the connections between the nodes.

3.3 Embedding Calculation

The representation of graph data in a lower dimension can be achieved through the use of embedding techniques, which are usually vector-based. Neo4j can implement these methods by using three different methods: node2vec, graph Sage, and fast random projection. In this study, we present a method that takes advantage of random network walks to calculate the embedding. For a graph to be represented in a lower dimension, it should be embedded using a vector representation.

The node2vec method is used to generate a list of nodes with identities and then combine these into a set of phrases. It is then used to calculate the embedding vectors. The approach makes a transition between two search strategies based on random walks. In a network, Neo4j can handle the computation of embedding between two nodes up to n dimensions, although this study only counted ten dimensions.

3.4 Relationship Criteria

The graph database is designed to visualize and implement various graph building blocks. It features interactive features such as relationship attributes, nodes, and cypher queries.

The Neo4j Sandbox tool was utilized to create the Global Terrorism Database as a graph database. After importing the dataset in CSV format, the program was able to split the data into seven groups. These include event time, event location, attack type, causality, target type, and weapon type. The Figures 3 and 4 show the graphs with nodes and relationships established.

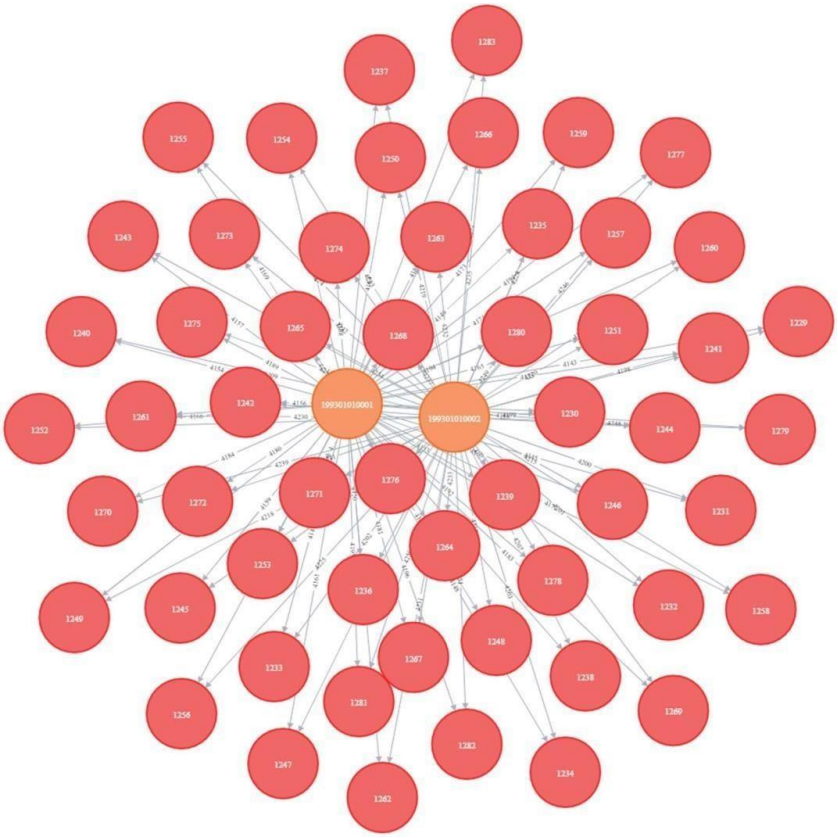

Figure 2 Graph of Attack Id with respect to Event Id.

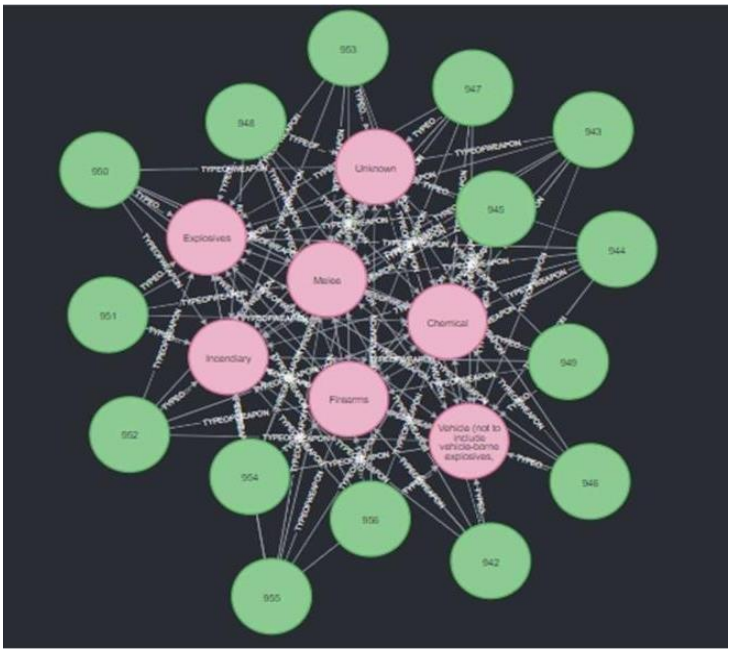

Figure 3 Graph showing nodes and relationships.

3.5 Data Pre-processing

The non-coding method was used to remove the null values from the dataset. The Orange data mining tool can perform various types of imputation. These include random value, model-based, fixed value, average, and most frequent. The most frequent and average imputation algorithms were utilized to exclude the null values. The data mining tool of Orange was able to perform a purge domain block to remove redundant attributes from the dataset. It offers three options to choose from, and it can perform various tasks to decrease, remove, and sort the features.

Figure 4 Graph showing common types of attacks in the Dominican Republic and Mexico.

One of the most common issues that researchers encounter when it comes to developing machine learning systems is overfitting (Khoje and Shinde 2023). This occurs when the training model starts to learn the underlying relationships and patterns instead of learning the details of the training (Shewale et al 2023). Two of the techniques that are commonly used to address this issue are LDA and PCA. In this study, the researchers utilized the PCA technique in the data mining tool to minimize the dimensions of the two components.

The researchers utilized the data processor block of the Orange data mining tool to perform scaling to reduce the size of the data. This process helps the model learn and categorize the data more quickly. Figure 5 shows the data table block in the Orange data mining tool.

The researchers utilized a coding-based approach to create a global terrorism database and a graph that is embedded in a CSV file. They then combined the two pieces of data using a Python script. The resulting dataset has over 136 attributes. To remove the null values, the researchers conducted a count of all the attributes in the data set. They were able to eliminate more than 35% of the null value count.

The researchers were able to maintain the remaining properties in the dataset if the number of null values was less than 35%. They performed data scaling with the usual approach.

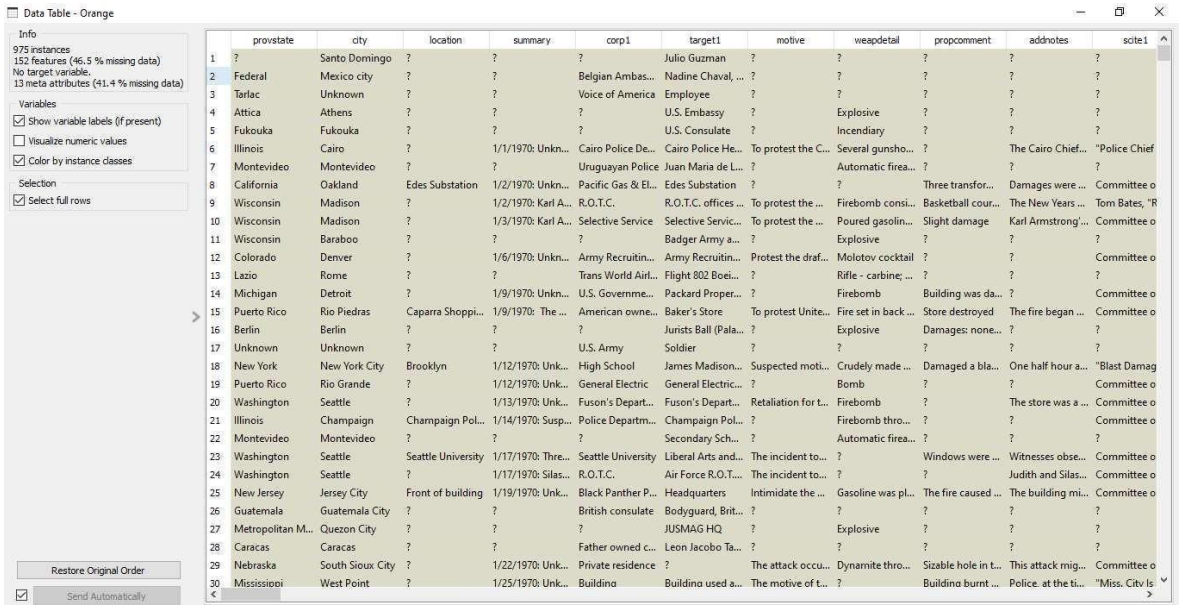

Figure 5 Data table block in Orange tool.

4. Results and Discussion

The goal of the machine learning models was to develop a representation of the various datasets that were imported into the orange mining tool using the CSV import block. The first two datasets that were imported were global terrorism and calculated embedding. A Data table block was used to input the data into the merge data block.

Due to the large number of null values in the combined dataset, an imputation block was utilized to remove them. The properties of the dataset can be selected and the target variable can be set. An 8:2 sampling ratio was then used to split the data into tests and training. The training and testing phases were carried out with a sample ratio of 8:2. About 80% of the collected data was utilized for training, while 20% was used for testing. A confusion matrix was then generated using a Python script to display the results of the learning model. Figure 6 shows the confusion matrix generated for Random Forest model, Figure 7 shows the confusion matrix for Gradient Boosting model, Figure 8 shows the confusion matrix for KNN, Figure 9 shows the confusion matrix for Decision Tree model and Figure 10 shows the confusion matrix for SVM model. Similarly, Figures 11 and 12 show the confusion matrices generated for Naïve Bayes and Adaboost models respectively.

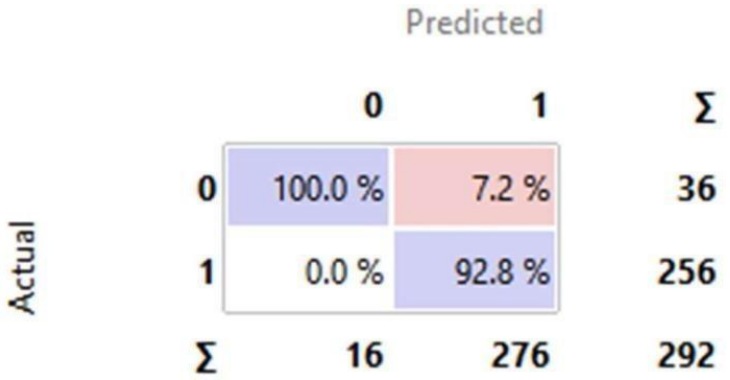

Figure 6 Confusion Matrix for Radom Forest

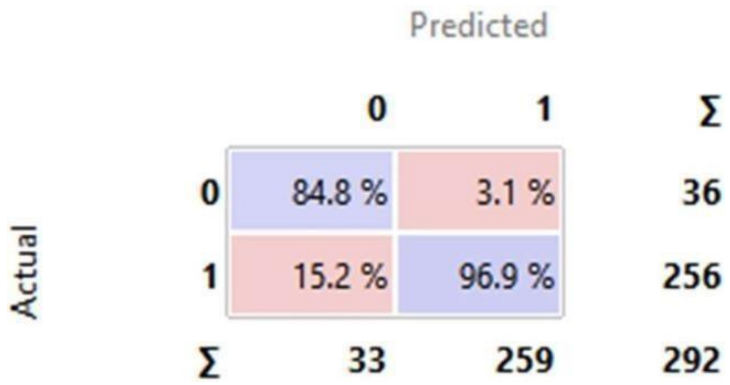

Figure 7 Confusion Matrix for Gradient Boosting.

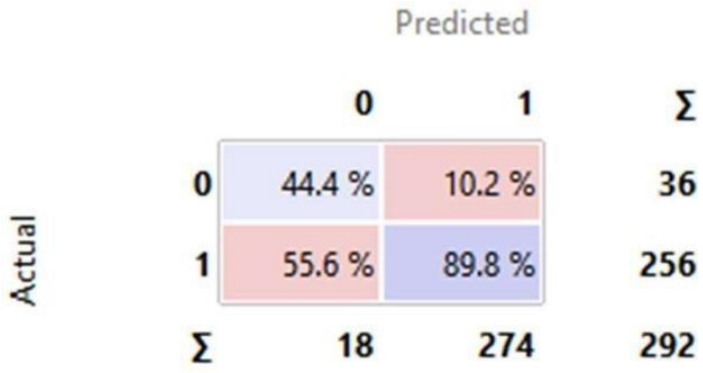

Figure 8 Confusion Matrix for KNN.

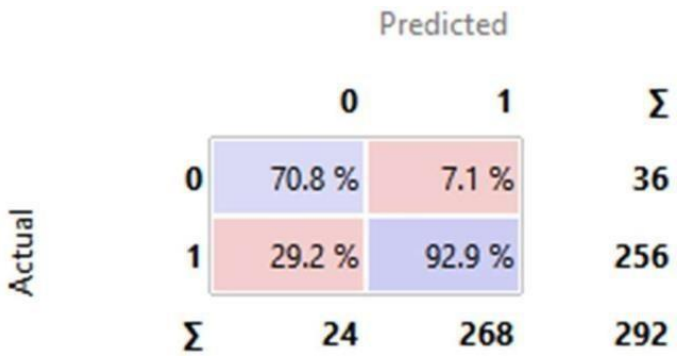

Figure 9 Confusion Matrix for Decision Tree.

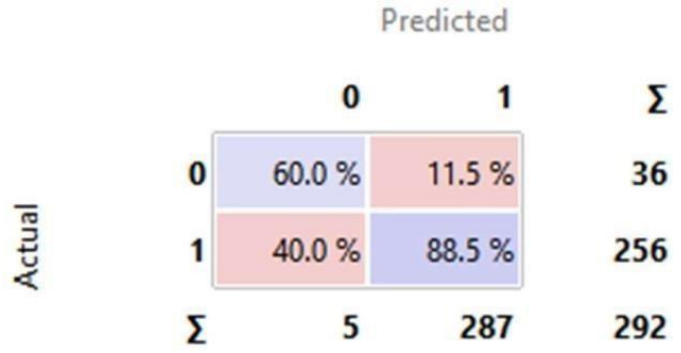

Figure 10 Confusion Matrix for SVM.

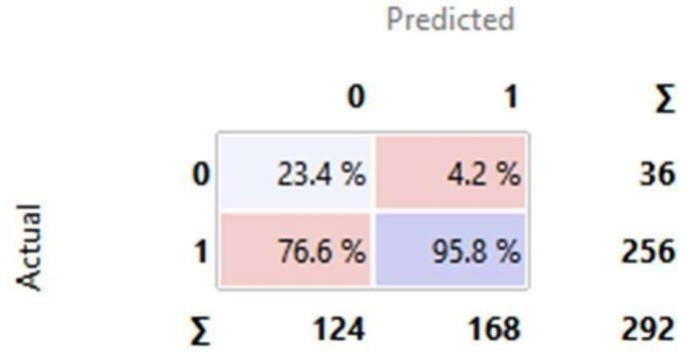

Figure 11 Confusion Matrix for Naïve Bayes.

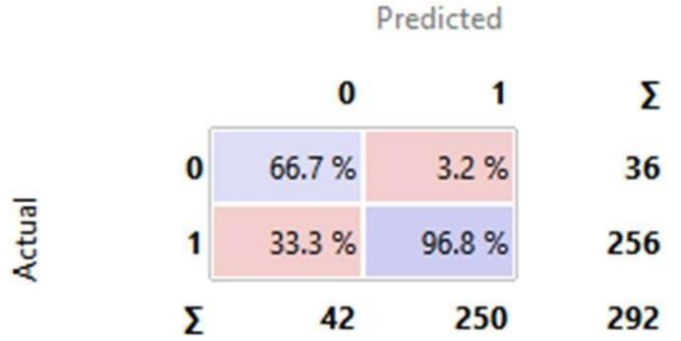

Figure 12 Confusion Matrix for Adaboost.

Table 1 shows the performance parameters for Random Forest, gradient boosting, Naive Bayes, decision trees, AdaBoost, K-nearest neighbour (KNN), and Support Vector Machine (SVM). The F1 score which is a combination of precision and recall and measures the model's accuracy shows the highest value for gradient boosting and the lowest for Naive Bayes.

Table 1 Performance parameters for different models.

Model	AUC	F1 Score	Precision	Recall
RF	0.938	0.920	0.936	0.932
GB	0.931	0.955	0.954	0.955
KNN	0.743	0.850	0.842	0.870
Tree	0.572	0.903	0.902	0.911
SVM	0.554	0.830	0.850	0.880
NB	0.833	0.710	0.869	0.651
Adaboost	0.862	0.927	0.931	0.925

5. Conclusion and Future work

The goal of this study was to analyze the graphical representation created from a relational database using seven different machine learning models. It was done by reviewing the matrix of confusion for the different models. The accuracy of a model is computed by analyzing the value of its confusion matrix. The results of the study indicate that gradient boosting is the most suitable model for this type of dataset.

Although the evaluation of the embedded models yielded promising results, they need to be tested on the latest and real-world datasets. The accuracy of the machine learning model is still promising, but it can be improved through the use of deep learning techniques such as Recurrent Neural Networks and graph adversarial networks (Kadam et al 2022; Shinde et al 2021).

The system can be implemented on diverse and significant datasets to ensure its validity. For instance, the work done on global terror attack predictions was carried out in the GTD relational database. The future of terrorist activity prediction is bright with the advent of deep learning and graph databases. The results of the evaluation can be tested on various graph databases, such as Redis Graph and Tiger Graph. These platforms can extract various characteristics, which can be used to train deep learning models.

Ethical considerations

Not applicable.

Conflict of Interest

The authors declare no conflicts of interest.

Funding

This research did not receive any financial support.

References

Abdalsalam M, Li C, Dahou A, Noor S (2021) A Study of the Effects of Textual Features on Prediction of Terrorism Attacks in GTD Dataset. Engineering Letters 29:2.

Feng Y, Wang D, Yin Y, Li Z, Hu Z (2020) An XGBoost-based casualty prediction method for terrorist attacks. Complex Intell Syst 6:721–740. DOI: 10.1007/s40747-020-00173-0

Global Terrorism Index. Available in: https://www.visionofhumanity.org/wp-content/uploads/2023/05/GTI-2023-web-190523.pdf. Accessed on: May 20, 2023.

Huamani EL, Mantari A, Roman-Gonzalez A (2020) Machine learning techniques to visualize and predict terrorist attacks worldwide using the global terrorism database. Int J Adv Comput Sci Appl 11. DOI: 10.14569/ijacsa.2020.0110474

Kadam SU, Shinde SB, Gurav YB, Dambhare SB, Shewale CR (2022) A novel prediction model for compiler optimization with hybrid meta-heuristic optimization algorithm. Int J Adv Comput Sci Appl 13. DOI: 10.14569/ijacsa.2022.0131068

Khoje S, Shinde S (2023) Evaluation of ripplet transform as a texture characterization for Iris recognition. J Inst Eng (India) Ser B 104:369–380. DOI: 10.1007/s40031-023-00863-6

Kumar V, Mazzara M, Messina A, Lee J (2020) A conjoint application of data mining techniques for analysis of global terrorist attacks: Prevention and prediction for combating terrorism. In: Advances in Intelligent Systems and Computing. Cham: Springer International Publishing 146–158.

Olabanjo OA, Aribisala BS, Mazzara M, Wusu AS (2021) An ensemble machine learning model for the prediction of danger zones: Towards a global counter-terrorism. Soft Computing Letters 3:100020. DOI: 10.1016/j.socl.2021.100020

Pan X (2021) Quantitative analysis and prediction of global terrorist attacks based on machine learning. Sci Program 2021:1–15. DOI: 10.1155/2021/7890923

Sardeshmukh M, Chakkaravarthy M, Shinde S, Chakkaravarthy D (2023) Crop image classification using convolutional neural network. Multidisciplinary Science Journal 2023039. DOI: 10.31893/multiscience.2023039

Sarker A, Chakraborty P, Sha SMS, Khatun M, Hasan MR, Banerjee K (2020) Improvised technique for analyzing data and detecting terrorist attack using machine learning approach based on twitter data. J Comput Commun 8:50–62. DOI: 10.4236/jcc.2020.87005

Shewale C, Shinde SB, Gurav YB, Partil RJ, Kadam SU (2023) Compiler optimization prediction with new self-improved optimization model. Int J Adv Comput Sci Appl 14. DOI: 10.14569/ijacsa.2023.0140267

Shinde S, Wadhwa L, Bhalke D (2021) Feedforward back propagation neural network (FFBPNN) based approach for the identification of handwritten math equations. In: Advances in Intelligent Systems and Computing. Cham: Springer International Publishing 757–775. DOI: 10.1007/978-3-030-51859-2_69

Shinde SB, Alagirisamy M, Wagh K, Dhore P (2022) Math accessibility for blind people in society using machine learning. ECS Trans 107:18071–18090. DOI: 10.1149/10701.18071ecst

Uddin MI, Zada N, Aziz F, Saeed Y, Zeb A, Ali Shah SA, Al-Khasawneh MA, Mahmoud M (2020) Prediction of future terrorist activities using deep neural networks. Complexity 2020:1–16. DOI: 10.1155/2020/1373087