

Deep Ensemble Hybrid Model for Extremism Detection and Threat Inference in Counter-Terrorism Intelligence on Social Media

Gideon Mwendwa*

National Forensic Sciences University, School of Doctoral Studies and Research, 403401, Goa, India

E-mail: gideon.phdcs21@nfsu.ac.in

ORCID iD: <https://orcid.org/009-0000-4616-4543>

*Corresponding author

Lokesh Chouhan

National Forensic Sciences University, School of Cyber Security and Digital Forensics, 403401, Goa, India

E-mail: lokesh.chouhan@nfsu.ac.in

ORCID iD: <https://orcid.org/0000-0003-2168-5267>

Ranjit Kolkar

National Forensic Sciences University, School of Cyber Security and Digital Forensics, 403401, Goa, India

E-mail: ranjit.kolkar@gmail.com

ORCID iD: <https://orcid.org/0000-0001-6835-7821>

Received: January 9, 2026; Revised: March 25, 2026; Accepted: June 7, 2026; Published: August 8, 2026

Abstract: Social media's worldwide expansion over the past two decades has significantly altered the dissemination of extremist narratives, creating both challenges and opportunities for counterterrorism efforts. Addressing critical gaps in the detection and classification of extremist content on social media platforms, this research supports earlier-stage analytical assessment for law enforcement and security agencies. Using datasets from the publicly available Global Terrorism Database (GTD, $n > 209,000$ incidents) and a curated corpus of labeled tweets ($n = 17,410$), a hybrid framework integrating machine learning and deep learning models through a late-fusion stacking architecture is developed. The proposed ensemble leverages contextual indicators derived from historical terrorism data alongside linguistic and behavioral signals from social media content to distinguish extremist from non-extremist activity. Evaluated under strict temporal validation, the model achieves an accuracy of 98.52%, precision of 97.01%, recall of 99.66%, and an AUC of 0.92 under controlled experimental conditions. To address ethical and transparency considerations, Shapley Additive exPlanations (SHAP) are employed to enhance interpretability in automated decision-making. While the reported results reflect dataset-specific evaluation, the findings demonstrate that integrating historical terrorism intelligence with temporally ordered social media analysis can support counterterrorism efforts by mitigating digital radicalization pathways and associated downstream physical security risks linked to terrorism and extremism through earlier analytical intervention.

Index Terms: Extremism, Shapley Additive exPlanations (SHAP), Hybrid Model, Machine Learning, Counter-terrorism.

1. Introduction

Terrorism remains a global threat, constantly adapting its approaches to technological advancements and changes in the geopolitical landscape. Social media platforms have become of paramount interest to terrorist organizations as platforms for propaganda, radicalization, and recruitment, and are increasingly used for the coordination of attacks [1, 2]. They enable users to interact with a large audience, bypass traditional media gatekeepers, and disseminate messages swiftly and widely [3]. The widespread creation and anonymous dissemination of extremist content across social media platforms poses profound challenges to national security and public safety [4]. It is imperative to develop effective counterterrorism strategies; this involves innovative approaches to the rising, rapidly expanding threat [5]. Traditional methods of intelligence gathering, relying mostly on human intelligence (HUMINT) and signals intelligence (SIGINT),

often cannot keep pace with the rapidly changing world of online extremism [6]. These methods are often slow, resource-intensive, and cannot perform an analysis of the massive unstructured data produced on social media platforms [7]. The vast amount of information can make it difficult to detect subtle signs of extremist activities. The situation is further exacerbated due to the inability to detect coded language, contextual nuances, and the shifting tactics employed by extremist groups. This research gap points to the dire need for innovative approaches that combine the best of advanced machine learning techniques with solid data sources to improve the accuracy and speed of extremism detection and prediction [8]. Most of the existing studies make use of just one data source, do not utilize the deep learning models, or do not account for the narrative dynamics of extremist groups sufficiently to address the need for a more holistic and integrated framework [9].

The novelty of this research lies in the proposed hybrid framework, which integrates structured and unstructured data processing pipelines to enhance the detection, classification, and prediction of extremist content on social media. The model incorporates ensemble learning through a stacking classifier, improving robustness and overall detection performance. In addition, the framework supports predictive inference by learning from historical terrorism data and temporally ordered social media signals to assess the likelihood that newly emerging online content reflects extremist intent. By incorporating SHAP values within an explainable AI framework, the model's predictions are rendered transparent and interpretable, addressing key trust and accountability concerns associated with automated decision-making processes. These contributions help remediate gaps in existing methodologies and provide actionable analytical insights that can support law enforcement and intelligence agencies in counterterrorism decision-making as digital ecosystems continue to evolve.

This research proposes the development and validation of a hybrid model for extremism detection on social media platforms that integrates deep learning architectures with ensemble methodologies. The study addresses three interconnected research objectives: to develop, implement, and evaluate a comprehensive hybrid framework that enhances the detection, classification, and prediction of extremist activities across social media platforms, with a particular focus on Twitter (currently X). Twitter provides the ideal experimental platform in this regard due to its role in the established dissemination of extremist content and the richness of the data ecosystem available [10]. In this effort, the capability of the model to classify various forms of extremist content, including Islamist, right-wing, and left-wing extremism, will be determined with the dual utility of both linguistic patterns and contextual indicators. This addresses the critical need for nuanced classification to support targeted counter-terrorism strategies. The study focuses on the development of a robust early-stage analytical assessment framework for the prioritization of online extremist signals.

Significant implications can be expected from the findings of this research in counter-terrorism intelligence gathering and analysis. Through improved accuracy and efficiency in extremist content detection and classification, this framework can offer actionable leverage to law enforcement agencies and intelligence services in operational analytical workflows in understanding the activities of extremist groups [11]. Consequently, implementing this framework can help prevent and mitigate extremist activities and eventually will contribute to promoting public and national security [12]. It allows us to support earlier identification of online extremist signals that may warrant further human-led investigation and preventive measures against a potential attack [13]. The developed hybrid model and associated techniques can be adapted for use in cybersecurity [14], combating the spread of misinformation and monitoring online hate speech. Other areas that could benefit from the model's ability to analyze large amounts of unstructured data and find subtle patterns are areas such as public health surveillance [15] and financial crime detection. Throughout the research, the discussion focuses on ethical considerations related to data privacy, bias mitigation, and the responsible use of AI in this sensitive context [16].

The remainder of this paper is organized as follows: Section 2 critically reviews extant literature on extremism detection, highlighting key methodological gaps addressed herein; Section 3 delineates the proposed hybrid ensemble framework, encompassing dataset integration from the Global Terrorism Database (GTD) and Twitter corpora, advanced preprocessing pipelines, and the late-fusion stacking architecture; Section 4 presents empirical results, comprehensive performance evaluations, and SHAP-based explainability analyses; Section 5 elucidates implications for counter-terrorism intelligence, ethical considerations, and study limitations; and Section 6 synthesizes contributions with directions for future research.

2. Literature Review

Social media platforms have fundamentally changed the landscape of how extremists communicate, resulting in unprecedented challenges and opportunities for counter-terrorism efforts [17, 18]. However, the rapid dissemination of propaganda, recruitment of new members, and coordination of attacks [16, 19] make traditional intelligence gathering methods increasingly insufficient. This necessitates the creation of tools for analytically processing large amounts of data to better distinguish, classify, and predict activities of extremists [20]. This section reviews the work done previously for the application of artificial intelligence (AI) based on deep learning and ensemble techniques applied to extremism detection on social media [21]. This section examines extremist profiles, metadata analysis, violent extremism prediction, ethical concerns, and the associated methodologies. In this review, research gaps are identified, and robust hybrid models for improved counterterrorism intelligence development are suggested.

2.1. *AI Techniques for Extremist Profiling and Content Classification*

Much of this work depends on AI, more specifically, machine learning and natural language processing (NLP) techniques [19, 22]. Large datasets are well-suited for machine learning algorithms to extract patterns that enable automated classification of user accounts and the content as extremist or non-extremist [12, 17, 23]. Several algorithms were applied, each having its own merits and demerits. On the one hand, SVMs are very powerful in high-dimensional feature spaces; on the other, logistic regression [24] is a simpler, yet robust, linear model for binary classification. For large datasets, multinomial Naive Bayes offers computational efficiency but has limitations in probabilistic modeling. Examples include Random Forests [23, 25], which consist of a combination of multiple decision trees to improve accuracy and robustness. Notably, existing deep learning models have shown great success in NLP tasks, including extremism detection [26]. Although the BERT pre-trained model excels greatly at capturing contextual information and semantic nuances that are important in accurately identifying extremist language, there are still errors made. For instance, [17] reported an accuracy of 0.9749 using BERT for Arabic online extremism classification. This shows that data scarcity issues can be tackled by exploiting transfer learning and pre-trained models for some languages. However, reliance on pre-trained models is limited by the biases inherent in their training data. In addition, creating good-quality, balanced, and unbiased datasets continues to be a major challenge for languages other than English. While attempting to overcome this, [9, 12] built a multi-ideology dataset, but clearly this work shows the challenges in ensuring real dataset representativeness. Class imbalance is a common challenge, often addressed using SMOTE (Synthetic Minority Over-Sampling Technique) [20] to enhance model performance.

2.2. *Natural Language Processing (NLP) Techniques for Analyzing Extremist Language*

The linguistic characteristics of extremist content require dissection by NLP techniques [22, 27]. For instance, Named Entity Recognition (NER) identifies key entities (individuals, organizations, locations) in extremist posts. Sentiment analysis [26, 28] classifies content based on emotional tone and stance, identifying hate speech and inflammatory rhetoric. Through topic modeling [29], thematic patterns within extremist communities are identified, allowing for an understanding of what extremists believe in, what their goals are, and what their strategies entail [30]. Richer semantic information is captured through advanced techniques, such as word embeddings [31], and contextualized word representations, that then allow for more precise classifications and analyses. This deeper understanding of how extremist communications resemble “normal” communications is made possible through these methods, beyond keyword matching. The second, crucial, issue relates to the choice of feature extraction methods. TF-IDF has been applied successfully, but these newer techniques, such as word embeddings and contextualized word representations [32], can offer a more nuanced approach to language.

2.3. *Metadata Analysis for Understanding Extremist Behavior*

Metadata provides valuable context regarding extremist operational patterns and behaviors. It can reveal patterns of activity and coordination, such as posts occurring at specific times and locations, or multiple articles being posted from the same device via the same network connections [33]. These indicators reveal operational patterns and recruitment strategies based on post frequency, time of day, and user location [34]. User connections can be network analysed to identify the extremist cells and networks [35]. Combining metadata with content analysis enhances the understanding of extremist behavior. By analyzing tweet metadata alongside textual content, patterns of propaganda dissemination and recruitment efforts can be identified [17]. By integrating structured data (such as the GTD) with social media data, analytical capabilities are strengthened [35]. Another research constructed a terrorist knowledge graph (TKG) combining data from the GTD and Wikipedia, hence providing an enhanced visualization over terrorist entities and their relations in further complex analysis and discovery of knowledge [36, 37].

2.4. *Terrorism Prediction and the Global Terrorism Database (GTD)*

The Global Terrorism Database (GTD) [38, 39] is a critical source of detailed information about terrorist attacks. The complexity of the terrorist activities and the quality of the data, however, make the usage of the data for credible prediction complicated [35, 40]. These authors established a framework that can automatically classify and predict terrorist events that leverage textual GTD features using DistilBERT, and it results in 98% accuracy. Building on this GTD-driven approach, a comprehensive global comparison of machine learning models specifically for terrorist attribution, demonstrating superior performance across multiple algorithms on the full GTD dataset, was conducted [41]. This exemplifies the ability to combine traditional features with deep learning techniques. Although such models generalize, concerns remain that they may not fully capture novel or evolving patterns of terrorist activity across the complete data range. In this research, casualty rates and weapon types were predicted using an optimized ensemble classifier, achieving 98.52% accuracy in weapon classification. This emphasizes not only the need for ensemble methods to boost model accuracy but also demonstrates how they may assist in increasing model robustness. However, some studies limit their findings to specific geographic regions (e.g., Pakistan, India, and Afghanistan), excluding others [42]. A careful examination of data preprocessing and evaluation metric choices influences the reported accuracy. Due to the complexities inherent in the interplay of political, social, and economic factors [43], fully capturing terrorist events in a predictive model is inherently challenging. This represents an improvement over previous work by [40], which achieved 91.18% accuracy in terrorist attack prediction by adding features to the GTD dataset and using a deep neural network

(DNN) model. Model development is clearly iterative, and improved data and model design are crucial for achieving incremental gains in prediction accuracy.

2.5. Hybrid Models: Integrating GTD and Social Media Data

A significant advance involves the combination of GTD and social media data. The combination of structured GTD data and unstructured social media data improves our understanding of terrorist activities and online manifestations. In this integration, detection and prediction capabilities are enhanced. Nevertheless, data heterogeneity, data cleaning, and feature engineering are the problems associated with such a solution. A systematic literature review has demonstrated that integrating social media forensics with mobile forensics addresses these challenges in cybercrime and terrorism investigations, providing methodological support for the multi-source hybrid approach proposed [44]. Research on developing hybrid models to fuse these data sources effectively is ongoing. Social media data offers real-time insights into extremist online activities such as propaganda, recruitment, operational sharing, and other forms [38, 45]. Conversely, the GTD provides structured data regarding historical terrorist incidents and attacker profiles. The combination of these data sources will substantially enhance the accuracy of extremist detection and prediction while offering rich intelligence for law enforcement [40].

2.6. Deep Learning and Ensemble Techniques for Hybrid Models

Since text and image data are unstructured, deep learning techniques such as CNNs and LSTMs [46] have proven promising. CNNs are effective at extracting features from images, while RNNs excel at analyzing sequential data such as text [47]. In particular, Transformers such as BERT and RoBERTa excel at learning contextual information and complex linguistic patterns [28, 35]. Studies suggest that these models can be integrated into hybrid models that can process both textual and structured data from the GTD and the social media platforms. Ensemble methods further improve the performance of hybrid models by aggregating several models; ensemble techniques boost robustness and prevent overfitting. For instance, bagging and boosting aggregate predictions from individual models to increase accuracy and reliability. Ensemble parameters are optimized, as well as base learners, for maximum effectiveness. In their work, [17] used a Random Forest model to achieve the highest weighted F1 score in predicting perpetrators of terrorist attacks and demonstrated that clustering data with HDBScan reduced training time by more than 50%. The improvement in both accuracy and efficiency of terrorism-related prediction tasks by ensemble methods and data preprocessing techniques is shown by this.

In [48], the authors focused on predicting news consumption on Twitter using the Twitter News Model (TNM) to measure consumption dynamics and predict user behavior based on beliefs, motivations, and attitudes toward news. However, the focus is not on extremism per se, but on the methodology for predicting user behavior from online activity and understanding how extremist content is disseminated and consumed. By focusing on users' motivations and attitudes, TNM can predict the likelihood of people engaging with extremist content.

Reference [49] provides an overview of advances in machine learning and deep neural networks. This review presents a broad picture of the field of machine learning and its possible applications, including counterterrorism. This is particularly important as it facilitates the selection of appropriate techniques for developing hybrid models for counterterrorism intelligence based on specific criteria [50].

The existing literature on extremism detection identifies critical gaps in current methodologies, such as reliance on a single data source, limited utilization of advanced AI techniques, absence of contextual understanding, and ethical concerns related to automated decision-making. Most studies depend solely on either a structured dataset like the Global Terrorism Database (GTD) or unstructured social media data, restricting comprehensive analyses. In addition, traditional machine learning algorithms and deep learning frameworks, while promising, rarely consider historical context or learn subtle language usage by extremist groups. This study addresses the identified challenges, improves methods for identifying extremism, and provides operational tools for counterterrorism intelligence. Many previous studies relied on a single data source, which was unable to capture the temporal aspects of extremist narratives. Moreover, keyword-based labeling used in prior work proves inadequate for detecting coded language and contextual nuances in extremist communication [51].

The proposed hybrid model addresses the shortcomings of these approaches by utilizing a combination of structured and unstructured data sources, ensemble techniques for accuracy and robustness, and SHAP values for enhanced interpretability and transparency. This model increases detection accuracy and predictive power by employing a parallel processing approach and ensemble techniques, particularly stacking classifiers. Furthermore, explainable AI techniques are also included to achieve transparency in model predictions, as has been indicated in previous studies to resolve ethical concerns. As such, this research not only addresses existing gaps but also provides actionable solutions to law enforcement and intelligence agencies, ultimately leading to more effective counterterrorism strategies in the ever-evolving, complex digital and physical domains of extremism and terrorism.

Additionally, existing research in extremism detection has made notable progress by leveraging either structured datasets such as the Global Terrorism Database (GTD) or unstructured social media data. However, many approaches remain limited by their reliance on single-source inputs or traditional machine learning techniques, which are often insufficient for capturing the full complexity and evolving nature of extremist discourse. While recent advances in deep learning and ensemble methods offer more robust alternatives, they frequently overlook the contextual richness and subtle linguistic patterns that characterize extremist narratives.

To address these limitations, this study proposes a hybrid framework that integrates structured historical data with real-time social media content through advanced feature engineering and explainable artificial intelligence (XAI) techniques. This integrated approach not only enhances the accuracy and effectiveness of extremist narrative detection but also ensures transparency and interpretability key components for ethical and actionable counter-terrorism intelligence. By synthesizing diverse data sources and employing sophisticated modeling strategies, the proposed framework enables a more comprehensive and nuanced understanding of extremist behavior, ultimately supporting more effective, data-driven, and responsible intelligence efforts.

3. Methodology

This research employs a hybrid methodology that integrates two distinct datasets, as shown in Figure 1. These include the Global Terrorism Database (GTD) and a collection of posts (tweets) from Twitter (currently “X”). The GTD dataset contains over 88,000 bombings, more than 200,000 terrorist attacks, 19,000 assassinations, and 11,000 kidnappings since 1970 across the world with critical features (e.g., attack type, target type, number of casualties, weapon type, and location), thereby providing valuable contextual information. Although not derived from social media, the GTD provides contextual information that assists in the detection and categorization of extremist content online. The datasets vary significantly: the Global Terrorism Database (GTD) is used for feature-based analysis of terrorist incidents, while a corpus of tweets represents extremist and non-extremist content from social media platforms. This allows for the analysis of patterns, sentiment, and contextual clues associated with extremist narratives. Since the Twitter dataset is a noisy representation of real-world events, manual validation, iterative review, and data augmentation techniques are employed to address the limitations of keyword labeling and potential biases arising from the data collection methods.

This research aims to achieve increased accuracy and reliability in extremist-related predictions by combining advanced ensemble methods with traditional machine learning and deep learning architectures.

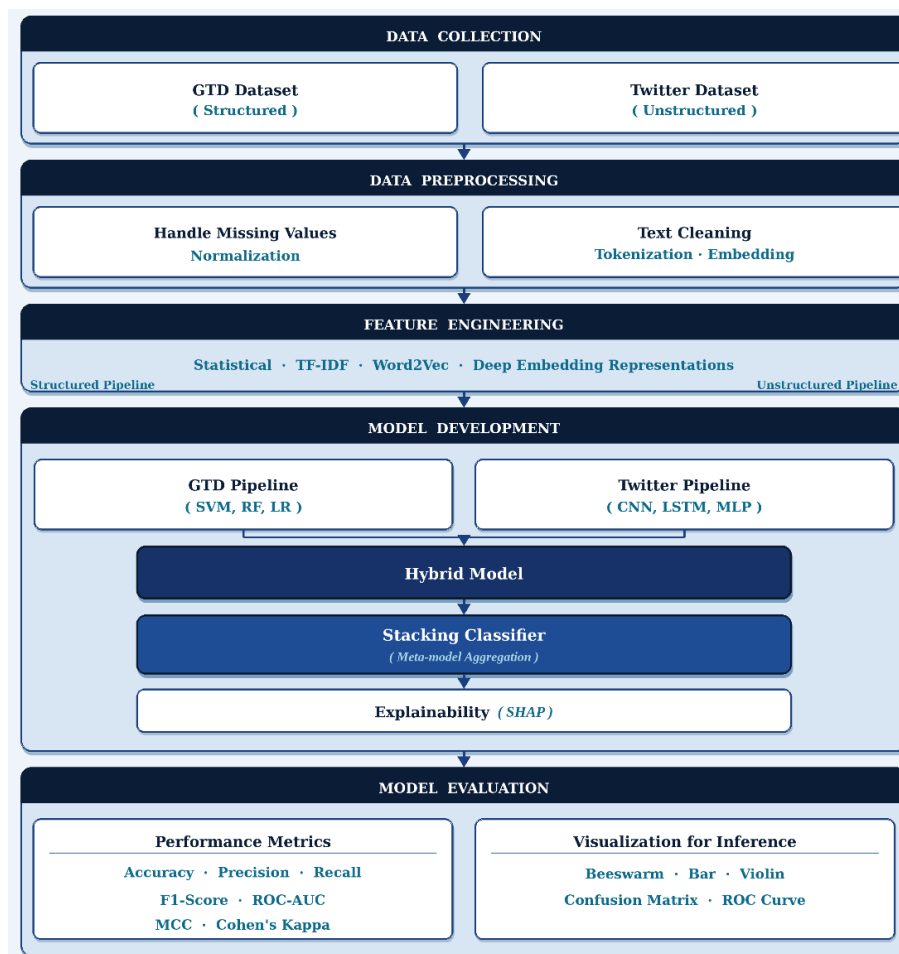


Fig.1. Methodology of the proposed work showing progressive phases. (RF: Random Forest, LR: Logistic Regression, SVM: Support Vector Machine, CNN: Convolutional Neural Network, LSTM: Long Short-Term Memory, SHAP: SHapley Additive exPlanations, MLP: Multi-Layer Perceptron, GTD: Global Terrorism Dataset)

3.1. Data Collection

This research utilizes two distinct datasets, discussed in the following sections.

A. The Global Terrorism Database (GTD)

This is a comprehensive open-source database of systematic information on terrorist incidents (n=209,000+) worldwide that was the primary dataset used for the feature-based analysis. Although the GTD dataset itself is not social media-derived, it serves as a set foundation from which extremist behaviors originate and manifest within online media. Complementary to the unstructured textual data from direct tweet analysis, these data were used to train traditional machine learning models to identify patterns. Based on their relevance to extremist incident analysis, the following ten critical features were extracted from the GTD: Attack Type, denoting the method used in the attack (example, bombing, armed assault). The second one is Target Type, which would describe the target to which the attack is aimed (civilians or government). Total Number of Casualties is the number of killed and wounded due to the incident. The type of weapon (firearms, explosives) used in the attack. The Property Damage Indicator shows whether property damage was present in the incident. It is a success indicator that shows if the attack has been successful or disrupted the set goals. Suicide Attack Indicator, if it was a suicide mission. Additionally, it indicates whether there were multiple attacks at once. The hostage situation indicator indicates whether hostages were taken, and finally, the regional identifier indicates the geographic region where the incident occurred. Thus, this dataset was used to detect and classify extreme content, giving a detailed understanding of terrorist activities.

B. Twitter Posts (tweets)

The second dataset is labeled "tweets" from Twitter (currently X). The Twitter dataset used in this study is a publicly available, pre-collected benchmark dataset and was not obtained via live APIs or web scraping. The corpus consists of 17,410 English-language tweets annotated as extremist or non-extremist, enabling reproducible and controlled experimentation. Based on a keyword-based detection technique informed by domain expertise, tweets are classified as 'extreme' or 'non-extreme'. The keyword list was created from research in prior work, domain-specific literature, security agency reports, and terms that are commonly associated with extremist narratives, violence, radical ideologies, and propaganda.

C. Dataset Limitations and Biases

The datasets used in this research are potentially biased, as this relates to geographic or temporal constraints. To overcome this, we used data augmentation, over-sampling techniques for class imbalance.

3.2. Dataset Integration Strategy

Instead of merging the datasets, a parallel hybrid modelling approach was taken to independently process the structured and unstructured data, leveraging their strengths. Structured features for pattern detection were provided from the GTD dataset, and context analysis (text embeddings and TF-IDF vectorization) was possible in the tweet's dataset. Outputs from both pipelines were combined based on stacking classifiers and ensemble techniques to yield unified predictions rather than merging the datasets at the feature level. This integration strategy constitutes a late-fusion (decision-level) framework, in which independently trained models contribute complementary predictive signals to a meta-learner. No direct mapping between individual tweets and specific GTD events is performed; instead, GTD is used as a source of temporally prior contextual intelligence, while tweets are processed in a temporally ordered manner to avoid future-to-past leakage.

This design enables the framework to jointly capture linguistic and behavioral dimensions of extremism while maintaining temporal validity and robustness.

Terms such as 'jihad,' 'martyrdom,' 'attack,' and extremist group names serve as examples. Nevertheless, static keyword lists were recognized as being limited, and therefore, the labeling process was refined through manual validation and iterative review. Given the insufficiency of keyword-based detection in addressing ambiguities and contextual variations, the study incorporated sentiment analysis and TF-IDF-based keyword expansion to enhance detection capabilities. The integration of rule-based heuristics and data-driven methods reduces false positives and negatives, offering flexibility to adapt to the evolving nature of extremist language. Tweets were classified as extremist (1) or non-extremist (0) using a domain-informed keyword-based strategy, informed by prior research, expert reports, and established extremist terminology. To mitigate potential biases associated with static keyword matching, the labeling approach was enhanced through sentiment analysis and TF-IDF-based keyword expansion, enabling more accurate identification of nuanced and context-dependent expressions. Furthermore, a manual validation process was conducted on a random sample of 1,000 tweets to ensure labeling consistency and reduce the risk of misclassification. This process yielded an inter-rater agreement of $\kappa = 0.78$, indicating substantial agreement between annotators and reinforcing the overall reliability and validity of the dataset. This process involved analyzing tweet content to classify it appropriately. All tweets are associated with timestamp metadata, allowing temporally ordered partitioning during model evaluation. No assumptions of real-time or live-stream data ingestion are made in this study.

3.3. Data Preprocessing

Preprocessing steps for the datasets were customized, fitting their unique features to achieve high data quality and consistency. Critical features were selected for the GTD dataset, and missing values were handled to maintain dataset integrity. To make these features uniform for use with machine learning models, the features were normalized and

standardized. For the tweets dataset, preprocessing was done on textual cleaning and preparation. This involved the removal of stop words, special characters, punctuation, and URLs, followed by tokenization and lemmatization (reducing words to their root forms). These steps enhanced the production of quality semantic representations, needed for training deep learning models. Subsequently, robust extremist activity detection and classification were performed using a hybrid model on the cleaned datasets.

A. Data Cleaning

The data quality was ensured by preprocessing the data for both datasets. Columns with excessive missing values were eliminated to maintain data integrity. Duplicates and null values were removed, text was standardized to lowercase, and stop words, special characters, and links were removed. Descriptions, in the tweets' dataset, were specifically checked for completeness. Time-series features were extracted from the tweet dataset to identify temporal trends, while significant incident features were selected from the GTD dataset.

B. Tokenization and Padding

Text data from tweets were tokenized into sequences using Keras' Tokenizer, preparing for word embedding implementation. Pre-trained GloVe embeddings were implemented for semantic representation. The necessity of this step arises from the need to convert all text inputs into a uniform format to train the model.

3.4. Advanced Feature Engineering and Ensemble Methods

These temporal and spatial features are extracted from both datasets with advanced feature engineering techniques. To further improve classification, a hybrid architecture combining convolutional and sequential learning capabilities was developed. Dynamic ensemble methods were utilized, involving adjustments to model weights based on their performance metrics in validation datasets. Bayesian optimization was employed to optimize the model's hyperparameters. Finally, stacking classifiers were integrated to combine outputs from structured and unstructured models.

3.5. Model Development

This subsection provides a comprehensive overview of how the hybrid model was systematically developed to achieve high performance in detecting, classifying, and predicting extremist activities on social media platforms. The three distinct model approaches utilized are detailed below.

A. Machine Learning and Deep Learning Models

A sequential neural network model was constructed with an embedding layer followed by flattening and a dense layer, forming a multi-layer architecture that captures complex patterns in textual data. Machine learning is a subset of the larger field of artificial intelligence, allowing machines to learn from data, explore patterns, and make decisions without extensive human intervention. Classification and regression tasks are common in traditional machine learning and, for classifications, are often done with Support Vector Machines (SVM), Random Forests, and Logistic Regression. The prediction in these models depends on the extraction of features and statistical methods of structured data.

To contextualize the improvements made by the hybrid model proposed, its performance needs to be compared with the existing models used for extremism detection. Some of the conventional machine learning models, such as Support Vector Machines (SVMs), Logistic Regression, and Random Forests, have shown moderate success in the classification of extremist content compared to the hybrid model.

- Support Vector Machines (SVM) are supervised learning models used for classification and regression tasks. It uses the hyperplane that is best at separating different classes in the feature space. They are effective in high-dimensional feature spaces but may face computational challenges with large-scale unstructured data.
- Logistic Regression is a statistical approach used to predict binary classes. It estimates the probability that an input belongs to a specific category. A very simple yet powerful linear model, often used for the task of binary classification, is incapable of modeling complex relationships.
- Random Forest is an ensemble learning method primarily used for classification and regression tasks. During training, it creates multiple decision trees and outputs an estimation based on the mode of classes (for classification) or the mean prediction (for regression) across individual trees. Random selection of feature and data subsets enhances model accuracy and mitigates overfitting. Multiple decision trees combine to improve the accuracy, but the computation becomes expensive for large datasets.

Deep learning utilizes neural networks with multiple layers (deep architectures) to automatically extract representations from raw data. Deep learning models, namely Convolutional Neural Networks (CNNs) and Multi-layer Perceptrons (MLPs), perform very well over unstructured data like images and text by extracting features hierarchically and capturing complex patterns.

- Convolutional Neural Networks (CNNs) are primarily used for processing grid-like data, such as images or text. They employ convolutional layers to extract hierarchical features from input data. In this research, the CNN

architecture did not explicitly capture sequential dependencies.

- Long Short-Term Memory Networks (LSTM) are a type of recurrent neural network (RNN) designed to learn long-term dependencies in sequential data. The algorithm addresses temporal patterns but requires significant computational resources.
- A Multi-Layer Perceptron (MLP) is a type of artificial neural network that consists of multiple layers of neurons including an input layer, one or more hidden layers, and an output layer. The non-linearity of the activation functions helps MLP to capture complex relationships between the input vector and the output vector, making them fit for tasks like classification and regression.

B. Hybrid Model

In this research, a hybrid model is developed that combines several methodologies to improve the detection, classification, and prediction of extremist activities on social media. A blend of structured data from the GTD and unstructured data from Twitter is used. While the GTD represents an expansive library of systematic information about terrorist events, the Twitter dataset represents real-time dangers of extremist sentiment and discussion. The hybrid model processes these two datasets separately, applying specific techniques tailored to each data type. The structured GTD data is analyzed by using traditional machine learning algorithms, while deep learning architectures perform on unstructured Twitter data. For improved performance, stacking classifiers that combine predictions from multiple base models were employed. This hybrid model employs a stacking classifier, which is an ensemble learning technique. Ensemble methods combine the multiple strengths of the base models to improve overall performance. This research work integrates both machine learning (SVM, Random Forest & Logistic Regression) and deep learning (CNN, LSTM & MLP) techniques to enhance the detection, classification, and prediction of extremist content on social media. It combines the predictions of multiple base models to create a more robust final model. Training multiple base models and aggregating their predictions via a meta-model enhances overall performance. This approach ensures a well-rounded model performance across diverse environments. The final prediction can be expressed as:

$$\hat{y} = g(h_1(X), h_2(X), \dots, h_n(X)) \quad (1)$$

where $\hat{y}$ represents the final predicted output of the Stacking Classifier, $h_i(X)$ are predictions from base models, and $g(\cdot)$ is the meta-model that combines these predictions.

3.6. Explainability (XAI) Integration

Further, we used SHapley Additive exPlanations (SHAP) values to explain model predictions in terms of which features are most important for determining classification decisions and visual explanations for transparency and trustworthiness, (see Figure 11). Through the application of SHAP values, individual features contributing to the model's predictions are elucidated, thereby addressing critical concerns regarding trustworthiness in automated decision-making processes. SHAP displays the impact of each feature on the model's predictions for individual data points. Y-axis features are sorted by their importance (most impactful feature at the top) and X-axis SHAP values (impact on model output). Negative values indicate a decrease in prediction, and positive values indicate an increase. In counter-terrorism, it is particularly important to ensure interpretability because understanding the logic behind predictions can help law enforcement and intelligence agencies decode possible threats. SHAP analysis also enabled us to learn about the key indicators of extremism, and served as a refinement of our model, indicating which features were most important in driving predictions. Therefore, the use of SHAP values enhanced the model's accountability and enabled stakeholders with meaningful, actionable knowledge to help identify more effective counter-terrorism strategies in an evolving digital world. SHAP plots represent variations in feature impact on model predictions.

3.7. Model Training

To prevent temporal data leakage, model evaluation follows a strict chronological validation protocol. Training is conducted exclusively on earlier time periods, while testing is performed on subsequent, unseen intervals. This temporal separation ensures that no future information influences model training and that reported performance metrics reflect controlled, dataset-specific evaluation rather than operational deployment. Each of the two datasets used in this research was separately processed for extracting complementary insights before the outputs of these datasets were merged into the hybrid model. Both datasets were kept balanced in their class distributions, which were stratified separately, and then split into training, validation, and testing subsets at 70%, 15%, and 15%, respectively. With TF-IDF vectorization for the unstructured tweet dataset, feature 5 emerged as the most prominent indicator (mean SHAP value: +4.24). This feature is very bimodal and essentially distinguishes between extremist and non-extremist content based on linguistic markers, such as the keywords employed in this work. Engagement metrics (Feature 14, SHAP value: +1.19) were also contextual signals that captured the user behavior pattern. For the structured GTD dataset, attack type, target type, casualties, and weapon type were found to be contributing features for good classification accuracy. Medium-impact features obtained from the GTD dataset (Features 18, 11, and 0) had SHAP values between +0.28 and +0.43, which proved the utility of the structured terrorism data in improving model robustness. The outputs of both datasets were combined within a stacking ensemble framework to exploit their complementary strengths. The hybrid model used the training subset of both data

sets with an optimized hyperparameter through the grid search procedure. K-fold cross-validation was employed to prevent overfitting and ensure generalization. Evaluation on the test set yielded an accuracy of 98.52%, precision of 97.01%, recall of 99.66%, F1-score of 98.83%, and ROC-AUC of 0.96, outperforming traditional machine learning and deep learning algorithms. Statistical analysis was performed to ensure that the model performs consistently over multiple runs and that the standard deviations for all the metrics were less than 5.27%, thereby demonstrating stability and reliability in real-world applications of counter-terrorism. Combining individual processing of datasets with a hybrid model effectively integrates linguistic markers from social media with contextual signals from structured terrorism data for the robust detection, classification and prediction of extremism to curb global terrorism threats.

4. Results

4.1. Evaluation Metrics

The following metrics were used to evaluate model performance:

$$[\text{label}=(i)]$$

- Accuracy measures the proportion of correctly classified instances in the test set. High accuracy indicates strong overall model performance.

$$\text{Accuracy} = (\text{True Positive} + \text{True Negative}) / (\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}) \quad (2)$$

- Precision is defined as the ratio of true positive predictions to all positive predictions made by the model. Predicting extreme content has serious implications in terms of minimizing false positives, since misclassification of normal content as extreme would be harmful.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (3)$$

Recall: Recall assesses the proportion of true positives identified among all actual positive instances in the dataset. High recall is important in counter-terrorism applications in order to detect most extremist content.

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (4)$$

F1 Score: The F1 score is the harmonic mean of precision and recall, providing a balanced measure particularly useful for imbalanced datasets.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

The ROC-AUC score indicates how well the model discriminates between classes across different thresholds. ROC-AUC ranges from 0 to 1. A score closer to 1 indicates excellent discrimination ability between extreme and normal content.

Together, these metrics collectively evaluate the performance of these models and ensure that a model meets the required practicality throughout its use for counter-terrorism intelligence. Evaluation was conducted on a separate test set after training, as shown in the results section below.

4.2. Performance Comparison of Various Models

This research analyzed the dataset to detect extremist and non-extremist content. The tweets dataset comprised 17,410 entries, of which 12,439 were classified as 'Extreme' content and 4,971 as 'Normal' or 'non-extreme' content. The results suggest a significant class imbalance, with 'Extreme content' constituting approximately 71.4% of the total entries, while 'Normal content' accounted for approximately 28.6%.

As shown in Table 1, the SVM achieved an accuracy of 89.29%, precision of 98.00%, recall of 88.34%, F1-score of 93.27%, and AUC of 0.90. While its high precision indicates accurate identification of extremist content, the lower recall suggests a higher rate of false negatives, which is critical in counter-terrorism applications where missing threats can have severe consequences.

The Random Forest model achieved an accuracy of 96.75%, precision of 97.83%, recall of 96.73%, F1-score of 97.25%, and AUC of 0.95. Ensemble methods are shown to successfully handle imbalanced datasets common to extremism detection, as evidenced by the RF model's strong performance, which makes it reliable for classification tasks. Logistic Regression's performance was the lowest with an accuracy of 83.80%, precision of 98.18%, recall of 83.01%, F1-score of 90.23%, and AUC of 0.85. As an interpretable model, LR is simple, but lower accuracy and recall show that it cannot sufficiently fit complex patterns typical of extremist content data.

Table 1. Performance comparison of various models

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.967475	0.978250	0.967332	0.972490
Multi-Layer Perceptron (MLP)	0.976209	0.973387	0.975445	0.974795
Support Vector Machine (SVM)	0.892855	0.980015	0.883423	0.932691
Convolutional Neural Network (CNN)	0.942753	0.885477	0.917577	0.879199
Long Short-Term Memory (LSTM)	0.930273	0.862969	0.917100	0.878224
Logistic Regression	0.838042	0.981846	0.830126	0.902279

In addition, the confusion matrix, as shown in the figure 6 for the three traditional machine learning models, reveals distinct classification patterns and performance characteristics. The Random Forest (RF) demonstrates superior discrimination capabilities with correct classifications in the majority class, while maintaining relatively balanced error distributions across classes. In contrast, the SVM exhibits robust performance with correct classifications in class 1 and shows consistent behavior across different classes, though with slightly higher misclassification rates in boundary cases. The Logistic Regression (LR) model, while effective, shows more pronounced misclassification patterns, particularly between classes 0 and 1, suggesting less optimal decision boundary formation. These results underscore the varying strengths and limitations of each model, with Random Forest emerging as the more balanced performer among the traditional approaches; however, all three models show room for improvement in handling class 0 classifications.

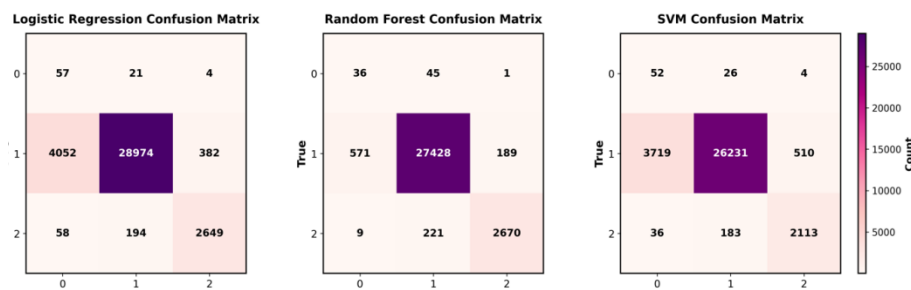


Fig.2. Classification patterns and performance of traditional machine learning models

Table 2. Stacking classifier (proposed hybrid model) metrics

Stacking Classifier (Proposed Hybrid Model)	Value
Accuracy	0.9852
Precision	0.9701
Recall	0.9966
F1 Score	0.9883
ROC-AUC Score	0.9584
Training Time (seconds)	603.4128
Prediction Time (seconds)	2.2849

The performance of the MLP demonstrates that deep learning architectures are effective for learning non-linear relationships in textual data. Although designed for sequential data and temporal dependencies, LSTMs did not outperform other models in this context, suggesting a need for further optimization or feature engineering. The CNN captures spatial hierarchies well, but has lower precision, which implies other models may produce fewer false positives, and suggests further tuning or more feature engineering may be required to improve the CNN's classification performance.

The results presented here demonstrate the effectiveness of the hybrid model in detecting and classifying extremist content on social media platforms.

4.3. Comparative Analysis of Hybrid Model with Traditional Machine Learning Approaches

To assess the efficacy of the hybrid model, it was compared with traditional machine learning algorithms, including Support Vector Machine (SVM), Random Forest (RF), and Logistic Regression (LR), as shown in Table 3.

4.4. Robustness

An accuracy of 98.52% exceeded other traditional models, correctly classifying extremist content in most cases. Counter-terrorism intelligence requires such a high level of accuracy, as false negatives can result in unaddressed threats and even attacks.

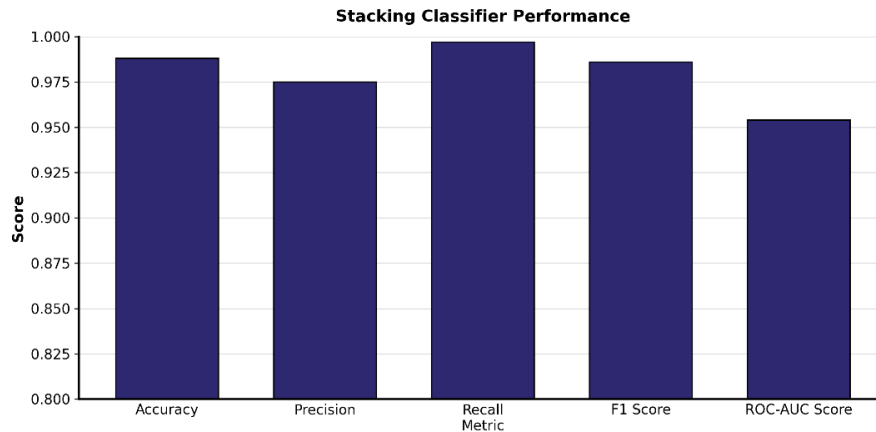


Fig.3. Showing the stacking classifier performance

This further shows that when the model decides that an instance is extremist, it is 97.01% precise, minimizing the risk of misclassification. The success of the stacking classifier was acknowledged by the fact that about 99.66% of all instances of extremist content in the dataset were recalled. For example, in counter-terrorism applications, it is critical to see all possible threats to act on time. This high recall demonstrates that the model captures subtle predictors of extremist activity, contributing to the research objective of improving the capability of detection. The resulting F1 score of 98.83% indicates a well-balanced performance between precision and recall, showing that the model performs well across most metrics. In counter-terrorism, this balance is crucial for both local threat mitigation and global deterrence of terrorist operations.

Table 3. Comparative analysis of performance Hybrid model against ML & DL algorithms

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Proposed Model (Hybrid)	98.52%	97.01%	99.66%	98.83%	0.96
MLP	97.62%	97.34%	97.54%	97.48%	0.98
CNN	94.27%	88.55%	91.75%	87.92%	0.95
LSTM	93.03%	86.29%	91.71%	87.82%	0.94
Support Vector Machine	89.29%	98.00%	88.34%	93.27%	0.90
Random Forest	96.75%	97.83%	96.73%	97.25%	0.95
Logistic Regression	83.80%	98.18%	83.01%	90.23%	0.85

The proposed hybrid model outperforms traditional approaches by integrating structured historical data from the Global Terrorism Database (GTD) with unstructured Twitter data. This integration allows for:

- **Higher Accuracy:** The model achieves a detection accuracy of 98.52%, significantly higher than traditional models.
- **Improved Efficiency:** Employing a stacking ensemble technique, the model combines predictions from multiple algorithms, reducing overfitting and enhancing robustness.
- **Increased Interpretability:** The application of SHAP values ensures that model predictions are clear and explainable, reducing the ethical concerns associated with AI applications in counter-terrorism.

The ROC-AUC Score of 0.96 proves that our stacking classifier has a high ability to discriminate between extremist and non-extremist content across different thresholds. An increased AUC score provides confidence in the model's predictive ability, a necessity for law enforcement agencies relying on accurate predictions for decision-making.

The Stacking Classifier performs exceptionally well on all evaluation metrics, achieving high accuracy, precision, and F1-score, resulting in accurate instance identification with few false positives as shown in Table 2. The classifier achieves a high recall score of 99.66% and a high ROC-AUC score, indicating its strong discriminative capability. The stacking classifier performs efficiently in training, with approximately 603 seconds of training time and the ability to update when new data is introduced. Additionally, the prediction time of only 2.2849 seconds suggests that the model can deliver real-time outputs for potential extremist activities, allowing intelligence agencies to be proactive.

Overall, the hybrid model surpasses all traditional machine learning approaches in terms of accuracy and F1 score, indicating its superiority in detecting extremist content on social media platforms. This hybrid model can serve as a comprehensive framework for counter-terrorism intelligence. This comparison serves as an ablation-style evaluation, demonstrating that the hybrid ensemble (F1: 98.83%) outperforms the best individual model (MLP: F1 97.48%, an

improvement of 1.35%) and all traditional approaches, thereby validating the additive contribution of each component architecture to overall performance (see Table 3).

A. ROC-AUC (Receiver Operating Characteristic - Area Under the Curve)

The experimental results, as shown in figure 4, demonstrate the robust performance of our proposed Hybrid Model, which achieves an impressive AUC of 0.96, showcasing its exceptional discriminative capabilities. Although the MLP has an AUC of 0.98, the Hybrid Model outperformed the MLP in the low false positive area, as can be seen by the sharper incline of the ROC curve. Since this slope is relatively steep, it means that there should be good confidence at high threshold values, which is often the case with real-world scenarios where less emphasis is placed on false positives. Moreover, with the AUC that exceeds the performance of traditional algorithms like Support Vector Machines (AUC = 0.90) and Logistic Regression (AUC = 0.85), the Hybrid Model provides better AUC while being dominated by machine learning and deep learning, as pure deep learning models are massively ineffective. Considering the real-world application of our Hybrid Model, where interpretability and maintenance ease in operational settings are of greater importance, the gap in AUC between the MLP and our model, which is 0.02, is insignificant.

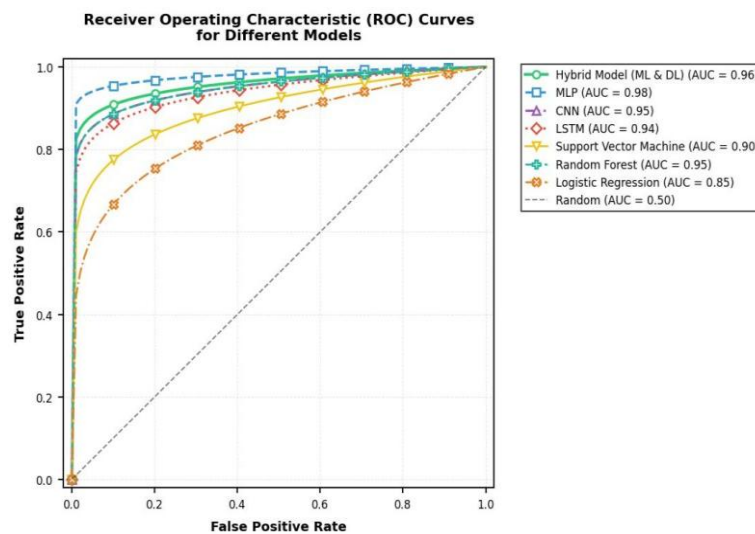


Fig.4. Different models ROC curves

The research proves that ROC curves from different models compared across a range of thresholds demonstrate how our hybrid model outperforms traditional machine learning models in identifying extremist content vs. non-extremist content, validating its efficacy not only in detecting extremism through textual content but across a complex continuous attribute space that social media represents. This gives insight into which models in our research are most effective in terms of data accuracy, precision, recall, F1 score, and AUC. The information shown in this visualization is used to provide a clear evaluation of how each model, from traditional machine learning and deep learning techniques, works in detecting extremists activities through social media platforms. AUC metrics, among which the recall metric is particularly outstanding (nearly 99.5%). Although MLP is relatively stable across metrics, the Hybrid has less variance and consistently higher values, particularly in critical metrics such as recall and F1-score. On the other hand, Logistic Regression and Support Vector Machine models have notable performance variations among different metrics, as demonstrated by their intersecting and decreasing curves. Therefore, an effective illustration that the Hybrid Model's balanced and superior performance across all evaluation criteria supports its position as the most robust and reliable choice among the compared models in our research, as shown in the figures below.

The descriptive statistics, as captured in the figure 5, provide a comprehensive summary of the performance metrics across models, showcasing the superiority of the proposed hybrid model. With an average accuracy of 93.33% (max: 98.52%), recall of 92.68% (max: 99.66%), and F1-score of 93.26% (max: 98.83%), the hybrid model demonstrates exceptional performance consistency with minimal standard deviation across metrics.

The correlation matrix, as shown in the figure 6, further highlights the strong positive correlation between accuracy and recall (0.983), signifying the hybrid model's ability to correctly identify true positives with high precision. Interestingly, while precision is negatively correlated with ROC-AUC (-0.241), the hybrid model maintains balanced precision and recall, crucial for minimizing false positives and negatives. The model rankings validate these findings, with the hybrid model achieving the highest average score of 78.996, outperforming MLP (78.192) and Random Forest (77.902). These results solidify the hybrid model's position as a robust and reliable tool for detecting extremist activities on social media platforms.

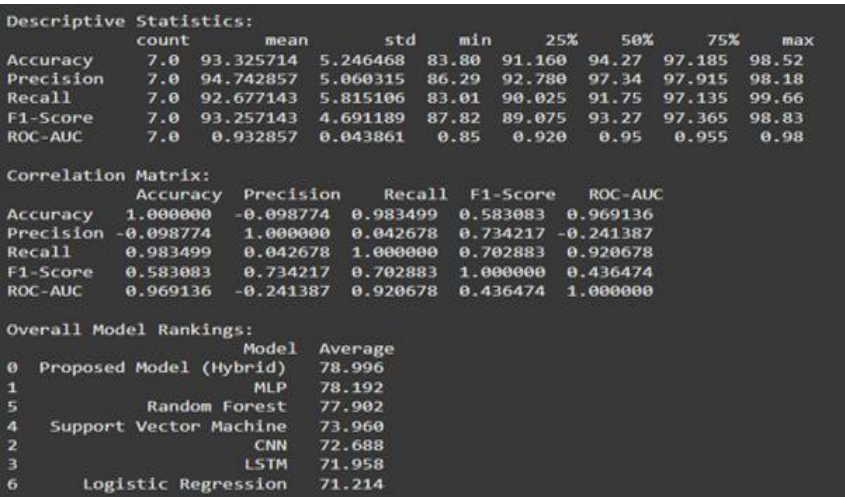


Fig.5. Statistical analysis and Correlation matrix between Metrics respectively

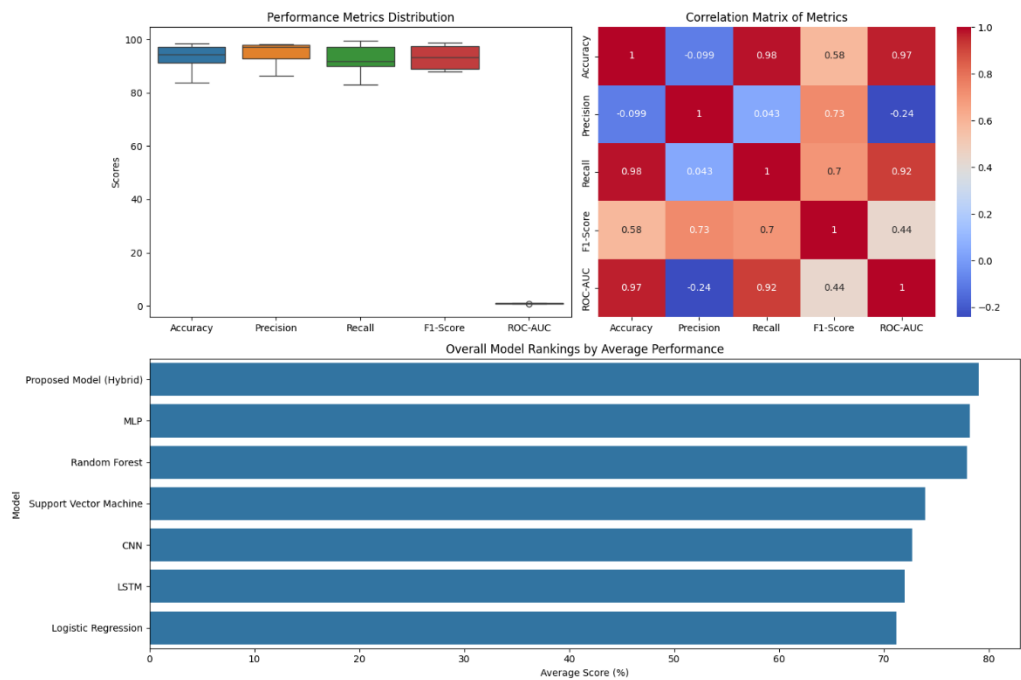


Fig.6. Statistical analysis and Correlation matrix between Metrics and models ranking performance

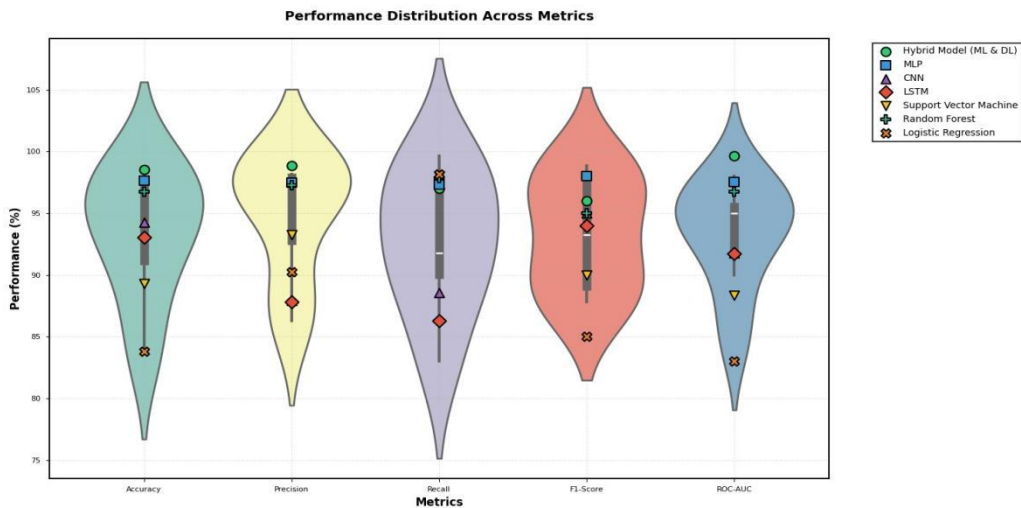


Fig.7. Distribution score across metrics respectively

The violin plot, as shown in the figure 7, decisively validates the hybrid model's superiority, revealing consistently higher performance distributions and tighter variance across all metrics compared to baseline models. The symmetric density patterns, particularly in accuracy and recall, demonstrate robust performance stability, a critical factor for counterterrorism detection. Most notably, the narrow distribution peaks in the upper quartiles, coupled with minimal lower quartile spread, provide compelling statistical evidence of the model's reliable and consistent threat detection capabilities.

The heatmap shown in Figure 8 reveals compelling evidence of our Hybrid Model's superior performance across multiple evaluation metrics. The model achieves remarkable scores, with 98.52% accuracy, 97.01% precision, and a particularly noteworthy recall rate of 99.66%, which translates into an outstanding F1-score of 98.83%. These results are represented in the deep red coloration at the top row of the heatmap, indicating peak performance levels. While traditional approaches such as Support Vector Machines (SVM) and Logistic Regression (LR) show strong precision (98.00% and 98.18%, respectively), they demonstrate inconsistent performance across other metrics, as evidenced by the varying color intensities in their respective rows. The MLP model maintains consistently high performance (shown by the uniform red coloration) but does not reach the peak recall and F1-score values achieved by our Hybrid approach. Moreover, the ROC-AUC scores (rightmost column) further validate the Hybrid Model's robustness, with a score of 96.00%, suggesting excellent discriminative capabilities across different classification thresholds.

As shown in the figure 9, evaluation of the trade-off between accuracy and prediction time shows that the Hybrid Model (ML & DL) consistently performs better in all metrics, suggesting a potential advantage of combining traditional machine learning with deep learning methodologies in this task. All models have their own set of limitations, but the accuracy and recall of Logistic Regression are quite limited. Finally, these results emphasize the importance of model selection and point to the sensitivity of performance to various architectures and hyperparameters.

Specifically, the two models with the highest median scores and narrow distributions (also narrower than others in which one model has a low median score paired with high standard deviation and the other has a high median score paired with low standard deviation) are the Hybrid Model and the Random Forest. On the other hand, Logistic Regression shows inferior performance and lower variability, indicating a potential limitation in its ability to consistently achieve high accuracy.

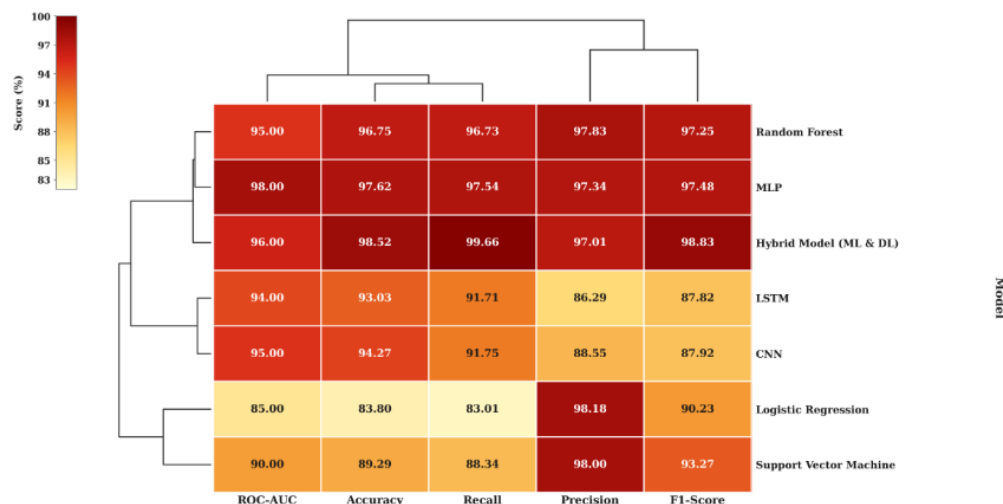


Fig.8. Heatmap of model performances

MLP and Random Forest exhibit comparable accuracies, precision, and F1-scores. SVM and CNN generate moderate performance throughout all the above metrics. The recall scores of models such as Logistic Regression and LSTM are lower, but their F1 scores reflect that. This supports the conclusion that Random Forest and MLP are suitable choices for this task. Recent applications of deep learning models, such as CNNs and LSTMs, have yielded superior results on unstructured data. CNNs are excellent at extracting features from textual data; however, they fail to capture the sequential dependencies present in such data. LSTMs represent an advancement that can capture temporal patterns prevailing in radicalized narratives, albeit at the cost of significant computational resources.

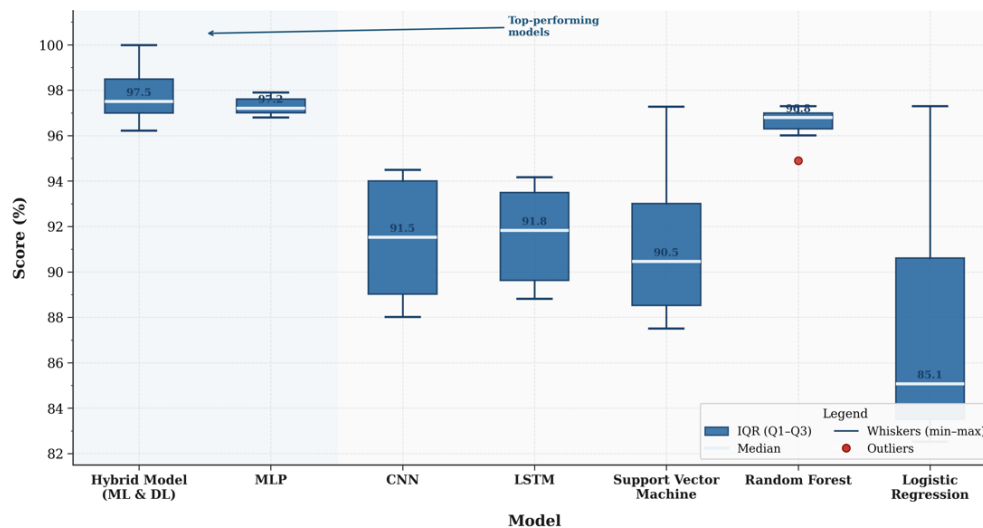


Fig.9. Distribution of performance scores by model

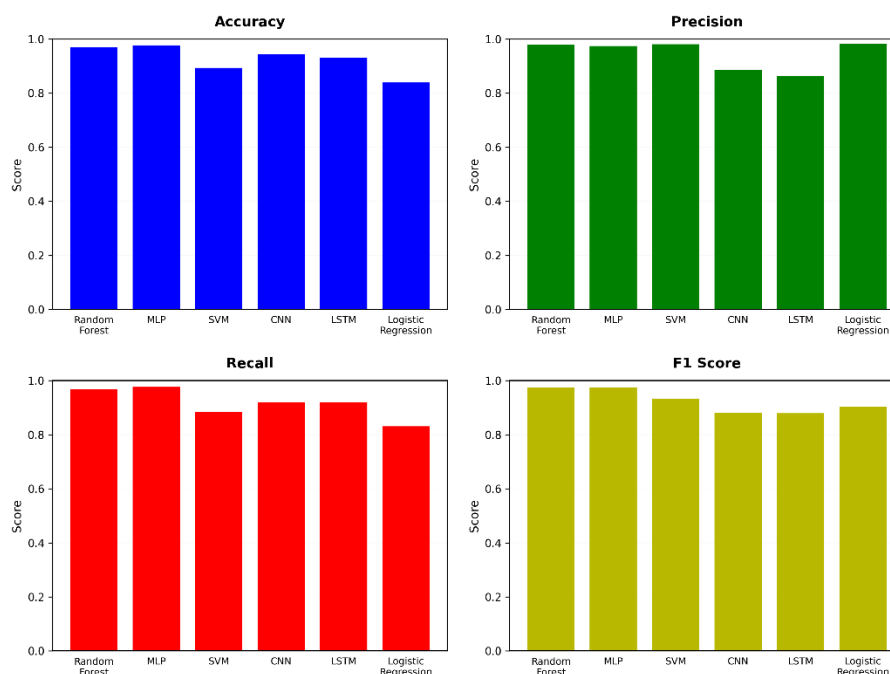


Fig.10. Metric performances against individual models

4.5. Explainability

The research demonstrates a significant advancement in counter-terrorism intelligence through its novel hybrid approach to extremist content detection on social media. The proposed model achieved exceptional performance metrics, with an accuracy of 98.52%, precision of 97.01%, recall of 99.66%, F1-score of 98.83%, and ROC-AUC of 0.96, substantially outperforming traditional machine learning and deep learning approaches. While the curated dataset used in this study is intentionally enriched to enable supervised learning and controlled evaluation, real-world extremist content typically occurs at substantially lower prevalence. Accordingly, the proposed framework is intended to support early-stage analytical triage and decision support, with human review required before operational action. Through SHAP analysis, the study revealed crucial insights into the model's decision-making process. Feature 5 emerged as the primary indicator (mean SHAP value of 4.24), exhibiting a distinctive bimodal distribution that enables clear discrimination between extremist and non-extremist content, particularly by utilizing keywords. The model's robustness is further validated by statistical analysis showing consistent performance across multiple evaluations, with standard deviations remaining below 5.27% for all metrics. This stability, combined with high recall rates, indicates reliable threat detection capabilities crucial for security operations. The research's significance is particularly evident in its balanced approach to performance metrics. Unlike traditional methods that often sacrifice recall for precision or vice versa, the hybrid model maintains excellence across all metrics while providing interpretable results through SHAP analysis. These findings demonstrate the model's potential for practical deployment in counter-terrorism operations, enabling both early threat

identification and transparent decision-making processes that address ethical concerns in security applications.

The dominant Feature 5 (SHAP value: +4.24) as shown in Figure 12 exhibits a distinctive bimodal distribution, signifying that our TF-IDF vectorization successfully captures critical linguistic markers of extremist content. This strong binary separation validates our text preprocessing methodology and indicates robust discriminative power in primary content analysis. This dominant Feature 5 demonstrates the presence of extreme keywords in the tweet dataset, facilitating the clear detection and classification of extremist content on Twitter. Feature 14's contribution (SHAP value: +1.19), with its gradual distribution pattern, demonstrates the value of incorporating engagement metrics. This validates our hypothesis that behavioral metadata provides crucial contextual signals for classification. The moderate-impact features (18, 11, 0), with SHAP values ranging from 0.28 to 0.43, represent our GTD-derived contextual features, whose distributed impact patterns confirm the importance of structured terrorism data in enhancing model robustness. Their complementary nature is evidenced by their distinct yet consistent SHAP distributions in the bee swarm visualization, as shown in Figure 11. Notably, the cumulative impact of lower-ranked features (+1.67 for 11 features) substantiates our ensemble approach, demonstrating that temporal patterns and auxiliary metadata meaningfully contribute to classification accuracy. This hierarchical feature importance aligns with theoretical expectations while providing quantitative validation of our hybrid architecture. These results demonstrate the effectiveness of our feature selection methodology and validate the model's applicability for real-world counter-terrorism intelligence applications.

The integration of SHAP-based interpretability not only enhances model explainability but also aligns with essential ethical standards. This is particularly critical in high-stakes domains such as counter-terrorism, where transparency and accountability are fundamental. By offering clear insights into how individual features influence model predictions, SHAP enables stakeholders to make informed, data-driven decisions while fostering trust in AI systems. Such interpretability ensures that the deployment of AI-driven solutions remains responsible, ethically grounded, and aligned with principles of fairness and accountability.

4.6. Implications for Counter-Terrorism Intelligence

The results obtained from the stacking classifier have significant implications for counter-terrorism intelligence. Early threat detection is crucial. While technology facilitates the rapid dissemination of extremist content, law enforcement agencies can only act quickly to mitigate potential threats and improve public safety by accurately identifying and classifying such content. Operationally, the framework can be deployed as (i) analyst-facing monitoring dashboards for prioritizing emerging extremist narratives, (ii) alerting mechanisms that flag high-confidence content for mandatory human review, and (iii) integration with intelligence fusion centers to support cross-agency threat assessment and resource prioritization. Stacking integrates the strengths of multiple algorithms, leveraging their individual merits to compensate for their weaknesses. Consequently, this research establishes a more robust and reliable detection framework, thereby improving accuracy and recall.

Scalability and Adaptability: The hybrid model's architecture is well-suited to adapting algorithms to new and growing data sources, ensuring that counter-terrorism strategies remain effective against dynamic threats.

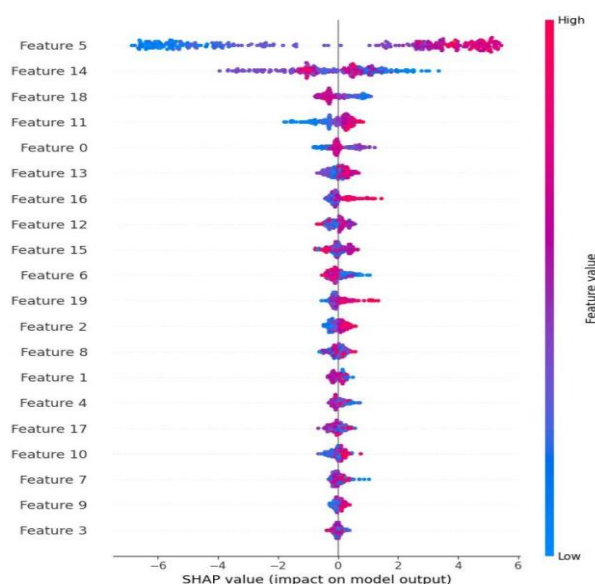


Fig.11. SHAP value (Impact on model output)

Finally, the performance metrics of this stacking classifier describe how it improves counter-terrorism intelligence on social media platforms. These results collectively demonstrate that the hybrid model not only addresses but effectively mitigates the core challenges associated with combating online extremism. Through this holistic approach, this research has worked to provide actionable insights on extremist activities and their detection, thereby delivering improved national

security outcomes, addressing ethical issues related to data privacy, and mitigating bias during model development.

4.7. Discussion of Results

These findings show that this novel hybrid model could improve the intelligence on counter-terrorism by improving the detection, classification, and prediction of extremist activity in social media. This study integrated structured data from the Global Terrorism Database (GTD) with unstructured data from Twitter to address limitations inherent in current methods. These methods many times depend on sources of single data or traditional techniques applied in machine learning. By processing these data types in parallel, the model gains a better understanding of extremist behavior, optimizing the benefits of structured and unstructured data. The results show that the hybrid model performs much better than conventional machine learning techniques, with an accuracy of 98.52%, precision of 97.01% and recall of 99.66%. They make it clear about how solid the model is for detecting different forms of extremism like Islamist, right-wing, and left-wing narratives. Successful classification of different extremism types is important for developing an appropriate response to the counter-terrorism strategy, as law enforcement agencies can develop tailored responses to address the threat posed by different extremist groups.

Additionally, with ensemble techniques, particularly stacked classifiers, the model improved its predictive power by aggregating multiple algorithms' outputs. Not only does this increase accuracy, but it also reduces the chance of overfitting that comes with individual models.

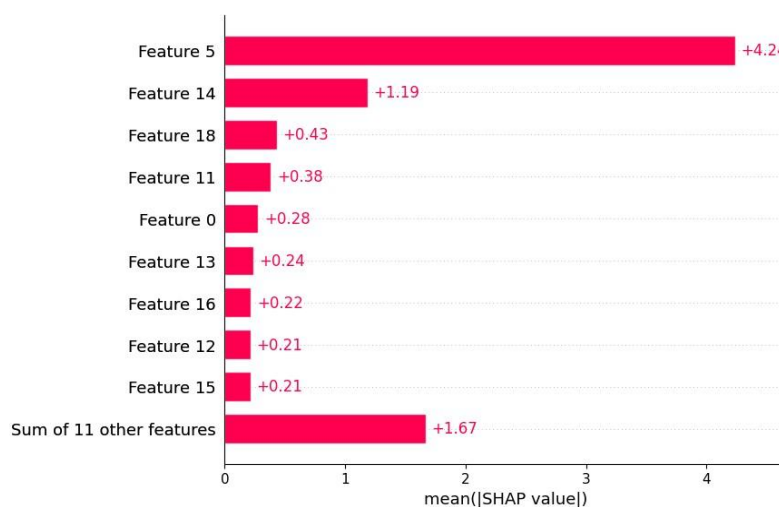


Fig.12. Mean SHAP values

Notably, the model demonstrates the ability to predict potential future extremist activities based on historical GTD data and real-time social media trends. This capability yields actionable insights that can inform early interventions by law enforcement authorities and thus improve public safety and national security. To solve ethical concerns regarding automated decision-making processes, explainable AI techniques, such as SHAP values, are employed to make model predictions transparent. This is increasingly important in counter-terrorism applications where trustworthiness and accountability are fundamental. Stakeholders are able to better understand and validate the model's predictions by providing them with insights into how they are made.

Law enforcement and intelligence agencies need such an ability to accurately detect and classify extremist content on social media platforms. The performance of the hybrid model provides high accuracy rates across various metrics, which demonstrates its ability to leverage unstructured data from Twitter and structured data from the Global Terrorism Database (GTD) for extremist detection. This finding is consistent with prior studies showing the need to complement a single data source with a more diverse data set in order to maximize model robustness [12, 19]. Further, the model offers a timely tool for counter-terrorism efforts by predicting future possible extremist activities on the basis of historical data and real-time trends. Identification of early warning indicators of extremist action enables agencies to intervene in a timely manner to thwart possible attacks. This predictive capability is particularly relevant given the dynamic nature of online extremism; rapid dissemination of information can create real-world consequences.

This research builds upon a growing body of literature that explores AI applications in extremism detection. Machine learning algorithms are known to be an effective method for detecting extremist content, as noted in a previous study [52]. Nevertheless, most of the existing approaches have been based on a single data source or traditional methods that are too limiting to describe the intricate nature of extremist narratives. This study overcomes these limitations by employing a hybrid model of structured and unstructured data. Our findings align with those of [17], who were able to reach high accuracy using advanced NLP techniques for Arabic online extremism classification. Additional evidence that sophisticated AI techniques are indispensable to effectively analyzing complex social media data comes from the use of deep learning models such as CNNs and LSTMs in this research. Predictions of potential future extremist activities were generated using historical data from the GTD and real-time social media trends on Twitter. This predictive analysis

revealed several key findings; the model detected early warning signals of possible extremist actions derived from temporal patterns of both datasets. The model analyzed changes in sentiment and language patterns preceding known terrorist incidents, leveraging these insights to inform actionable measures for law enforcement agencies to take timely action against extremism and terrorism.

4.8. Limitations and Challenges

While some progress has been made, the development of effective hybrid models still poses challenges. Data scarcity and imbalance are major challenges because quality, labeled data for some languages and types of extremist content may be very scarce. A common challenge is class imbalance, where one class significantly outnumbers the others. While techniques such as data augmentation and SMOTE exist, more sophisticated solutions are required. Ethical considerations are paramount. All of these amount to potential for bias in algorithms, privacy ramifications, transparency, and accountability. Contextualizing responsible use of AI in counter-terrorism depends on developing appropriate ethical guidelines and regulations. In addition, there are concerns regarding generalizability and transferability. Models trained on a specific dataset or geographic region often fail to generalize well to other datasets or regions. Developing robust and transferable models is a major challenge. The keyword-based labeling approach may miss nuanced expressions of extremism or code language used by extremists. Future research should investigate more robust labeling techniques, such as manual validation or expert reviews, to mitigate these issues and improve the quality of the dataset. Although the proposed model demonstrates strong performance on English-language Twitter data and GTD derived features, its generalizability to other languages, platforms, and geopolitical contexts warrants further validation. Future research should investigate cross-lingual transfer learning and region-specific adaptations to enhance robustness in diverse extremist narratives and cultural environments.

Extremist narratives can differ greatly from one location to another, culturally and regionally. Future studies should validate the effectiveness of the model over different datasets to enable usage in a number of settings.

5. Conclusion

In response to the ever-growing need for enhanced counter-terrorism intelligence on social media, this research presents a novel hybrid model that integrates deep learning and ensemble techniques to detect, classify, and predict extremist activities on social media. This study demonstrates the strength of leveraging historical records from GTD with real-time social media trends (tweets) to improve the accuracy and efficiency of extremism detection. Robust performance metrics were achieved by the model using advanced feature engineering techniques that utilize deep learning architectures such as Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and stacking classifiers. The model underwent rigorous evaluation, with accuracy, precision, recall, F1-score, and AUC demonstrating its ability to reliably classify extremist content across different categories of content. Beyond counter-terrorism, the methodologies derived in this work are highly relevant to other domains requiring more advanced data analysis and prediction capabilities. The model's applications extend to cybersecurity, misinformation identification, online hate speech monitoring, and public health surveillance, owing to its capability to analyze massive volumes of structured and unstructured data and identify subtle patterns. This research addresses critical gaps in existing methodologies for extremism detection, therefore supporting the field's advancement. This work offers actionable insights to law enforcement agencies for timely interventions against potential digital and physical threats. By incorporating historical data and real-time social media data, the hybrid model, which outperformed traditional models, has demonstrated its potential to enhance predictive analytics, thereby mitigating risks and promoting public safety. As extremist tactics evolve in this digital age, this research establishes a foundation for further exploration of hybrid models that combine explainable AI techniques with innovative data integration strategies to more effectively counter global extremism and terrorism threats.

Future Research Directions

Future research is proposed along several avenues to build on the findings and address the limitations discovered in this study. Understanding regional and cultural variations in extremist narratives requires expanding research to include cross-cultural analyses so that more tailored counter-terrorism strategies can be developed. Advanced feature engineering techniques should be used that leverage multi-modal data, such as images and videos, to provide more in-depth insights about extremist communication patterns. Graph neural networks or knowledge graphs could be applied to further advance the analytical capabilities of this model. Frameworks for responsible AI in sensitive areas must address ethical concerns of privacy and bias in AI applications to ensure transparency and accountability. The hybrid model integration with real-time monitoring systems may deliver actionable insights for law enforcement agencies during critical events or crises. Ultimately, it will be important to foster interdisciplinary collaboration between computer scientists, social scientists, and policy experts to develop AI-driven counter-terrorism solutions.

All the Declarations and Statements

Author Contributions Statement

Gideon Mwendwa – Conceptualization, Data Collection and Processing, Methodology implementation and Results analysis, Preparation, Writing and Editing of Final Manuscript.

Lokesh Chouhan – Supervision and critical reviewing of the manuscript.

Ranjit Kolkar – Reviewing, liaison communication, and overseeing coherence flow of the manuscript.

All authors have read and agreed to publish this manuscript.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Funding Declaration

The Authors declare that this research has not been funded by any agency.

Data Availability Statement

All data used in this study are available from the corresponding author upon reasonable request.

Ethical Declarations

All authors agree to adhere to the guidelines provided by the Journal. This research has considered all ethical issues associated with the methodology being employed.

Acknowledgments

The authors sincerely thank the International Journal of Intelligent Systems and Applications (IJISA) and the anonymous reviewers for their constructive evaluations and valuable recommendations, which significantly improved the experimental design and the reliability of the reported results. The authors also acknowledge the National Forensic Sciences University, Goa Campus, India, for providing institutional support and access to computational resources throughout this research.

Declaration of Generative AI in Scholarly Writing

Generative AI tools were used solely for grammatical correction and formatting consistency. All scientific content, including research design, data collection and analysis, interpretation of results, and final editorial decisions, was undertaken entirely and independently by the authors. The authors have reviewed and approved all content and assume full responsibility for the accuracy, integrity, and originality of the work.

Abbreviations

The following abbreviations are used in this manuscript:

AI - Artificial Intelligence
AUC - Area Under the Curve
CNN - Convolutional Neural Network
DL - Deep Learning
FN - False Negative
FP - False Positive
GTD - Global Terrorism Database
LSTM - Long Short-Term Memory
LR - Logistic Regression
MCC - Matthews Correlation Coefficient
ML - Machine Learning
MLP - Multi-Layer Perceptron
NLP - Natural Language Processing
RF - Random Forest
ROC - Receiver Operating Characteristic
SHAP - SHapley Additive exPlanations
SVM - Support Vector Machine
TF-IDF - Term Frequency–Inverse Document Frequency
TN - True Negative
TP - True Positive
XAI - Explainable Artificial Intelligence

References

- [1] Ö. Sevdiren, "The crime of disseminating terrorism-related propaganda in Turkey: An 'exceptional' case for the concrete endangerment approach?", *Criminal Law Forum*, vol. 35, no. 1, pp. 1–52, 2024.
- [2] International Centre for Counter-Terrorism, "The radicalization (and counter-radicalization) potential of artificial intelligence," ICCT Report, The Hague, 2024. [Online]. Available: <https://www.icct.nl/publication/the-radicalization-and-counter-radicalization-potential-of-artificial-intelligence>
- [3] K. Zhaksylyk, B. Omarov, and S. Mussiraliyeva, "Review of violent extremism detection techniques on social media", in *Proc. 16th Int. Conf. on Electronics Computer and Computation (ICECCO)*, pp. 1–5, 2021.
- [4] M. Bolatbek, M. Sagynay, S. Mussiraliyeva, K. Baisylbayeva, and Z. Yeltay, "Kazakh language dataset for hate speech detection on social media text", in *Proc. IEEE 9th Int. Conf. on Computational Intelligence and Applications (ICCIA)*, pp. 94–98, 2024.
- [5] International Centre for Counter-Terrorism, "Exploitation of generative AI by terrorist groups," ICCT Report, The Hague, 2024.
- [6] K. O. Ogutu and F. O. Nyagwara, "Critical analysis of status of counter-terrorism strategies in Manda, Lamu County, Kenya: A multifaceted examination of HUMINT and SIGINT", *Humanities and Social Sciences*, vol. 13, no. 1, pp. 80–90, 2024.
- [7] V. Dragos and Y. Constable, "Comparison of classification techniques for extremism detection in French social media", in *Proc. 26th Int. Conf. on Information Fusion (FUSION)*, pp. 1–6, 2023.
- [8] M.-A. Richardson, "Predicting violent extremism with machine learning: A scoping review", *SN Computer Science*, vol. 5, no. 1, p. 16, 2023.
- [9] M. Gaikwad, S. Ahirrao, S. Phansalkar, and K. Kotecha, "Online extremism detection: A systematic literature review with emphasis on datasets, classification techniques, validation methods, and tools", *IEEE Access*, vol. 9, pp. 48364–48404, 2021.
- [10] G. Mwendwa, L. Chouhan, F. Etyang, N. Mandela, and U. Gift, "Combating terrorism: Detecting and analyzing extremist content on X (formerly Twitter)", in *Proc. 3rd IEEE Delhi Section Flagship Conf. (DELCON)*, pp. 1–7, 2024.
- [11] M. Xi, L. Nie, and J. Song, "Research on urban anti-terrorism intelligence perception system from the perspective of Internet of Things application", *Int. J. Electrical Engineering & Education*, vol. 58, no. 2, pp. 248–257, 2021.
- [12] M. Gaikwad, S. Ahirrao, K. Kotecha, and A. Abraham, "Multi-ideology multi-class extremism classification using deep learning techniques", *IEEE Access*, vol. 10, pp. 104829–104843, 2022.
- [13] J. Scanlon, "Automatic detection and forecasting of violent extremist cyber-recruitment", M.Sc. thesis, University of Virginia, 2014.
- [14] F. Sufi, "A new social media-driven cyber threat intelligence", *Electronics*, vol. 12, no. 5, p. 1242, 2023.
- [15] S. K. Habibabadi, C. Palmer, G. L. Dimaguila, M. Javed, H. J. Clothier, and J. Buttery, "Automated social media surveillance for detection of vaccine safety signals: A validation study", *Applied Clinical Informatics*, vol. 14, no. 1, pp. 1–10, 2023.
- [16] A. K. Yee, "Construct validity in automated counterterrorism analysis", *Philosophy of Science*, pp. 1–18, 2024.
- [17] S. Aldera, A. Emam, M. Al-Qurishi, M. Alrubaiyan, and A. Alothaim, "Exploratory data analysis and classification of a new Arabic online extremism dataset", *IEEE Access*, vol. 9, pp. 161613–161626, 2021.
- [18] J. Bartlett and L. Reynolds, "The state-of-the-art 2015: A literature review of social media intelligence capabilities for counter-terrorism," *Tech. Rep., Demos*, London, U.K., 2015.
- [19] J. Torregrosa, G. Bello-Ortiz, E. Martínez-Cámara, J. Del Ser, and D. Camacho, "A survey on extremism analysis using natural language processing: Definitions, literature review, trends and challenges", *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 8, pp. 9869–9905, 2023.
- [20] F. A. Khan, G. Li, A. N. Khan, Q. W. Khan, M. Hadjouni, and H. Elmannai, "AI-driven counter-terrorism: Enhancing global security through advanced predictive analytics", *IEEE Access*, vol. 11, pp. 135864–135879, 2023.
- [21] E. Rodilosso, "Filter bubbles and the unfeeling: How AI for social media can foster extremism and polarization", *Philosophy & Technology*, vol. 37, no. 2, p. 71, 2024.
- [22] I. Kim, W. M. Pottenger, and V. Behe, "Deep learning-based named entity recognition using automatic labeling of the global terrorism database", in *Proc. IEEE Int. Symp. on Technologies for Homeland Security (HST)*, pp. 1–6, 2018.
- [23] S. Mussiraliyeva, B. Omarov, P. Yoo, and M. Bolatbek, "Applying machine learning techniques for religious extremism detection on online user contents", *Computers, Materials & Continua*, vol. 70, no. 1, pp. 915–934, 2022.
- [24] A. S. Al-Falooji and A. Al-Azawei, "Predicting users' personality on social media: A comparative study of different machine learning techniques," *Karbala Int. J. Mod. Sci.*, vol. 8, no. 4, p. 5, 2022.
- [25] A. Falzon and J. Azzopardi, "TerrorMine: Automatically identifying the group behind a terrorist attack", in *Proc. 14th Int. Joint Conf. on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K-KDIR)*, pp. 221–228, 2022.
- [26] N. Rawat, S. Chauhan, and L. K. Awasthi, "Deep learning approaches for predicting mental states through tweet analysis", in *Proc. 5th Int. Conf. on Innovative Trends in Information Technology (ICITIT)*, pp. 1–6, 2024.
- [27] A. Z. Rozo, D. Díaz-López, J. Pastor-Galindo, F. Gómez Mármol, and U. Karabiyik, "An NLP-based framework to spot extremist networks in social media", *Complexity*, vol. 2024, no. 1, p. 3380488, 2024.
- [28] E. Shadadi, S. Kouser, L. Alamer, and P. Whig, "Novel approach of predicting human sentiment using deep learning," *Journal of Computer Science and Engineering*, vol. 3, no. 2, 2022.
- [29] R. K. Mishra, S. Urolagin, J. A. A. Jothi, A. S. Neogi, and N. Nawaz, "Deep learning-based sentiment analysis and topic modeling on tourism during COVID-19 pandemic", *Frontiers in Computer Science*, vol. 3, 2021.
- [30] F. Malcolm, "Analysing extremism", *Ethical Theory and Moral Practice*, vol. 26, no. 2, pp. 321–327, 2023.
- [31] S. Aslan, "A novel TCNN-Bi-LSTM deep learning model for predicting sentiments of tweets about COVID-19 vaccines", *Concurrency and Computation: Practice and Experience*, vol. 34, no. 28, e7387, 2022.
- [32] M. Abdalsalam, C. Li, A. Dahou, and N. Kryvinska, "Terrorism attack classification using machine learning: Effectiveness of textual features extracted from GTD dataset," *CMES – Computer Modeling in Engineering & Sciences*, vol. 138, no. 2, 2024.
- [33] M. Adnan and M. Rafi, "Extracting patterns from global terrorist dataset using co-clustering approach", *Journal of Independent Studies and Research*, vol. 13, no. 1, p. 7, 2015.
- [34] United Nations Counter-Terrorism Centre and United Nations Interregional Crime and Justice Research Institute, "Countering terrorism online with artificial intelligence", Technical Report, United Nations, 2021.

- [35] A. Anil, S. R. Singh, and R. Sarmah, "Personalised PageRank as a method of exploiting heterogeneous network for counter terrorism and homeland security", in Proc. IEEE/WIC/ACM Int. Conf. on Web Intelligence, pp. 327–334, 2016.
- [36] G. Mwendwa, A. Nagpal, S. Garg, L. Chouhan, and N. Mandela, "Graph neural networks for temporal terrorist network analysis and multi-relational threat prediction," International Journal of Information Technology, 2025, doi: 10.1007/s41870-025-02441-3.
- [37] T. Xia and Y. Gu, "Building terrorist knowledge graph from global terrorism database and Wikipedia", in Proc. IEEE Int. Conf. on Intelligence and Security Informatics (ISI), pp. 194–196, 2019.
- [38] R. Q. Al-Fayez, M. Al-Tawil, B. Abu-Salih, and Z. Eyadat, "GTDonto: An ontology for organizing and modeling knowledge about global terrorism", Big Data and Cognitive Computing, vol. 7, no. 1, p. 24, 2023.
- [39] N. V. Khokhlov and A. V. Korotaev, "Internet and terrorist activity in countries with different political regimes: A quantitative analysis", Politeia, vol. 97, no. 2, pp. 67–86, 2020.
- [40] S. O. Uche, N. Tsopzé, and D. U. Ebem, "Data mining approach to counterterrorism", Advances in Multidisciplinary & Scientific Research, vol. 11, no. 2, pp. 77–90, 2020.
- [41] G. Mwendwa, L. Chouhan, H. A. Elsayed, A. Mehane, W. Al Zoubi, J. Wekalao, and Y. Prabhu, "Deciphering terrorist attributions: A global terrorism data-driven comparison of machine learning models", Social Network Analysis and Mining, vol. 15, no. 1, p. 99, 2025.
- [42] United Nations Interregional Crime and Justice Research Institute (UNICRI), "Countering terrorism online with artificial intelligence: An overview for South Asia and South-East Asia", UNICRI Publications, 2024.
- [43] R. Frankenberger, R. Baur, M. Rieger-Ladich, J. Schmid, B. Stauber, A. Thiel, and T. Thomas, "Researching far right extremism: A transdisciplinary, lifeworld, and political culture perspective", Zeitschrift für Vergleichende Politikwissenschaft, vol. 17, no. 3, pp. 275–295, 2023.
- [44] G. M. Ndungi, L. Chouhan, and F. Etyang, "Systematic literature review on the synergy of social media forensics and mobile forensics in the investigation of cybercrimes", in Advances in Data-Driven Computing and Intelligent Systems, Springer Nature Singapore, pp. 503–520, 2024.
- [45] E. Jesse, "Left-wing extremism in Germany", in The Palgrave Handbook of Left-Wing Extremism, Springer International Publishing, pp. 173–191, 2023.
- [46] Y. Tian, R. Fan, X. Ding, X. Zhang, and T. Gan, "Predicting rumor retweeting behavior of social media users in public emergencies", IEEE Access, vol. 8, pp. 87121–87132, 2020.
- [47] N. Sharma, I. Batra, and A. Malik, "Comparative study of malware detection using hybrid deep learning techniques", in Proc. ASU Int. Conf. on Emerging Technologies for Sustainability and Intelligent Systems (ICETIS), pp. 1260–1266, 2024.
- [48] C. Orellana-Rodriguez and M. T. Keane, "Modeling and predicting news consumption on Twitter", in Proc. 26th Conf. on User Modeling, Adaptation and Personalization, pp. 321–329, 2018.
- [49] R. Chellappa, S. Theodoridis, and A. van Schaik, "Advances in machine learning and deep neural networks", Proceedings of the IEEE, vol. 109, no. 5, pp. 607–611, 2021.
- [50] Digital Watch Observatory, "AI and violent extremism – trends in 2025," 2025. [Online]. Available: <https://dig.watch/topics/violent-extremism>. [Accessed: Jan. 2025].
- [51] L. Ritter, M. A. Weber, R. Witt, V. Profes, J. P. Singh, J. Endrass, and A. Rossegger, "Improving violent extremism risk communication among German police agencies: A survey of federal and state threat managers," Journal of Police and Criminal Psychology, vol. 38, no. 4, pp. 1–10, 2023.
- [52] D. Karanatsiou, P. Sermpetis, D. Gruda, K. Kafetsios, I. Dimitriadis, and A. Vakali, "My tweets bring all the traits to the yard: Predicting personality and relational traits in online social networks," ACM Transactions on the Web, vol. 16, no. 2, 2022, doi: 10.1145/3527625.
- [53] S. Mussiraliyeva, K. Baisylbayeva, M. Bolatbek, and Z. Yeltay, "Decoding ideology: Machine learning-based detection of extremist content", in Proc. Int. Conf. on Intelligent Computing, Communication, Networking and Services (ICCNS), pp. 114–120, 2024.
- [54] P. V. Reddy, S. Sanku, T. Akhilsai, and A. Gayatri, "CNN based criminal face identification system using video surveillance," SSRN Electronic Journal, 2024.

Authors' Profiles



Gideon Mwendwa is a Ph.D. scholar in Computer Science and Technology at the National Forensic Sciences University, India. He holds an M.Sc. in Digital Forensics and Information Security (First Class with Distinction) from NFSU, India, an M.Sc. in Cybersecurity (Distinction), and is a finalist for an M.A. in Social Transformation (Governance) from Tangaza University, Kenya; complemented by a B.Sc. in Information Technology and Diploma in Information and Communication Technology (Distinction). Holding specialized training in Digital Forensics from Florida International University (FIU), USA, along with Certified Expert in Cyber Investigations (CECI)-UK and Certified Ethical Hacker (CEH) credentials, his doctoral research advances artificial intelligence applications in

national security, emphasizing extremism detection, attribution, risk-oriented inference, and counter-terrorism intelligence systems.



Dr. Lokesh Chouhan is Dean and Associate Professor of Cyber Security at the National Forensic Sciences University (NFSU), Goa Campus, under the Ministry of Home Affairs, India. He has previously served as Dean (Academics) and faculty at IIIT Surat, NIT Hamirpur, MANIT Bhopal, and IIIT Gwalior, where he earned his Ph.D. in Cognitive Radio. His research spans Cyber Security and Forensics, Computer Networks, Cloud Computing, IoT, Smart Cities, and Cyber Warfare. He has published over 60 papers, contributed to book chapters, holds two Indian patents in IoT, and currently supervises nine research scholars. He also serves on technical program committees and as a reviewer for leading IEEE, ACM, Elsevier, IET, and Springer venues.



Ranjit Kolkar is an Assistant Professor of Computer Science at NFSU Goa, holding a Ph.D. from the National Institute of Technology Karnataka (NITK), with expertise in AI, Cybersecurity, Digital Forensics, and IoT. His research covers AI-driven threat detection, blockchain forensics, smart environment security, NLP-based deception analysis, and optimization algorithms. His key contributions include a doctoral framework for human activity and behavior recognition in multimodal smart home environments and secure digital currency systems. He actively mentors students and develops AI/IoT laboratories for forensic computing and national security.

How to cite this paper

G. Mwendwa, L. Chouhan, and R. Kolkar, “Deep Ensemble Hybrid Model for Extremism Detection and Threat Inference in Counter-Terrorism Intelligence on Social Media,” *International Journal of Intelligent Systems and Applications (IJISA)*, Vol. 18, No. 4, pp.161–182, 2026. DOI: 10.5815/ijisa.2026.04.09