HYBERT-X: A Hybrid Deep Learning Framework for Cyberbullying and Cyber Threat Detection

Jayapriya J. [1, *], Kavitha S. [1], Manimekala B. [1], and Nileem Kaveramma [2]

[1] Department of Computer Science, Christ University, Bangalore, India
[2] Data Science Intern, Aditya Birla Group, Karnataka, India
Email: jayapriya.j@christuniversity (J.J.); kavitha.s@christuniversity.in (K.S.); manimekala.b@christuniversity.in (M.B.); nileem.kaveramma@msds.christuniversity.in (N.K.)
*Corresponding author

***Abstract*—Cyberbullying and online threat detection are the new vulnerabilities of cybersecurity, and the conventional tools are unable to keep pace with the development of harmful content. The paper discusses HYBERT-X as a hybrid deep learning model that integrates Bidirectional Long Short-Term Memory (BiLSTM), Graph Convolutional Network (GCN) based learning, Bidirectional Encoder Representations from Transformers (BERT) and Extreme Gradient Boosting (XGBoost) to detect threats with high accuracy and interpretability. The proposed framework uses an ensemble classification layer to boost the accuracy and interpretability of detection based on semantic, temporal, and structural representation. The experimental assessment done on the typical benchmark datasets has shown that HYBERT-X is up to 95.53% accurate in comparison with the currently available baseline models both in generalization ability and explainability. The findings suggest that this work is computationally effective, scalable, and can be adapted to real-time cybersecurity applications and is applicable to next-generation cybersecurity systems against online threats and cyberbullying.**

***Keywords*—cyberbullying detection, cybersecurity, hybrid deep learning, Bidirectional Encoder Representations from Transformers (BERT), Extreme Gradient Boosting (XGBoost)**

I. Introduction

The period of the internet age has created a situation of being ever-connected, and this new form of communication on the internet has not only made life more accessible but also raised its own level of concerns regarding security and abuse of the internet. Amongst the many crimes that are increasing in number, cyberbullying and online threats stand out as a threat to the safety of individuals. A threat to one's safety also means a threat to one's mental health, which in turn can become a threat to organizational integrity or national security [1]. Given that the platforms are unregulated and the internet offers anonymity to all evil acts, there is a technical urgent scenario that needs to address the immediate detection and resolution of the threats in real time [2].

Natural Language Processing (NLP) is an artificial intelligence discipline concerned with the comprehension by computers and understanding of human language. It allows machines to process natural language in a meaningful way. The last ten years have been years of great advancement for NLP, with a substantial impact on areas such as virtual assistants, language translation, sentiment analysis, and intelligent search engines. Cybersecurity has introduced NLP as an innovative tool in the analysis and securing of digital communications [3, 4]. The rapid increase in user-generated data on social media platforms, forums, emails, and chats has increased the risk of cyber threats such as phishing, cyberbullying, fake news, hate speech, etc. It can also be used for the detection of suspicious patterns, classification of harmful intent, most importantly key indicator extraction and eventually, all those together using real-time online threat monitoring. NLP endows the security systems with capabilities to process and analyze unstructured textual data at large scales, which usually remains beyond the reach of conventional security mechanisms.

Even with these improvements, cyber threats have become more complex and harder to identify using traditional methods due to the enormous increase in textual information on digital platforms. The conventional methods of machine learning and some of the deep learning models have demonstrated weaknesses in the complete embodiment of semantic, structural, and contextual traits of text in the context of cybersecurity-related information. These constraints show that more powerful and dynamic learning strategies are required to manage the changing online threat in the changing digital space [5].

To explore these limitations, the proposed work aims to integrate semantic, structural, and sequential representations for the detection of cyberbullying and

Manuscript received February 18, 2026; revised March 10, 2026; accepted April 27, 2026; published July 28, 2026.

online threats by developing a hybrid deep learning framework.

The existing hybrid models combine limited learning paradigms, whereas the proposed HYBERT-X framework integrates semantic (Bidirectional Encoder Representations from Transformers (BERT)), structural (Graph Convolutional Network (GCN)-style features), sequential (Bidirectional Long Short-Term Memory (BiLSTM)), and ensemble-based decision learning (Extreme Gradient Boosting (XGBoost)), introducing a multi-level feature fusion architecture. This consolidated combination results in capturing the portrayed cyberbullying content. The novelty of the framework lies in creating a comprehensive pipeline in which each component adds a unique and non-redundant representational benefit, making it more robust.

The main contributions of this work are:

(1) Combination of transformer-based embedding, graph-based features, sequence-based modeling, and ensemble classification;
(2) A property fusion model that is more robust to detection and generalization;
(3) Extensive experimental analysis based on baseline models.

The remaining sections of the paper are structured in the following manner. Section II will discuss the literature. The proposed framework is described in Section III. Section IV is the description of experimental design and evaluation measures. Section V interprets the results obtained. In section VI, the conclusions are given together with recommendations to be followed by future research study.

II. Literature Review

The sheer volume of user-created content in social media outlets has grown exponentially over the past few years and so has the horrendous growth of bullying, harassment, and online abuse among users. All these have made researchers and developers to develop advanced Artificial Intelligence (AI)-based ways of countering and reducing online harmful behavior [6–13]. Literature in this field indicates that there is active progression, first using old machine learning strategies and then eventually arriving at the deep learning-based approaches and transformer-based approaches. These strategies have been implemented in a variety of cybersecurity applications, such as cyberbullying, phishing, hate speech, and online abuse identification. Although deep learning models have significantly outperformed classical methods, the outgoing literature shows that most of the methods still cannot deal with implicit language and contextual ambiguity and require flexibility between platforms and domains [14–31].

To help limit themselves, recent research has shown more interest in hybrid modeling strategies involving the integration of various learning paradigms. The goal of such hybrid methods is to make use of the complementary strengths of the various methods, such as semantic understanding, sequential modeling, as well as feature-driven classification, to support the detection robustness. Some important ones are scarcity of data, interpretability, bias, multilingualism, privacy, adversarial vulnerability, and lack of standardization. These gaps indicate the necessity of more comprehensive and flexible learning designs that could be effective in meeting the complexity of the nature of cyber threats in online dynamic settings [32–37].

III. Proposed Work

The emergence of online platforms as the main place for cyberbullying and harassment has been a hurdle for automated content moderation systems to fight the issue effectively. Even though breakthroughs have been achieved in NLP and deep learning, the leading systems are still not even close to detecting all the different types of harmful speech that are subtle, implicit, or even new. The prevailing models are rigid, do not take the context into consideration, or struggle to extend their application to new areas, especially in capturing the fast movement of social media, where the language varies significantly from day to day.

A new method that focuses on strength and versatility has been introduced to deal with the aforementioned constraints. The proposed hybrid model alters the classical classification pipelines to an all-around detection framework that uses various learning strategies. The model is a hybrid model which incorporates semantic knowledge, structural representation and time sequence modelling frameworks into a multi-layer structure. This multi-faceted approach is focused on enhancing the capacity of the model to detect even the smallest signals that would not have been detected otherwise.

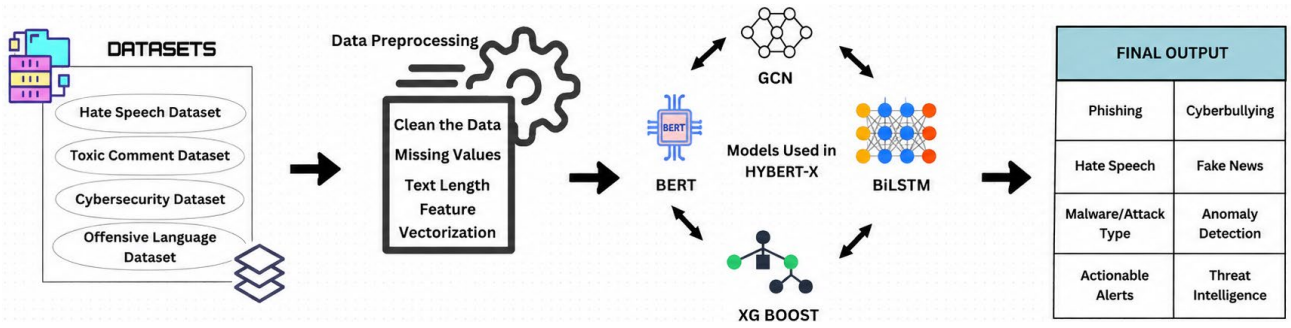

Fig. 1. Working of HyBERT-X model.

Fig. 1 shows the working of HYBERT-X, which has been set up with a bunch of datasets consisting of hate speech, toxic comments, cybersecurity content, and offensive language, thus offering a variety of bases for training and evaluation purposes. Data undergoes fully-fledged preprocessing, including cleaning operations and handling missing values, along with text-length analysis and vectorizations: the conversion of raw text into numerical features that can be fed into a machine-learning system.

The architecture followed a process in which four advanced components were intermingled:

(1) Embeddings extraction by BERT with deep context consideration;
(2) GCN capturing relational dependencies between features;
(3) BiLSTM, which models sequential patterns in text, and lastly;
(4) XGBoost that does classification on a robust and efficient basis.

These frameworks together produce the hybrid model called HYBERT-X. The detection at the output layer covers cybersecurity and online abuse, such as phishing, cyberbullying, hate speech, fake news, malware types, anomalies, actionable alerts, and threat intelligence.

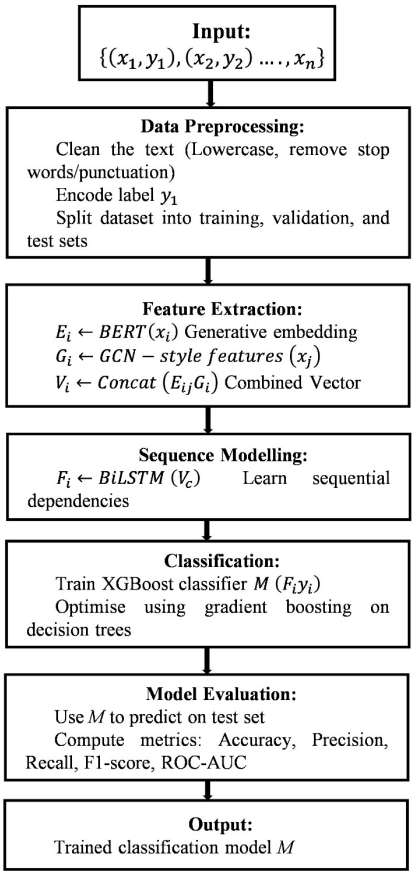

Fig. 2. Algorithmic workflow of the hybrid HYBERT-X model.

Fig. 2 shows the basic algorithmic workflow of the hybrid HYBERT-X model. It begins with a labelled dataset that includes text samples and their respective labels. The initial stage of data preprocessing involves cleaning texts via procedures such as lowercasing, stop word removal, and punctuation filtering. In addition, there is the task of encoding labels and splitting the dataset into training, validation, and testing sets. The feature extraction block extracts the contextual embeddings from BERT and structural features from GCN-style representations. The features are concatenated to produce a single vector capturing semantic as well as topological information.

This combined vector is submitted to the Sequence modelling stage, which then employs a BiLSTM to model temporal and contextual dependencies across the word sequences. BiLSTM outputs are forwarded to the Classification stage, wherein an XGBoost classifier is trained using gradient boosting over decision trees to construct the final set of hypotheses. During the Model Evaluation phase, the trained model is tested on the test set with metrics measuring accuracy, precision, recall, F1-Score.

The essence of hybridization comes from the fact that hybrid models attempt to capture as many broader and more complex models, thereby including distinctions that no single model type would be able to handle alone. In cyberbullying, the name-calling or worse behaviours could be either very explicit or heavily embedded in tone, context, or conversational history. Transformer-type models, such as BERT, will excel in contextual understanding but may lack the knowledge about the structure and relation in language. On the other hand, graph-based learning will excellently identify these innate user or linguistic connections, whereas sequence models like BiLSTM will still be able to help in tracing the flow and intent through time. Such collaboration will significantly influence the accuracy of the detection, since this will also facilitate the system to counter various forms of attacks hence rendering it even more applicable to practical application in quick platform changes.

The suggested construction views cyberbullying detection as a multiple perspective threat recognition problem that is not only a text classification task but also a text semantic meaning task, a structural relation task, and a temporal pattern task. This architecture triggers better robustness and adaptability of realistic content moderation systems.

IV. Materials and Methods

Among the mounting challenges in identifying cyber threats and cyberbullying, the paper applies a compounded methodological design that integrates some of the recent machine learning and deep learning algorithms. The method is called HYBERT-X which implies that the study deals with model combinations that provide optimal performance in detection, contextualization, and interpretability. The existing section will be a summary of the gradual approach that was followed when developing, training, and evaluating the model, and data preprocessing, feature extraction, model construction, and performance evaluation. By way of the methodology, the linguistic and structural layers of the online material will be thoroughly examined, which is essential for the effective detection of threats in real time.

A. Dataset Description

A publicly available cyberbullying detection dataset is taken from Kaggle, which integrates data from various social media platforms such as Twitter, Wikipedia Talk Pages, YouTube, and other sources. Binary class classification can be seen in the dataset, bifurcating the data into cyberbullying and non-cyberbullying.

The dataset consists of multiple forms of abusive language, such as aggression, hate speech, toxicity, and insults. This ensures the coverage of real-world cyberbullying scenarios and enhances the generalization capability of the proposed model. XGBoost classifier effectively handled the imbalance in class distribution caused from splitting of the dataset into training and testing sets using an 80:20 ratio. The summary of the dataset characteristics is provided in Table I.

TABLE I. SUMMARY OF THE CYBERBULLYING DETECTION DATASET

Dataset	Cyberbullying Detection Dataset
Source	Kaggle (multi-platform)
Task	Binary classification
Classes	Bullying/non-bullying
Split	80:20

B. Proposed Method: HYBERT-X

The model HYBERT-X proposed here is a hybrid framework for the detection of cyberbullying that merges semantic embeddings with graph-structured features, sequential modelling, and classification based on gradient boosting. This structure combines the advantages of the four most powerful elements, such as BERT, GCN-style feature extraction, BiLSTM, and XGBoost. The process is explained step by step in the forthcoming subsections.

1) Input and output

Let Eq. (1) be a labelled dataset consisting of social media text posts, each x_i and binary class labels $y_i \in \{0, 1\}$ to denote the classes (cyberbullying or not). The model results in a classifier that can anticipate y^i new inputs.

$$D = \{(x_i, y_i \mid i = 1,2, \ldots., n)\} \quad (1)$$

2) Data preprocessing

The normalization process for text inputs consists of changing the case to lower, eliminating stop words, punctuation, and special characters. The labels y^i are also encoded properly. In this manner, the dataset is divided into subsets for training, validation, and testing, which makes it possible to carry out a strong evaluation.

3) Feature extraction

Each input text x_i undergoes dual feature extraction:

Semantic Embedding via BERT: A pre-trained BERT model encodes contextual representations:

$$E_i = BRET\ (x_i) \quad (2)$$

where x_i is the *i*-th input text instance, $E_i \in R_d$ is the dense vector representation (embedding) of the input text produced by BERT. d denotes the dimensionality of the BERT embedding.

The BERT model captures both the contextual meaning and the semantic relationships of words in the sentence.

4) Structural features via GCN-style embedding

Graph-based features G_i are derived using syntactic dependency graphs or co-occurrence relations to capture structural semantics:

$$G_i = GCN\ (x_i) \quad (3)$$

where:

- x_i is the *i*-th input text instance.
- $G_i \in R_k$ is the graph-based feature vector extracted from x_i.
- $GCN(\cdot)$ refers to the Graph Convolutional Network applied over the graph representation of the sentence.The graph is constructed from syntactic dependencies (e.g., dependency parse trees) or word co-occurrence graphs.
- k denotes the dimensionality of the structural feature space.

This embedding captures relational structure and context beyond linear word order.

The resulting embeddings are concatenated to form a comprehensive feature vector:

$$V_i = [E_i \,||\, G_i\,] \quad (4)$$

where:

- E_i typically represents textual embeddings, such as those from BERT, (Term Frequency-Inverse Document Frequency) TF-IDF, or BiLSTM.
- G_i represents graph-based features, such as those from a GCN or any structural information.

To form the adjacency matrix of the GCN component, an approach that relies on co-occurrences is used to represent nodes (symbolizing tokens) and form edges between them through words sharing a considered context window. The co-occurrence frequency is used to determine the edge weights reflecting the strength of association between tokens.

Also, social media-specific features like user mentions (`@user) and hashtags (`#topic) are considered as important nodes, and the relationships are established between them and the words around them. This adjacency matrix is then normalized and fed into a multi-layer GCN to allow the model to encode higher order dependencies between structures.

5) Sequence modelling

To capture temporal and contextual dependencies, the concatenated vector V_i is passed through a BiLSTM network:

$$F_i = BiLSTM\ (V_i) \quad (5)$$

where V_i is the input at time step *i*. It is a fused vector made by concatenating: E_i is textual features (e.g., from BERT or TF-IDF); G_i: is Graph-based features (e.g., from GCN).

BiLSTM is a neural network architecture that processes sequences in both forward and backward directions, allowing the model to capture:

- Past context (via the forward LSTM)
- Future context (via the backward LSTM)

F_i is the output hidden state from the BiLSTM at time step i, which now encodes:

- The temporal dependencies (how current input relates to previous/future ones)
- The contextual relationships across the entire sequence

6) Classification with XGBoost

The feature representation F_i is used to train an XGBoost classifier:

$$y^i = XGBoost\ (F_i) \tag{6}$$

where:

- F_i is the contextual feature vector produced by the BiLSTM, encoding both semantic and structural information for the i-th instance.
- y^i is the predicted label (e.g., class of cyberbullying or not) for the i-th input.
- XGBoost is a powerful, tree-based ensemble learning algorithm that performs well on structured and tabular data and handles complex decision boundaries.

XGBoost leverages gradient boosting over decision trees to optimize classification performance and handle complex, non-linear decision boundaries.

7) HYBERT X model equation

Let:

- x_i = Input text sample;
- y^i = True label;
- $E_i = BERT\ (x_i)$ Semantic embedding from BERT;
- $G_i = GCN\ (x_i)$ Structural embedding from GCN-style feature extractor;
- $V_i = [E_i \,||\, G_i]$ Concatenated feature vector;
- $F_i = BiLSTM\ (V_i)$ Feature output from BiLSTM;
- $y^i = M(F_i)$ Predicted label from XGBoost model MMM;

The overall flow of the HYBERT-X model can be mathematically represented as Eq. (7):

$$HYBERT - X\left(x_i\right) = XGBoost\left(BiLSTM\left(\left[BERT\left(x_i\right) GCN\left(x_i\right)\right]\right)\right) \tag{7}$$

where:

- $BERT\ (x_i)$ Embeds input text semantically.
- $GCN\ (x_i)$ Extracts graph-based structural features.
- $[\cdot \,||\, \cdot]$ Denotes concatenation of features.
- BiLSTM: Captures sequential dependencies over the combined vector.
- XGBoost: Performs final classification using boosted decision trees.

C. Architecture and the Workflow of this System

This study is about a hybrid deep learning analyser of cyber-bullying that amalgamates varied components to extract semantic, sequential, and graph-based information. The strategy has the transformer-based language models, graph-inspired features extraction, sequential modelling, and ensemble classification embedded in the detection pipeline, which ensures robustness and scalability. The architecture and the workflow of this system are as follows.

1) Data preprocessing and representation

The dataset passed through a pre-processing phase in which common NLP methods were used such as lowercasing, removal of stop words, tokenization and filtering of noise among others. Applying label encoding to categorical variables and standard scaling for the other numerical inputs, the entire dataset has been shuffled randomly, segregating it into train, validation, and test sets to avoid overfitting and to generalize well.

2) Contextual embedding via BERT

The raw textual data has been converted to high-dimensional semantic vectors with BERT. The model of BERT that is pre-trained and fine-tuned on the cyberbully dataset provides context embeddings that capture very slight nuances in terms of language cues and patterns of discourse. These embeddings would thus serve as the brain of the model. It is shown in Fig. 3.

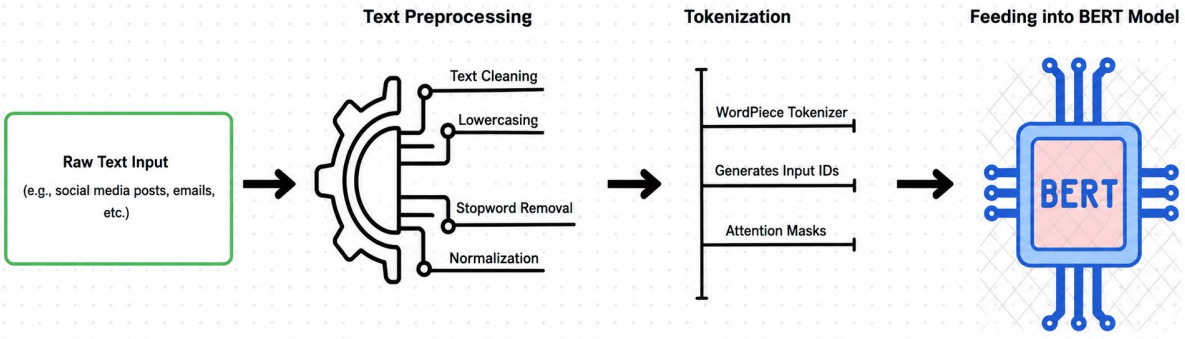

Fig. 3. BERT model input processing pipeline.

3) GCN-inspired feature engineering

Graph-structured attributes abstracted from the textual data augment semantic features. Such features are typically GCN-style features characterizing co-occurrence matrices or interaction graphs; hence, user mentions, hashtags and proximity are ways of emulating topological relationships without requiring a clear graph convolutional network. This stage introduces relational awareness, thus

improving pattern recognition in social context scenarios as well. The flow of activity of GCN is represented in Fig. 4.

4) Temporal modelling using BiLSTM

Fig. 5 shows the BiLSTM-based sequence modelling architecture. The concatenated semantic and structural features are passed through a BiLSTM network, which allows processing or flow of data in both forward and backward directions. BiLSTM is primarily fit for capturing long-range dependencies and temporal sequences in conversational or narrative texts. It, thus, increases the system's capacity in evolving patterns in language, slang, and threats.

5) Final classification with XGBoost

To utilize both deep and structured features, output representations from BiLSTM and GCN-like encodings are input to XGBoost, a scalable and high-performance ensemble learning classifier. By combining weak learners into a strong predictor, XGBoost improves model interpretability and performance, especially when dealing with imbalanced classifications such as cyberbullying. The XGBoost classification pipeline is shown in Fig. 6.

6) Model evaluation and comparative analysis

The proposed hybrid model for its evaluation includes the following metrics: accuracy, precision, recall, F1-Score. This hybrid model achieved an accuracy of 95.53% with great prediction of cyberbullying. Comparative experimentation has been conducted with baseline models that included, but were not limited to, convolutional neural networks, traditional machine learning classifiers such as support vector machines and random forests, and other deep learning approaches. This validated the superiority and generalization capacity of the hybrid.

The HYBERT-X model has been evaluated across multiple runs with various initializations to ensure robustness. The stability of the proposed framework is proved as the performance metrics remained constant with minimal variance.

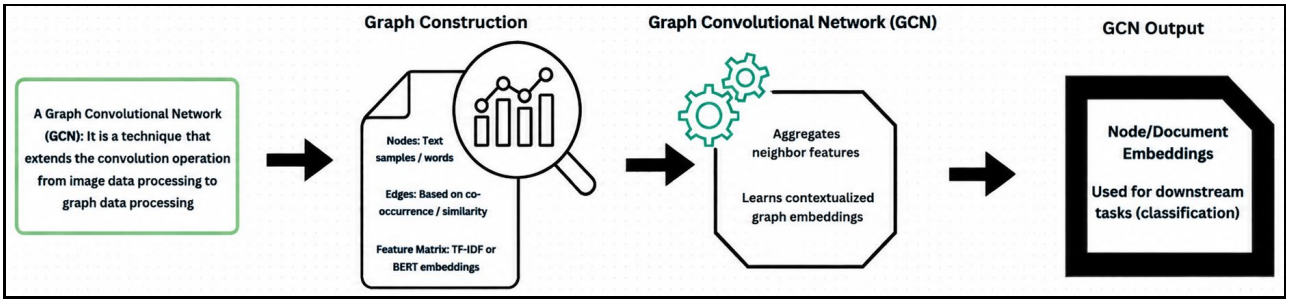

Fig. 4. Workflow of GCN.

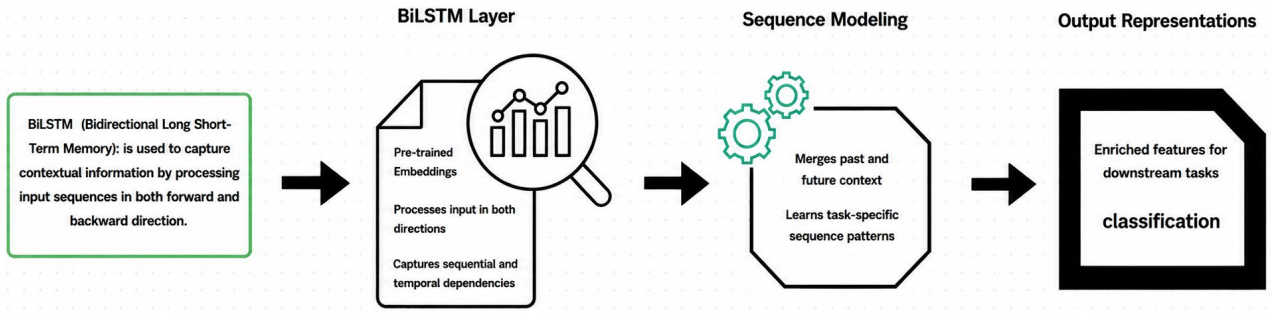

Fig. 5. BiLSTM-based sequence modelling architecture.

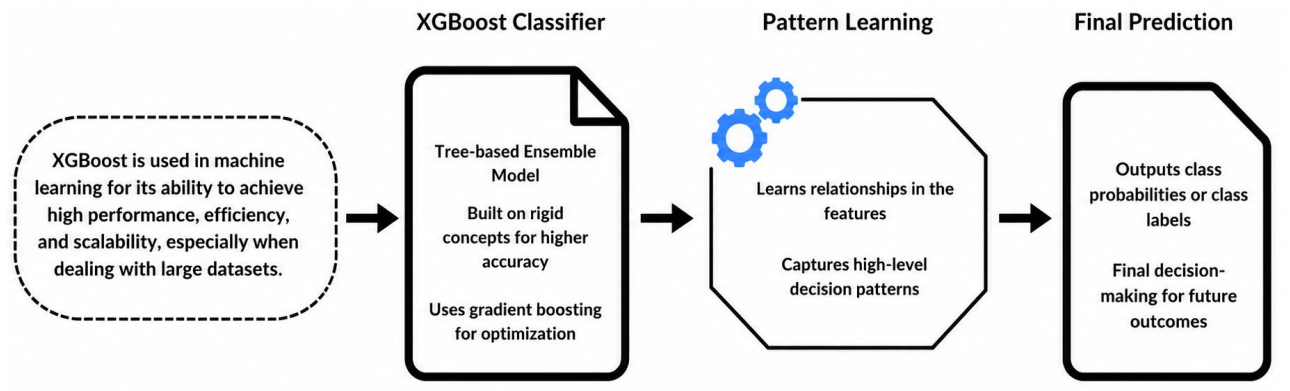

Fig. 6. XGBoost classification pipeline for threat detection.

V. Result and Discussion

This work provides a well-comparative study between standalone models, BERT, GCN, BiLSTM, and XGBoost, and hybridized performance when fed into the proposed hybrid model, HYBERT-X. BERT, GCN, BiLSTM, and XGBoost are four models with diverse capabilities toward better performance of the hybrid: semantic representation, structural learning, temporal sequencing, and robust classification, respectively. This section imparts standalone performance, critical parameters, and synergistic effect in a hybrid model.

Table II represents the individual constituents of the HYBERT-X hybrid framework and their standalone performance metrics together with functional roles. Each model brings its capabilities to the hybrid system. BERT, with 94% accuracy and F1-Score, is the most suitable for improving semantic interpretation and gaining contextual and precision and recall capacity over various cyberbullying examples. A GCN adds further value with 87% accuracy, thereby uncovering structural and network-level dependencies that a sequential mode may not do. The BiLSTM enhances the contextual flow of the review by capturing bidirectional dependencies, in terms of sequence and structure, better than the standalone model, except for its lower value of 85%. XGBoost, on the other hand, is a very strong classifier, with an accuracy and F1-Score of 94% and 95%, respectively, ensuring good generalization from the fused feature set of the deep learning methods. Together within HYBERT-X, this suite of models complements each other in building a strong, end-to-end cyberbullying detection system.

TABLE II. Component Models and Their Roles in the Hybert-X Hybrid Framework

Model	Standalone Performance	Role in HYBERT-X
BERT	Accuracy: 94% F1-Score: 95% Learning Rate: 2e−5 Batch Size: 32 Epochs: 3	Enhances the semantic understanding, improves contextual richness, and boosts precision and recall.
GCN	Accuracy: 87% 2-layer architecture Hidden Dim: 64 Learning Rate: 0.01	Captures network-level patterns and latent relationships not visible through sequential models alone.
BiLSTM	Accuracy: 85% Hidden Units: 128 Dropout: 0.5 Optimizer: Adam	Refines contextual flow and enhances the model's ability to interpret the order and structure of text.
XGBoost	Accuracy: 94% F1-Score: 95%	Delivers robust classification, leveraging fused features for improved accuracy and generalization.

Table III exhibits the comparative analysis in terms of accuracy, precision, recall, and F1-Score for each individual model towards the performance of the proposed HYBERT-X hybrid model. BERT was a top performer even as an independent model and secured the best scores across the board making its robust semantic comprehension and context analysis capabilities very clear. XGBoost is almost as good, with stronger recall (95.1%) and F1-Score (95%) in the classification task. GCN was mediocre with comparatively balanced scores, contributing by means of capturing relational data patterns, while BiLSTM had the lowest F1-Score of 80%, accounting for sequential dependency and flow of text.

TABLE III. Performance Comparison of Individuals Models and the Hybert-X Hybrid Framework

Model	Accuracy	Precision	Recall	F1-Score
BERT	94%	94%	95%	95%
GCN	87%	86%	87%	87%
BiLISTM	85%	83%	84%	80%
XG Boost	94%	94.2	95.1	95
HYBERT-X	95.53%	96%	96%	96%

HYBERT-X clearly outperforms when compared to all the individual systems, also securing the highest scores: 95.53% accuracy, 96% precision, 96% recall, and 96% F1-Score. This indicates that each additional component contributes equally. Contextual semantic understanding by BERT, structural and relational dependencies by GCN, sequential patterns by BiLSTM, and enhanced classification by XGBoost, demonstrate the superiority yielded by multi-architecture integration involving semantic, structural, sequence, and gradient-boosted decision-making understandings to deliver a more holistic and effective cyberbullying detection system.

A. BERT: Standalone vs. in HYBERT-X

BERT, as a standalone model, showed good context understanding with an accuracy of 95% and F1-Score of 95%. The model was fine-tuned at a rate of 2e-5, with a batch size of 32, and for 3 epochs. BERT was used in semantic encoding in HYBERT-X, thereby contributing to the rich contextual embeddings in the feature space. Although the personal performance of BERT could not be evaluated when incorporated into HYBERT-X, it was noted that BERT contributed greatly to the enhancement of the semantic and contextual depth performance, with the overall model accuracy to up to 95.5%. It is shown in Fig. 7.

B. GCN: Standalone vs. in HYBERT-X

The GCN applied solo, sought to enhance the accuracy by focusing on the structural and relational features of the said text data, reaching 87%. It used a 2-layer architecture with a hidden dimension size of 64 and a learning rate of 0.01. Within the hybrid system, GCN contributed relational context by modelling word and feature-level graph relationships. While its individual output could not be quantified within the ensemble, its integration clearly supported HYBERT-X's ability to capture latent network patterns in cyberbullying detection. It is represented in Fig. 8.

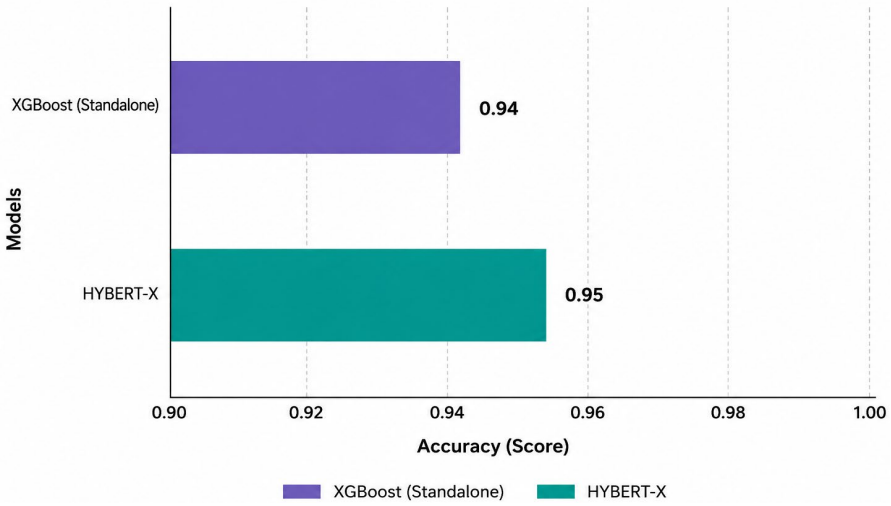

Fig. 7. Comparing accuracy between BERT and HYBERT-X.

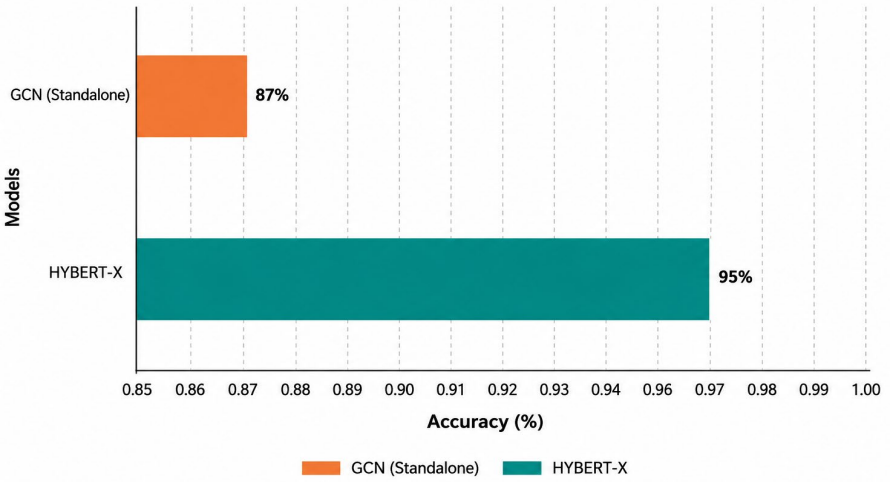

Fig. 8. Comparing accuracy between GCN and HYBERT-X.

C. *BiLSTM: Standalone vs. in HYBERT-X*

The BILSTM model has been trained with respect to temporal dependencies in sequential text and achieved an accuracy of 85%. It was built with 128 hidden units and a 0.5 dropout rate, and trained using the Adam optimizer. BiLSTM here performed the sequential patterns after fusion of BERT and GCN in the HYBERT-X framework. While not measured individually for its single influence effect in the hybrid, its contribution to refine sequential coherence has efficaciously benefited the final prediction layer and error reduction. It is shown in Fig. 9.

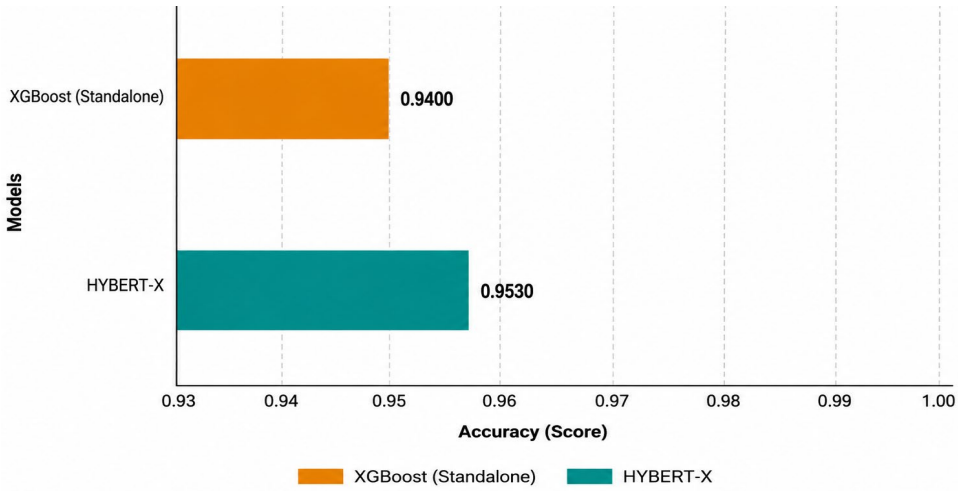

Fig. 9. Comparing accuracy between BiLSTM and HYBERT-X.

D. XGBoost: Standalone vs. in HYBERT-X

XGBoost alone performed exceptionally, achieving a performance of 94.43% accuracy and an F1-Score of 95.36% based on TF-IDF and engineered features. However, the performance of XGBoost was limited as it used a simple input representation. Here, XGBoost became the last classifier in the train, along with high dimensional features from BERT, GCN, and BiLSTM. This configuration was intentional and increased the generalization capability of XGBoost which finally raised the model accuracy up to 95.5%, a huge jump from its standalone performance. The comparison of accuracy between XGBoost standalone and HYBERT-X is shown in Fig. 10.

Although the current research is devoted to examining the role of each component in the framework of HYBERT-X, one should place it in perspective of its performance in comparison with such strong baseline models as RoBERTa, DistilBERT, and hybrid CNN-LSTM. Transformer-based models are only able to obtain semantic relationships, but HYBERT-X can achieve this and obtain structural and sequential representations. This multi-paradigm combination offers a richer feature space, which is likely to offer competitive/or better performance over single-paradigm or constrained hybrid methods.

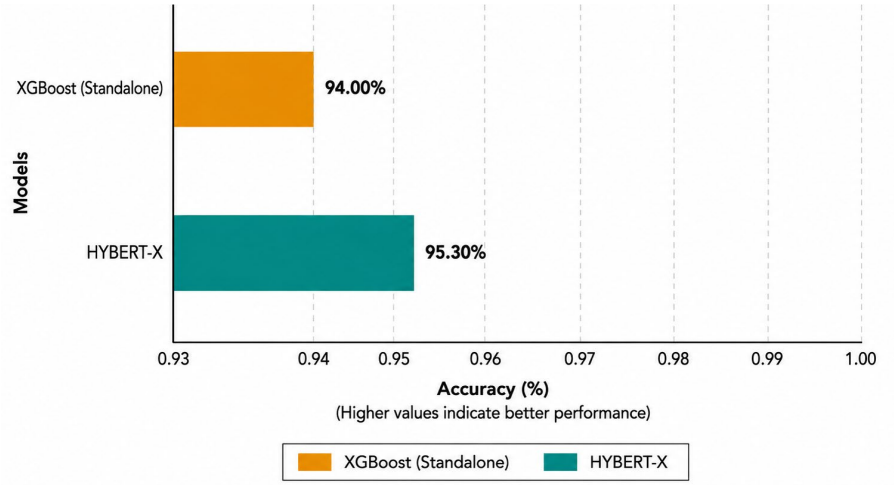

Fig. 10. Comparing accuracy between XGBOOST and HYBERT-X.

VI. Conclusion

This research proposes a novel hybrid deep learning model for cyberbullying and cybercrime content detection on social media, integrating BERT, GCN-style features, BiLSTM, and XGBoost. The shortcomings of the existing systems which, above all, are inflexible against the changing time of language, implicit threats, and complex contextual patterns, were revealed by systematic literature review and comparative analysis. The model will deal with these drawbacks by incorporating contextual, structural, and sequential learning functions in a single paradigm that will foster functionality in terms of precision and strength. Experimental findings established that the hybrid architecture that achieved 95.53% accuracy outperformed all classical machine learning frameworks and reliant deep neural networks through a vast number of evaluative metrics. The graph feature and ensemble booster further expand the ability of the model to detect linguistic subtlety and dynamic relationships and reduce false positives and false negatives of this real-world implementation. This does not only drive the content moderation of artificial intelligence but also provides a scalable and flexible solution to improving online safety. Though HYBERT-X yields superior accuracy and generalization, it may be the case that its output changes in a novel, unexplored domain and multilingual data sets. Moreover, the model at present only supports textual data, and does not have the possibility of multimodal analysis, comprised of images or audio. In addition to that, real-time deployment and scalability of inference will be optimized to achieve practical cybersecurity applications. Some of the directions of future analysis can be refining this work with multimodal representations like pictures, audio and using techniques in domain adaptation or using privacy-conscious learning methods like federated learning to promote ethical training are all potential subjects of interest.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Jayapriya J. conducted conceptualization, methodology, supervision, writing—original draft preparation, software, data curation, validation, formal analysis, visualization. Kavitha S. conducted supervision, project administration, resources, writing—reviewing and editing. Manimekala B. conducted investigation, data collection, experimental setup, validation. Nileem Kaveramma conducted data preprocessing, model training, visualization, writing—reviewing and editing. All authors had approved the final version.

ACKNOWLEDGMENT

The authors sincerely acknowledge CHRIST (Deemed to be University) for providing the institutional support and necessary funding for this research project.

REFERENCES

[1] N. V. S. S. R. Lakshmi and S. Dharan, "A comparative analysis of convolutional, sequential and their hybrid models in detecting cyber-attacks," in *Proc. 2024 1st International Conference on Trends in Engineering Systems and Technologies (ICTEST)*, 2024, pp. 1–8. doi: 10.1109/ictest60614.2024.10576092

[2] P. Maturure, A. Ali, and A. Gegov, "Hybrid machine learning model for phishing detection," in *Proc. 2024 IEEE 12th International Conference on Intelligent Systems (IS)*, 2024, pp. 1–7. doi: 10.1109/is61756.2024.10705257

[3] E. Daraghmi, S. Qadan, Y. Daraghmi, R. Yussuf, O. Cheikhrouhou, and M. Baz, "From text to insight: An integrated CNN-BiLSTM-GRU model for Arabic cyberbullying detection," *IEEE Access*, vol. 12, pp. 1–1, 2024. doi: 10.1109/access.2024.3431939

[4] A. R. Egigogo, I. Idris, M. Olalere, and A. O. Aderiike, "Evaluating deep learning models for website phishing attack detection: A comparative analysis," *Ceddi J. Inf. Syst. Technol.*, vol. 3, no. 2, pp. 19–29, 2024. doi: 10.56134/jst.v3i2.100

[5] R. Kimanzi, P. Kimanga, D. Cherori, and P. K. Gikunda, "Deep learning algorithms used in intrusion detection systems–A review," arXiv preprint, arXiv:2402.17020, 2024.

[6] S. Abarna, J. I. Sheeba, and S. Pradeep Devaneyan, "A novel ensemble model for identification and classification of cyber harassment on social media platform," *J. Intell. Fuzzy Syst.*, vol. 45, no. 1, pp. 13–36, 2023.

[7] M. T. Ahmed *et al*., "Cyberbullying detection based on hybrid ensemble method using deep learning technique in Bangla dataset," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 9, 2023.

[8] A. Akhter *et al*., "A robust hybrid machine learning model for Bengali cyber bullying detection in social media," *Natural Lang. Process. J.*, vol. 4, 10002, 2023.

[9] H. Al-Kaabi, F. A. Al-Ibraheemi, A. K. Jasim, M. Al-Rekabi, and J. Parchami, "Fusion-based hybrid model for SMS spam detection integrating local, sequential, and contextual features," *Research Square*, 2025. doi: 10.21203/rs.3.rs-5865706/v1

[10] A. Altayeva, R. Abdrakhmanov, A. Toktarova, and A. Tolep, "Cyberbullying detection on social networks using a hybrid deep learning architecture based on convolutional and recurrent models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 10, 2024. doi: 10.14569/ijacsa.2024.0151018

[11] J. P. Appadurai, L. S. Raveendran, R. Latha, E. Gangadevi, M. Shri, and S. Gite, "Cybersecurity using hybrid type model for classification through SCO optimization technique," in *Proc. 5th Int. Conf. Power, Control, Comput. Technol. (IC2PCT)*, 2024, pp. 1390–1396. doi: 10.1109/ic2pct60090.2024.10486307

[12] A. Baiganova, S. Toxanova, M. Yerekesheva, N. Nauryzova, Z. Zhumagalieva, and A. Tulendi, "Hybrid convolutional recurrent neural network for cyberbullying detection on textual data," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 5, 2024. doi: 10.14569/ijacsa.2024.0150584

[13] B. A. H. Murshed, Suresha, J. Abawajy, M. A. N. Saif, H. M. Abdulwahab, and F. A. Ghanem, "FAEO-ECNN: Cyberbullying detection in social media platforms using topic modelling and deep learning," *Multimedia Tools Appl.*, vol. 82, no. 30, pp. 46611–46650, 2023.

[14] S. Chen, J. Wang, and K. He, "Chinese cyberbullying detection using XLNet and deep Bi-LSTM hybrid model," *Information*, vol. 15, no. 2, 93, 2024. doi: 10.3390/info15020093

[15] S. P. Dandamudi, J. Sajja, and A. J. Khanna, "Advancing cybersecurity and data networking through machine learning-driven prediction models," *Int. J. Innov. Res. Comput. Sci. Technol.*, vol. 13, no. 1, pp. 26–33, 2025. doi: 10.55524/ijircst.2025.13.1.4

[16] J. A. Díaz-García and J. P. Carvalho, "A survey of textual cyber abuse detection using cutting-edge language models and large language models," arXiv preprint, arXiv:2501.05443, 2025.

[17] S. Dutta, M. Neog, and N. Baruah, "Towards safer social spaces: LSTM, Bi-LSTM and hybrid approach for cyberbullying detection in Assamese language on social networks," in *Proc. 2024 Second International Conference on Emerging Trends in Information Technology and Engineering (ICETITE)*, 2024, pp. 1–6. doi: 10.1109/ic-etite58242.2024.10493302

[18] F. B. Faruk Ananna, S. Barua, and M. N. Uddin, "Cyberbullying detection: An innovative multitask learning framework for analysing harmful memes," in *Proc. 2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, 2024, pp. 1–6. doi: 10.1109/compas60761.2024.10796421

[19] A. K. Gautam and A. Bansal, "Email-based cyberstalking detection on textual data using multi-model soft voting technique of machine learning approach," *J. Comput. Inf. Syst.*, vol. 63, no. 6, pp. 1362–1381, 2023.

[20] M. Gomathy and A. Vidhya, "Optimized hybrid model using machine learning to combat the prevalence of cybercrime," in *Proc. 7th Int. Conf. Electron., Commun. Aerosp. Technol. (ICECA)*, IEEE, 2023, pp. 739–746.

[21] W. Gong, C. S. Lee, S. Li, D. Adkison, L. Na, L. Wu, and X. Ye, "Cyber victimization in hybrid space: An analysis of employment scams using natural language processing and machine learning models," *Journal of Crime and Justice*, vol. 49, no. 1, pp. 132–153, 2025. doi: 10.1080/0735648x.2024.2448804

[22] T. N. Harshitha, M. Prabu, E. Suganya, S. Sountharrajan, D. P. Bavirisetti, N. Gadde, and L. S. Uppu, "ProTect: A hybrid deep learning model for proactive detection of cyberbullying on social media," *Front. Artif. Intell.*, vol. 7, 1269366, 2024. doi: 10.3389/frai.2024.1269366

[23] Y. W. Kassa, J. I. James, and E. G. Belay, "Cybercrime intention recognition: A systematic literature review," *Information*, vol. 15, no. 5, 263, 2024. doi: 10.3390/info15050263

[24] T. A. Khan *et al*., "Multi-source cyber intrusion detection using ensemble machine learning," *J. Comput. Sci.*, vol. 21, no. 1, pp. 111–123, 2025. doi: 10.3844/jcssp.2025.111.123

[25] D. E. O. Ogonji and W. Mwangi, "Hybrid phishing detecting with recommendation decision trees," *Int. J. Recent Technol. Eng.*, vol. 13, no. 2, 2024, doi: 10.35940/ijrte.b8120.13020724

[26] P. Paithane, "A systematic review of evaluating the efficiency of hybrid machine learning techniques in unmasking phishing URLs," *Indian Sci. J. Res. Eng. Manag.*, vol. 8, no. 3, pp. 1–5, 2024. doi: 10.55041/ijsrem29671

[27] V. L. Paruchuri and P. Rajesh, "CyberNet: A hybrid deep CNN with N-gram feature selection for cyberbullying detection in online social networks," *Evol. Intell.*, vol. 16, no. 6, pp. 1935–1949, 2023.

[28] A. Prabhu and N. R. Sunitha, "A literature review on machine learning methods used in intrusion detection system to detect cyber-attack," in *Proc. 2024 International Conference on Cybernation and Computation (CYBERCOM)*, 2024, pp. 94–97. doi: 10.1109/cybercom63683.2024.10803215

[29] M. Priyadharshini, H. J, A. F. Banu, B. Nithya, and V. Murugesh, "Advanced cyberbullying detection: A hybrid model integrated with naïve Bayes," in *Proc. 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, 2024, pp. 1–5.

[30] H. A. Rasool *et al*., "Evolutionary algorithm with graph neural network driven cyberbullying detection on low resource Asian languages," *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 2023.

[31] H. Saini, H. Mehra, R. Rani, G. Jaiswal, A. Sharma, and A. Dev, "Enhancing cyberbullying detection: A comparative study of ensemble CNN–SVM and BERT models," *Soc. Netw. Anal. Min.*, vol. 14, no. 1, 1, 2023.

[32] F. Shannaq, M. Shehab, A. Al-Shorman, M. Hammad, B. Hammo, and W. Alomari, "Exploring metaheuristic optimization algorithms in the context of textual cyberharassment: A systematic review," *Expert Syst.*, vol. 42, no. 2, 2025. doi: 10.1111/exsy.13826

[33] N. B. Sulikeri, "Billy-Buddy against cyberbullying," *Int. J. Sci. Technol. Eng.*, vol. 13, no. 1, pp. 767–777, 2025. doi: 10.22214/ijraset.2025.66441

[34] U. H. Sultana, "Hybrid model for cyberbullying detection from Bangla text in social media," M.Sc. thesis, Dept. Comput. Sci. Eng., Bangladesh Univ. Eng. Technol., Dhaka, Bangladesh, 2023.

[35] I. Tabassum and V. Nunavath, "A hybrid deep learning approach for multi-class cyberbullying classification using multi-modal social media data," *Appl. Sci.*, vol. 14, no. 24, 12007, 2024. doi: 10.3390/app142412007

[36] R. J. V. Geest, G. Cascavilla, J. Hulstijn, and N. Zannone, "The applicability of a hybrid framework for automated phishing

detection," *Comput. Secur.*, vol. 143, 103736, 2024. doi: 10.1016/j.cose.2024.103736

[37] Q. Yang, Y. Chen, Z. Xu, Y. Shang, S. Guo, and X. Zhang, "SCCD: A session-based dataset for Chinese cyberbullying detection," in *Proc. 31st International Conference on Computational Linguistics*, 2025, pp. 9533–9545.