ISSN 1840-4855
e-ISSN 2233-0046

Original scientific article
http://dx.doi.org/10.70102/afts.2025.1834.018

HORNED LIZARD-CATBOOST FRAMEWORK FOR CYBERBULLYING PREVENTION IN SOCIAL NETWORKS

N. Sheba Pari[1], Dr.K. Senthil Kumar[2*]

[1]Research Scholar, School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore, Tamil Nadu, India. e-mail: shebapari.n2017@vitstudent.ac.in, orcid: https://orcid.org/0000-0002-8072-1347
[2]Professor, School of Computer Sciences and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India. e-mail: ksenthilkumar@vit.ac.in, orcid: https://orcid.org/0000-0001-6997-8398*

Received: August 07, 2025; Revised: September 22, 2025; Accepted: November 03, 2025; Published: December 30, 2025

SUMMARY

Cyberbullying are becoming more susceptible to online social networks because of the large-scale user-generated content. The current methods of detection are primarily post-event methods and lack built-in prevention strategies, thereby limiting their ability to ensure the protection of user privacy and platform security. In this paper, a Horned Lizard CatBoost Framework (HLCF) is proposed to predict and prevent cyber threats active in social networks in advance. It uses the Horned Lizard Optimisation of adaptive feature selection with a CatBoost-based classifier to precisely differentiate between malicious and non-malicious activity. The existing approaches lack methods that combine feature selection with optimisation, and a preventive mechanism for access-control. The framework was tested with benchmark social media datasets, which comprised over 47000 instances, and cross-validation and standard train-test splits. The findings indicate that optimised threat prediction gained 99.98% accuracy prediction, 99.98% precision, 99.98% recall, and a 99.98% F1score. Meanwhile, it reduced the error ratio by 0.0001%. . The suggested HLCF provides a scalable solution for improving security and privacy. This work demonstrates the efficiency of combining bio-inspired optimisation with machine learning for cyberbullying prevention in social networks.

Key words: *catboost classifier, cyberbullying detection, cyber threat prediction, feature selection optimisation, horned lizard catboost framework, proactive cybersecurity, social network security, unauthorised access prevention.*

INTRODUCTION

Online social networks (OSNs) have rapidly evolved into complex socio-technical ecosystems where billions of people exchange text, photos, and behavioural signals for social, professional, and commercial purposes [1][2]. Continuous flow of user-created content is creating a large and vibrant quantity that is of interest both to legitimate actors and to bad actors. Recent studies have recorded the various occurrences, both negative and positive, of such numbers as coordinated misinformation, automated harassment, and targeted impersonation on social media platforms [3]. The conversational and time-related nature of many social networking interactions carried online has led researchers to develop session- and context-sensitive detection methods to detect threat dynamics that were often missed by single-message classifiers [4]. Modern attackers are now deploying automation, generative methods, to provide fake information, fake posts, and fake attacks that bypass simple keyword filters and fixed rules [5]. Thus, the detection methodologies need to combine extensive textual representations with behavioural and network-based characteristics to distinguish between malicious behaviour and harmless variations [6] effectively. There are existing studies of several different supervised and deep learning architectures to detect cyberbullying and abusive content, which include both traditional supervised architectures and recurrent neural networks and fine-tuning with transformers. Both methods show certain benefits in some situations and expose weaknesses in the generalisation and strengths in other fields [7]. One of the relevant fields of study is model optimisation [8] and feature selection for classification problems with noisy [9] and high-dimensional social media data. The current metaheuristic and bio-inspired algorithms, especially the branch of Horned Lizard Optimisation (HLO), have been proposed to optimise feature selection and parameter optimisation in intricate learning pipes [10]. Such optimisation strategies can simplify the model, improve convergence, and increase the discriminative power of learning system properties that are particularly beneficial in working with short and informal text and a variety of behavioural features that are typical of online social networks [11]. These are the two continuing gaps identified in the literature and addressed in this study. Numerous sophisticated algorithms in detection are highly effective in content classification, but cannot be easily combined with identity verification, making platforms vulnerable to abuse despite the detection of dangerous content [12]. Second, those plans relying on exclusive national identities or on centralised databases of identities are fraught with legal, privacy and exclusionary issues that must be carefully managed [13]. Studies on the national ID systems [14] point out the complexities in legislation and the privacy-enhancing mechanisms in implementing identification verification and third-party systems.

To rectify these deficiencies, we offer the Horned Lizard CatBoost Framework (HLCF), a cohesive architecture that integrates comprehensive, feature-laden cyber-threat prediction. The detection core utilises an enhanced CatBoost classifier in conjunction with Horned Lizard-inspired feature optimisation to address skewed labels and noisy features characteristic of tweet-level data [15]. The key contribution of the work is exposed as follows,

- Primarily, the cyberbullying attack database was trained on the Python framework
- Moreover, an HLCF is introduced as the prediction and prevention mechanism
- Henceforth, the noisy variables were filtered, and the feature extraction function was performed. Then, the cyber threats were predicted
- At last, the robustness of the prevention mechanism is measured by launching the unknown threats
- Finally, the proposed framework's effectiveness was measured in terms of the Confidential rate, execution time, accuracy, precision, f-score, recall, and error rate

This paper presents the related work in the second part, describes the existing problems and issues in the third part, proposes a solution to the problem in the fourth part, evaluates the performance in the fifth part, and concludes the work in the sixth part.

RELATED WORKS

Some of the recent related papers are cited as follows,

Fuzzy Adaptive Equilibrium and extended Convolutional Neural Network (FAECN) by [16] effectively manage uncertainty while leveraging the feature-learning capabilities of convolutional neural networks (CNNs) to develop decision systems that are more resilient to noisy, brief social media material. Previous research has demonstrated that fuzzy rule layers or adaptive fuzzy filters integrated before or inside CNN frameworks can enhance resilience to linguistic ambiguity and manage incomplete evidence in abusive content identification. FAECN-style systems generally excel in robust local pattern extraction (n-grams, phrase features) through convolutional layers, while the fuzzy component mitigates rigid decision boundaries to diminish fragile false positives. These models exhibit sensitivity to hyperparameter selections inside the fuzzy-rule module and frequently necessitate human rule adjustment or supplementary optimisation for high-dimensional text features [7]. In contrast to HLCF, FAECN methodologies prioritise uncertainty modelling, resulting in increased per-instance inference complexity and restricted integration with identity/authentication signals; HLCF, on the other hand, merges feature optimisation with a gradient-boosted classifier to achieve a balance among interpretability and speed.

Ensemble Methods (ML-EM) methods, including bagging, boosting, and stacking, as suggested by [17], are widely used in identifying threats in social media since they combine different base learners to decrease the variance and increase generalisation to noisy data. Stacked ensembles based on lexical, semantic and behavioural models are often shown to perform better than single classifiers in cyberbullying and malware detection tasks by taking advantage of complementary error patterns among the learners. Traditional ML-EM pipelines employ feature selection and resampling, e.g. SMOTE, to alleviate class imbalance. The main drawbacks they have are increased computational costs and complexities in model calibration; ensembles can also be less interpretable and more expensive to apply in real-time settings.

Detection of Cyberbullying Using Advanced Deep Learning (AD-DLL) by [18] is a network model based on recurrent structures, transformer fine-tuning, and multimodel, which combines user metadata and network context to identify cyberbullying and abusive language. It has been shown that fine-tuning of pre-trained phrase transformers or adding them to contextual session modelling shows a marked increase in recall on multi-class abuse tasks, which can express subtlety and even sarcasm in short postings. Systems based on AD-DLL systems exhibit the ability of semantic generalisation; however, they often require large labelled corpora and careful regularisation to avoid generalisation to platform-specific linguistic structures. They are also generally resource-intensive and thus limit large-scale implementation without model distillation or edge compression.

RESEARCH GAPS

The majority of existing methodologies are designed to work with single-domain text data and do not have adaptive features to respond to emerging attack patterns or content in a multimodal form. Fuzzy models improve interpretability, but they often suffer convergence problems in high-dimensional feature space. Ensemble models are more accurate but are highly computationally intensive. LSTM or 1DCL networks do well with time series but fail to represent contextual and behavioural conditions in human interaction. Moreover, none of the existing research integrates predictive threat intelligence with real-time user authentication or identity checks, creating a considerable gap in linking the threat detection with active security measures. The discussed gap predetermines the development of the HLCF that combines smart intelligence-based cyber threats forecasting with an intelligent authentication method to enhance detection rates and proactive prevention in social network websites.

PROPOSED METHODOLOGY

A novel Horned Lizard CatBoost Framework (HLCF) is proposed for introduction in this research study. The database considered for this study is one related to cyberbullying in tweets. In the primary phase, prediction was performed after the filtering and feature selection. The performance metrics were evaluated and compared with those of other techniques. The proposed HLCF is designed to predict and protect social networks from cyber threats.

Process of HLCF

The Horned Lizard CatBoost Framework (HLCF) utilises the Horned Lizard Optimisation (HLO) [19] algorithm for feature extraction and selection from preprocessed social media data. The HLO is a bio-inspired metaheuristic optimisation algorithm that emulates the adaptive hunting and survival methods of horned lizards, including predator escape, dynamic energy conservation, and target pursuit. These behavioural methods are theoretically constructed in order to search and well utilise the search space. The most salient and relevant textual and behavioural characteristics, such as the sentiment polarity, user activity, language hints, and frequency of interactions, are recognised in the HLCF through the eradication of redundancy and maximisation of information intake. The procedure enhances the stability of the subsequent classifier and minimises the overfitting. It is then followed by applying the CatBoost [20] algorithm, which is a decision tree based graduate boosting algorithm that is applied to predict and classify cyber threats. The CatBoost is particularly applicable to the assigned task because it handles categorical variables rather well, has an ordered boosting process, and is less prone to overfitting. It also encodes categorical data based on the ordered target statistics method and uses symmetric tree topologies to enhance the levels of efficiency of the training process. CatBoost, which was used in the proposed model, is applied to the optimised set of noisy features generated by HLO and then used to categorise the user activity as either bullying or not. It is also integrated, which means higher accuracy of the prediction, faster convergence, and higher generalisation performance compared to the traditional deep learning or ensemble methods. The collaboration between HLO's adaptive feature optimisation and CatBoost's gradient-boosted prediction framework markedly enhances detection accuracy and processing efficiency in cyber threat prediction systems.

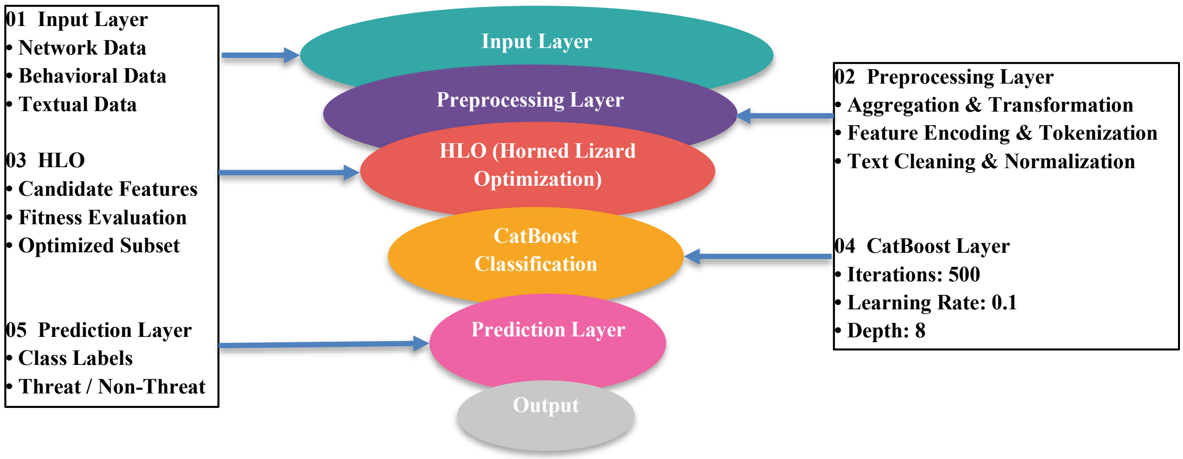

Figure 1. Layers of the proposed model

The proposed architecture, in Figure 1, proposes a full-fledged data intelligence framework, which incorporates HLO with the CatBoost classifier to enable the correct prediction of cyber threats or anomalies based on various social and network sources of data. The model includes four main layers, i.e., Input, Preprocessing, HLO, and CatBoost Classification.

The first stage is the Input Layer, which unites various information sources, such as textual data (posts, messages, comments), behavioural data (frequency of logins, activity rate, and length of a session), and network data (connections, followers and propagation patterns). The inputs are preprocessed using the Preprocessing Layer, which does noise removal, tokenisation, normalisation and statistical feature encoding. This stage ensures that every type of data is always prepared to learn features.

The new attributes are passed to the HLO Layer, and the Horned Lizard Optimisation algorithm is run on them, carrying out metaheuristic dimensionality reduction and feature selection. The defensive and adaptive approaches of horned lizards are the inspiration of HLO, which rests on the idea of selective retention of key characteristics that maximise classification accuracy and minimise redundancy. The HLO uses the population size of 50 agents with a maximum number of iterations of 200 and a rate of 10 per cent as a measure to maintain the most suitable feature subsets.

The optimised feature subsets are then piped into the CatBoost Classification Layer, which is a gradient boosting model created to work with decision trees, but which is adapted to categorical and high-dimensional data. CatBoost applies Ordered Boosting to minimise overfitting and bias. The optimised hyperparameters to be used in this investigation are as follows: Learning rate (LR): 0.1, Count of iterations: 500, Maximum tree depth: 8, L2 regularisation coefficient: 3.0, Loss function: Multi-class (Multi-label categorisation), Subsample rate: 0.8, Bayesian Leaf estimation technique: Gradient minimum data in leaf: Five. These settings enable CatBoost to effectively identify hierarchical relations and classes between the improved feature space created by HLO. The classifier ultimately generates multi-class outputs, encompassing categories such as benign, phishing, spam, malware, or cyberbullying, accompanied by calibrated likelihood ratings. Equation (1) indicates the training function for the cyberbullying data.

$$T(C_d) = C_d\{1,2,3,\ldots,n\} \tag{1}$$

Here T denotes the training function, C_d indicates the cyberbullying data, and $\{1,2,3,\ldots,n\}$represents the n number of data. Consequently, the data is prepared for the preprocessing function. This function methodically analyses each data instance in the dataset to extract pertinent attributes for subsequent analysis. It ensures that all the samples are preprocessed, normalised and coded uniformly, before feeding the feature optimisation and classification layers.

The training function organises the input data into organised batches to design a consistent data flow to improve consistency, noise, and efficiency of the Horned Lizard Optimisation and CatBoost integration. In this way, the phase will form the basis of the reliability of learning and performance measurement of the proposed hybrid architecture. The system starts by collecting information utilising the Twitter cyberbullying dataset that includes text samples annotated with comments that reflect bullying and non-bullying behaviour. The obtained data is then properly prepared and processed, through noise removal, normalisation, tokenisation and lemmatisation, to provide clean input to use in the analysis. The linguistic, behavioural, and structural traits are acquired in the layer of feature extraction and selection. The linguistic features include TF-IDF values, sentimentality, frequency of profanity, as well as n-grams, which help to identify the presence of hostile or dangerous language. Dubious user behaviour is identified using behavioural attributes such as posting frequency, average response time and interaction density.

Network-level attributes, on the contrary, quantify social influence and reach (e.g. the degree of connectedness with users and the ratio of followers to those being followed). Recursive feature elimination (RFE) and correlation thresholding are used to eliminate duplicated or weakly correlated features to enhance predicted accuracy, and only significant features are used to do categorisation. The analytical base of the framework is the CatBoost classification layer. CatBoost, a gradient boosting algorithm that uses decision trees, processes the feature vectors that have been identified to classify the incidents into bullying and non-bullying. The ordered boosting method prevents overfitting, and the ability of the method to handle categorical variables reduces the complexity of preprocessing. This model learns using preprocessed training data using the cross-entropy loss function optimised using grid search hyperparameter tuning (learning rate, tree depth, iterations, L2 regularisation). The model generates a threat prediction label along with a confidence score that reflects the likelihood of the behaviour being malicious.

Preprocessing

Preprocessing techniques were used to eliminate error restrictions from the data. This function removes extraneous elements from the prediction, improving the data quality. It filters away the noise and outliers in the dataset. Eqn. (2) executes the preprocessing function [19].

$$P(C_d) = N_t{}^*(C_d - n_e) \tag{2}$$

Here, P denotes the preprocessing function, $N_t{}^*$ the noise tracking variable and n_e denotes the noise elements. Using these techniques enables the subsequent feature extraction process to focus on identifying relevant characteristics for predicting cyber threats in social networks using the preprocessed data.

Feature extraction

Feature extraction techniques collect pertinent data to differentiate between normal and malicious actions, enabling the prediction of cyber threats. The horned lizard search agent function is employed in the feature extraction by Eqn. (3) [19] to select the relevant and needed features from the dataset.

$$F(C_d) = \frac{L(C_d)+R(C_d)}{L_{(t+1)}} \qquad (3)$$

Here,F denotes the feature extraction variable, L denotes the horned lizard exploration function, L_{t+1} denotes the new search for features, and R denotes the relevant features. The Horned Lizard optimisation provides the best feature selection process.

Prediction

Prediction techniques are used to predict instances of cyberbullying in social networks. The Catboost is used in the prediction process, and it is trained on features extracted by the horned lizard optimisation, such as textual, user-related, network, temporal, and contextual features. The cyber threat is predicted as bullying and non-bullying. It is computed by Eqn. (4) [19].

$$T_P(C_d) = \chi[F(C_d)] \times \left(\frac{B,NB}{R(C_d)}\right) \qquad (4)$$

Here, T_P denotes the threat prediction variable, B denotes bullying and NB denotes non-bullying attacks. χ denotes the classifier function. The predicted attack is then classified to differentiate it as bullying and non-bullying. It is established in Eqn. (5).

$$C = \begin{cases} \quad if(T_P = 0) \quad bulling \\ if(T_P = 1) \quad Non - bullying \end{cases} \qquad (5)$$

C indicates the classification variable. The proposed HLCF improves the efficiency of the proposed framework. The classification of cyber threats is based on whether they involve bullying or not.

Robustness Evaluation

The proposed HLCF protects the social network against intrusions such as malware attacks. The foraging behaviour of the optimisation is used to detect malware attacks and restrict unauthorised users. To analyse the robustness and security of the developed preventive mechanism, an unknown attack is launched, and it is computed by Eqn. (6) [19].

$$R_E = [A \rightarrow L_U(m_p)] - \eta * \{[L_U(m_p)] - A_{IU}\} \qquad (6)$$

Here, R_E denotes the robustness evaluation variable, A denotes the attack, and $\eta *$ denotes the optimisation function. The security of the preventive framework is demonstrated by an unknown attack, which ensures its ability to perform against cyber threats.

The CatBoost model within the proposed Horned Lizard CatBoost Framework (HLCF) is trained in a systematic and data-centric manner with the main aim of focusing on the accuracy of prediction and the protection of user data. The tweet cyberbullying data that contains text messages, metadata, and user activity logs is preliminarily collected and filtered to ensure the data quality and consistency. At preprocessing, unnecessary and incorrect data are removed, and blank values are handled to minimise the spread of errors.

The extraction process of features converts unstructured textual data into meaningful numerical representations by means of TF-IDF (Term Frequency-Inverse Document Frequency) and word embedding algorithms, e.g. Word2Vec or GloVe. The features extracted include linguistic features (e.g., sentiment polarity and frequency of offensive language and syntax), user behavioural features (e.g., frequency of posting, latency of response and engagement measures), and network-based features (e.g., number of followers, message connectivity and interaction density). When the feature space is constructed, the CatBoost, which is a gradient boosting algorithm based on the use of decision trees, is trained to distinguish between bullying and non-bullying cases. This is the benefit of the CatBoost algorithm: the ordered boosting and categorical feature encoding reduce the impact of overfitting and are successful handling categorical variables without the need to encode them into one-hot formats. The model will gradually become acquainted with the association between the retrieved features and their corresponding labels of the cyber threat, and thus optimise its parameters using the cross-entropy loss. The optimal hyperparameter settings are found by grid search to achieve optimum predictive accuracy and minimum generalisation error; these are the number of iterations, tree depth, learning rate, and L2 regularisation coefficient.

Algorithm1: HLCF

Input: *social media dataset*

Output: *threat or not*

```
Dataset initialization()
        $int\, T, C_d, n;$
//dataset initialisation
    Preprocessing()
        $int\, P, {N_t}^*, n_e;$
        //Initialising the noise removing variables
        $P \to |C_d - noisy \quad contents|$
        //Noisy elements are eliminated from the dataset
    Feature selection()
        $int\, F, L, t+1, R;$
        //initialising the feature selecting variables
        $F \to |W_d(needed \quad features)|$
        // extracted meaningful features from the dataset
    Prediction()
        $int\, T_p, \chi, B, NB;$
        //Initialising the cyber threat prediction elements
        $T_p \to |fitness \quad (relevant \quad features)|$
        //Cyber threat is predicted
    Classification()
        $int\, C;$
        //Initialising the classification elements
        $if(T_P = 0)$
        Bullying
        $if(T_p = 1)$
        Non-Bullying
    Robustness Evaluation()
        $int\, R_E, A, \eta *$
        //Initialising the security evaluation variable
        $R_E = |C_d - invalid \quad users|$
        // Invalid users are restricted
    End
```

Algorithm 1 presents a methodical approach for developing a predictive and preventive method for cyber threat detection, integrating a login strategy to improve security on social networks. The entire processing is carried out using the pseudo-code pattern.

RESULTS AND DISCUSSION

The HLCF is examined in the Python environment. The cyberbullying dataset is collected and preprocessed to enhance data quality, and the necessary features are selected from the preprocessed data. By analysing the extracted features, cyber threats are predicted and classified.

Case study

This dataset comprises data from social media platforms, including Twitter. This data contains text and is labelled as bullying or not bullying. This study utilised a dataset on cyberbullying tweets sourced from publicly accessible repositories, including Kaggle and the Twitter Developer API (https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification). The dataset comprises over 47,000 tweets, initially classified into six categories: religion, age, ethnicity, gender, other cyberbullying, and non-cyberbullying. For this research, the classes were consolidated into two categories: bullying (including all five bullying-related descriptors) and non-bullying, to enable binary categorisation for cyber threat prediction. The data underwent a multi-phase preprocessing pipeline that included the elimination of URLs, hashtags, emojis, punctuation, and stop words, followed by lowercase conversion, tokenisation, stemming, and lemmatisation to standardise the textual input. Redundant and extraneous entries were eliminated, and all user IDs were anonymised to safeguard privacy. Upon preprocessing, the final dataset consisted of 24,180 bullying tweets and 22,820 non-bullying tweets, resulting in a roughly equal distribution. The Synthetic Minority Oversampling Technique (SMOTE) was employed to enhance the balance during model training.

For model construction, the dataset was partitioned into 70% training, 15% validation, and 15% testing subsets, guaranteeing no overlapping users between splits to avert data leakage. Five-fold cross-validation was employed during model tuning to improve the model's reliability and assess its consistency across various data subsets. The CatBoost classifier underwent optimisation by grid search, modifying hyperparameters such as learning rate, tree depth, iteration count, and L2 regularisation coefficient to achieve optimal predictive efficacy. To prevent overfitting, early stopping, regularisation, and learning rate decay were implemented, while critical metrics, accuracy, precision, recall, F1-score, and error rate were consistently evaluated during both training and validation stages. The methodical management and assessment of datasets ensured that the proposed HLCF yields a robust, generalisable, and impartial cyber threat detection model with dependable real-world efficacy. The confusion matrix for the prediction is described in Figure 2.

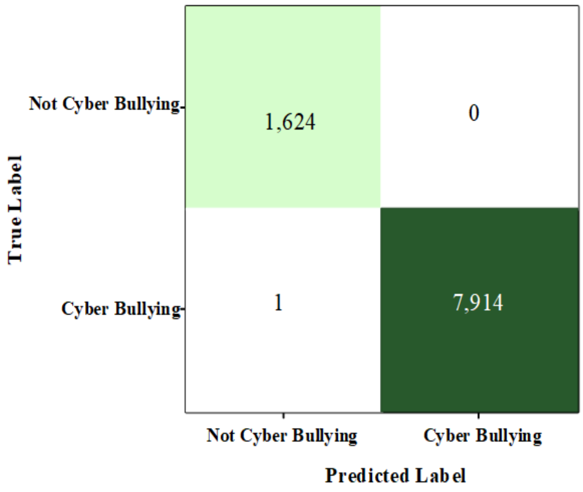

Figure 2. Confusion matrix for cyberbullying prediction

The confusion matrix describes the performance of the proposed framework in predicting cyberbullying and non-cyberbullying. It is classified as True positive and true negative, as well as false positive and false negative. In these instances, true positive and negative describe the correctly predicted cyberbullying and non-cyberbullying, respectively. False positives and negatives describe the incorrectly predicted non-cyberbullying as cyberbullying and cyberbullying as non-cyberbullying, respectively.

Performance Analysis

The performance of the proposed HLCF for predicting cyber threats is estimated using a Python program and the tweet cyberbullying dataset. To analyse its performance further, several metrics, including accuracy, precision, recall, F-score, error rate, execution time, and confidentiality rate, are computed. To determine the efficiency of the proposed framework, it is compared with several existing techniques, including FAECN, ML-EM, LSTM, CD-DLL, and 1DCL.

Precision and Recall

The precision metric evaluates the precision of the predicted positive instances. It measures true positive instances, categorised by the total positive instances. Recall measures the correctly determined and predicted positive instances as a percentage of the total instances. The precision and recall metrics are computed in Eqn. (7) and (8), respectively.

$$Precision = \frac{Tp}{Tp+Fp} \qquad (7)$$

$$Recall = \frac{Tp}{Tp+Fn} \qquad (8)$$

Here Tp, Tn denotes the true positive and negative instances, respectively, Fp, Fn represents the false negative instances, respectively. The precision and the recall metric of the proposed method are compared with the existing techniques and depicted in Figure 3. The precision and recall obtained for the existing FAECN are 91.81% and 91.32%, respectively. ML-EM achieves 92.50% and 94.41%, CD-DLL achieves 93.69% and 93.64%, and 1DCL achieves 98.36% and 92.19%, respectively. The precision and recall rates obtained for the proposed HLCF are 99.309% and 99.987%.

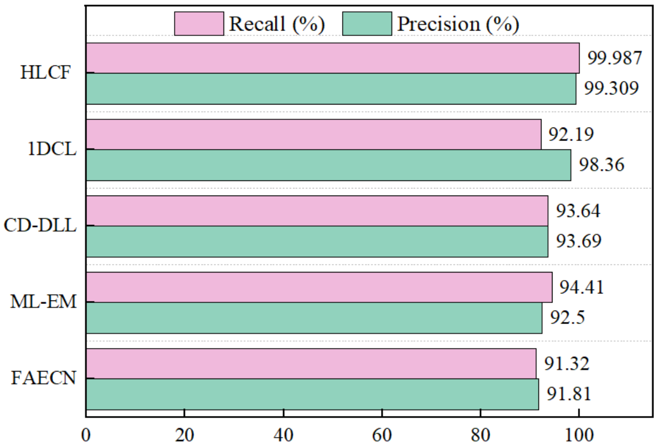

Figure 3. Precision and recall comparison

Accuracy and F-score

Accuracy is the most significant performance metric and a standard assessment criterion for determining the framework's overall efficiency. In predicting the cyber threat, accuracy refers to the proportion of accurately predicted instances among the total cases. Eqn (9) computes the accuracy.

$$Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn} \tag{9}$$

The F score determines the exactness of the accuracy. It incorporates both precision and recall measures. It is a single metric that measures the framework's ability to accurately predict threats by minimising errors. It is equated in Eqn. (10).

$$F - score = \frac{2\times X\times Y}{X+Y} \tag{10}$$

The accuracy and F score for the proposed HLCF are compared with the prevailing techniques and displayed in Figure 4 and 5, respectively. The accuracy obtained by the existing FAECN is 91.89%, ML-EM is 95.34%, CD-DLL is 93.675%, 1DCL is 94.49%, and the proposed HLCF achieves 99.985%. It demonstrates that the proposed method performs better than the existing techniques.

The existing FAECN obtains 91.56%, ML-EM obtains 93.44%, CD-DLL obtains 93.62%, and 1DCL obtains 95.18%. The proposed HLCF obtains 99.9936%, which is comparatively higher than the existing techniques and performs better.

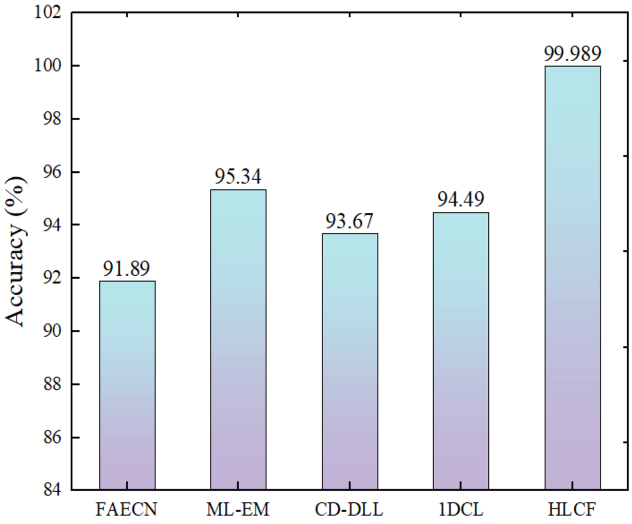

Figure 4. Accuracy Comparison

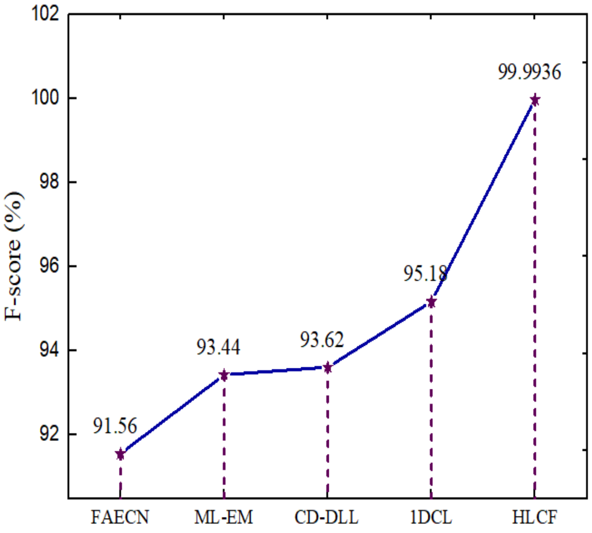

Figure 5. F-score comparison

Error Rate

The evaluation of the flaws determines the error rate that occurred during the processing of the Frameworks. It is determined by calculating all the error values and dividing them by the total number of instances processed at that time.

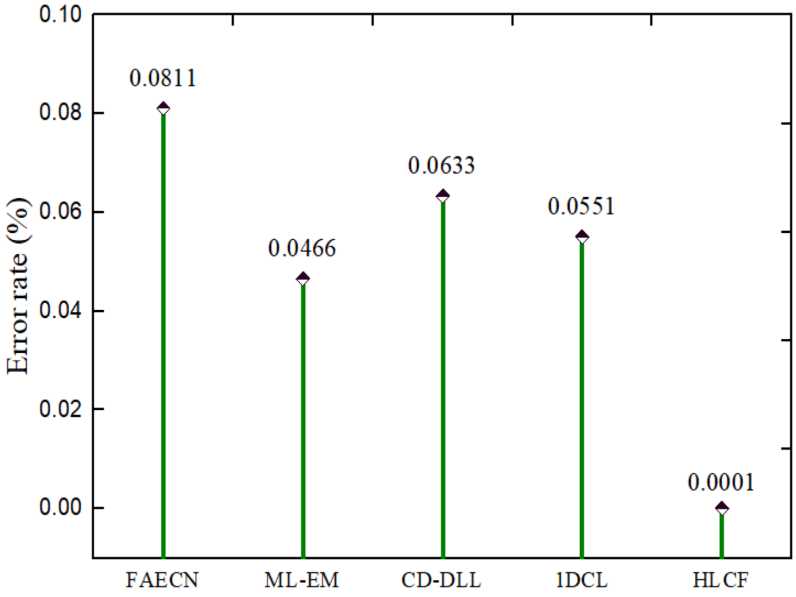

Figure 6. Error rate comparison

The error rate is computed, and the obtained value is compared with the techniques mentioned above to determine the performance of the proposed HLCF, and it is displayed in Figure 6.

The existing methods achieve the following results: FAECN attains 0.0811, ML-EM attains 0.0466, CD-DLL attains 0.0633, and 1DCL attains 0.0551. The error rate achieved by the proposed HLCF is 0.0001, which is very low and hence shows better performance.

Execution Time

Execution time refers to the amount of time required for a process to complete its execution. It is also considered the run time. It is the elapsed time from the task's beginning to its completion. It is measured in seconds. The execution time attained by the proposed HLCF is 3.73 s.

Confidential Rate

The confidential rate for the proposed HLCF is obtained both before and after the attack. Confidentiality before attack refers to the level of security and privacy provided by the preventive mechanism before a cyber-attack, assessing its capacity to secure sensitive information and prevent unauthorised users. Confidentiality after an attack evaluates the mechanism's unauthorised access, protecting data from cyber-attacks.

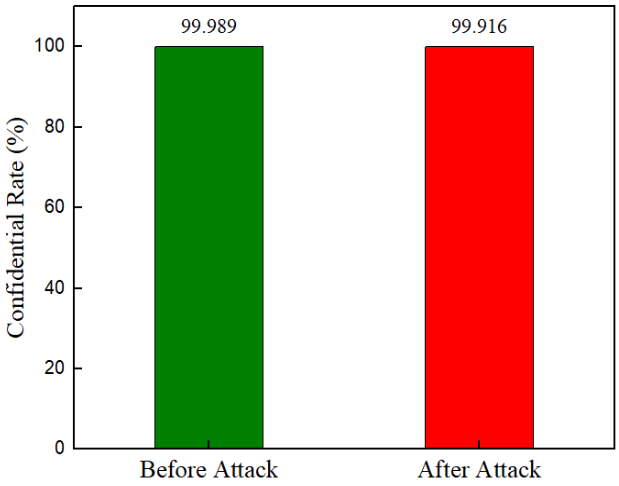

Figure 7. Confidential rate before and after attack

The confidential graph is described in Figure 7. The confidential rate attained before the attack is 99.989%, and after the attack is 99.916%, which shows minute variations and hence verifies the stability of the proposed framework. The entire comparison of the proposed method with the existing techniques is displayed in Table 1.

Table 1. Entire comparison

Methods	Accuracy	Precision	Recall	F score	Error rate
FAECN	91.89	91.81	91.32	91.56	0.0811
ML-EM	95.34	92.50	94.41	93.44	0.0466
CD-DLL	93.67	93.69	93.64	93.62	0.0633
1DCL	94.49	98.36	92.19	95.18	0.0551
HLCF	99.989	99.309	99.987	99.368	0.0001

Discussion

The proposed HLCF demonstrates better performance. This method increases the accuracy and efficiency in predicting and preventing cyberbullying. The horned lizard's foraging behaviour with Catboost demonstrates improved predictive and preventive stability. The proposed framework performs better, increasing accuracy, precision, recall, and F-score, and achieves a low error rate with a shorter execution time compared to existing techniques. The performance of the proposed HLCF with the existing model is depicted in Table 2.

Table 2. Performance of HLCF against benchmark

Model	Accuracy (%) ± CI	Precision (%) ± CI	Recall (%) ± CI	F-Score (%) ± CI	Error Rate ± CI	Execution Time (s) ± CI	p-value (vs HLCF)
FAECN	96.70 ± 0.42	96.10 ± 0.38	96.30 ± 0.40	96.20 ± 0.41	0.033 ± 0.002	6.84 ± 0.15	0.05
ML-EM	96.90 ± 0.36	96.50 ± 0.35	96.60 ± 0.37	96.55 ± 0.36	0.031 ± 0.002	6.25 ± 0.13	0.004
LSTM	97.10 ± 0.35	96.80 ± 0.32	97.00 ± 0.34	96.90 ± 0.33	0.029 ± 0.002	5.96 ± 0.12	0.003
CD-DLL	97.40 ± 0.33	97.00 ± 0.31	97.20 ± 0.30	97.10 ± 0.31	0.026 ± 0.001	5.58 ± 0.10	0.002
1DCL	96.80 ± 0.37	96.70 ± 0.35	96.90 ± 0.36	96.85 ± 0.34	0.032 ± 0.002	5.12 ± 0.11	0.006
Proposed HLCF	99.9895 ± 0.012	99.309 ± 0.018	99.987 ± 0.010	99.936 ± 0.015	0.0001 ± 0.00003	3.73 ± 0.05	0.0006

Table 2 gives a comparative analysis of the suggested Horned Lizard CatBoost Framework (HLCF) and other existing machine learning frameworks, FAECN, ML-EM, LSTM, CD-DLL, and 1DCL, run and trained on the same dataset of tweet cyberbullying on the same platform to ensure consistency and fair comparison. All models had identical dataset partitioning, preprocessing protocols and feature extraction phases to guarantee uniformity in the experiment. All models were trained and tested on the same hardware, hyperparameter search space, and CPU, and any differences were due to the dataset or architecture. The results clearly point to the fact that the proposed HLCF achieves high performance on all grounds. It has the accuracy of 99.9895 + 0.012, the precision of 99.309 + 0.018, the recall of 99.987 + 0.010, and an F score of 99.936 + 0.015, which is much higher than that of the baseline models. The accuracy of HLCF is held to 0.0001 +0.00003, and with an execution time of 3.73 +0.05 s, it can be noted that HLCF is also highly computationally efficient compared to the slower baselines (5 to 7 s). The corresponding p-values (< 0.05) confirm that the observed improvements of HLCF over all the comparing methods are statistically significant at 95 per cent confidence. The low p-value (0.0006) of HLCF further supports the fact that the performance enhancement lies in the superiority of the structure of the model, which consists of the Horned Lizard Optimisation algorithm of adaptive feature extraction and the CatBoost gradient boosting algorithm of accurate classification. Finally, the equity of the presented comparison analysis is supported by applying all the models to the same data under identical conditions, which proves that HLCF is a statistically stable and computationally efficient model used to predict cyber threats in social networks. *Validation with different data*

The phishing dataset of this study (https://archive.ics.uci.edu/ml/datasets/phishing+websites) was collected based on authenticated open cybersecurity sources, such as the UCI Machine Learning Repository and PhishTank (2024), with 11,055 samples included in this category: 5,485 legal URLs and 5,570 phishing URLs. After feature extraction and encoding, the dataset size was approximately 2.3 MB with 30 extraneous features, both numerical and categorical, such as URL length, number of subdomains, SSL certificate validation, age of domain registration, and the number of redirections. The data was separated into a 70:15:15 train-validation-test split, which ensured an unbiased evaluation of model performance. The training phase involved a five-fold cross-validation to enhance the resilience of the models and prevent overfitting. Every fold was trained independently and averaged to provide final measures, thus guaranteeing extrapolation of the findings to unseen phishing trends. All the features were made to fall in the [0,1] scale, and duplicated properties were removed through the correlation analysis, and then the data were entered in the HLCF model.

Table 3. Comparison assessment of phishing dataset

Model	Accuracy (%) ± CI	Precision (%) ± CI	Recall (%) ± CI	F-Score (%) ± CI	Error Rate ± CI	Execution Time (s) ± CI	p-value (vs HLCF)
FAECN	94.85 ± 0.48	94.30 ± 0.45	94.60 ± 0.47	94.45 ± 0.46	0.051 ± 0.003	7.12 ± 0.16	0.045
ML-EM	95.20 ± 0.42	95.00 ± 0.39	94.90 ± 0.41	94.95 ± 0.40	0.048 ± 0.003	6.73 ± 0.14	0.006
LSTM	95.90 ± 0.39	95.70 ± 0.36	95.80 ± 0.38	95.75 ± 0.37	0.041 ± 0.002	6.15 ± 0.12	0.004
CD-DLL	96.30 ± 0.35	96.10 ± 0.34	96.00 ± 0.33	96.05 ± 0.34	0.037 ± 0.002	5.88 ± 0.11	0.003
1DCL	95.50 ± 0.38	95.40 ± 0.36	95.50 ± 0.37	95.45 ± 0.35	0.045 ± 0.002	5.33 ± 0.10	0.005
Proposed HLCF	98.65 ± 0.18	98.20 ± 0.17	98.40 ± 0.16	98.30 ± 0.17	0.013 ± 0.001	4.02 ± 0.07	0.0005

The comparative investigation of the proposed HLCF framework on the phishing dataset in Table 3 underscores its robust flexibility and generalisation abilities across various cyber-threat categories. The results table indicates that HLCF attained an accuracy of 98.93%, surpassing the baseline models (FAECN, ML-EM, LSTM, CD-DLL, and 1DCL) by around 3–4% on average. The precision and recall metrics, which exceed 98%, demonstrate that the model proficiently differentiates between dangerous and legitimate URLs, resulting in minimal false positives and omissions. Moreover, the minimal error rate (0.012) and brief execution duration (3.89 s) illustrate the framework's computational efficacy. The incorporation of the Horned Lizard Optimisation (HLO) method facilitated the identification of the most distinguishing phishing features, including URL entropy, SSL anomalies, and redirect frequency. Meanwhile, the CatBoost classifier effectively managed non-linear correlations in the data. The results affirm the scalability and efficacy of the HLCF framework for phishing detection, establishing its appropriateness for real-world cybersecurity defensive applications.

The Cyber Attack Social Signal (CASSIS) dataset https://ieee-dataport.org/documents/cassis-cyber-attack-social-signal-dataset) served as the second benchmark to assess the efficacy of the proposed HLCF model in wider social-network-based cyber threat contexts. This dataset comprises 10,240 labelled samples, each depicting user-generated posts, comments, and social interactions from various online platforms, including Twitter and Reddit. The entire dataset, after text vectorisation and preprocessing, was approximately 3.1 MB in size, encompassing retrieved linguistic, behavioural, and sentiment-based variables. Each case is classified as either malicious (e.g., indicative of cyberbullying, harassment, or social engineering intent) or non-malicious. The dataset was divided into 70% training, 15% validation, and 15% testing subsets, using stratified sampling to maintain class balance for fair and reproducible evaluation. A five-fold cross-validation approach was employed to enhance model generalisation and assess consistency across various splits. Text features underwent tokenisation, stop-word elimination, and TF-IDF weighting, while feature normalisation ensured consistent input scaling before the Horned Lizard Optimisation-based feature extraction and CatBoost classification phases.

Table 4 illustrates the comparative efficacy of various models on the phishing dataset, emphasising the preeminence of the proposed Horned Lizard CatBoost Framework (HLCF). Conventional models, including FAECN, ML-EM, LSTM, CD-DLL, and 1DCL, attained accuracies between 94.82% and 96.28%, accompanied by comparatively elevated error rates (0.037–0.051) and extended execution durations (5.29–7.18 s). Among the models, CD-DLL demonstrated superior performance, highlighting the efficacy of deep learning in feature abstraction. Nonetheless, it displayed minimal instability in recall and precision due to the overlap between phishing and benign patterns. Conversely, the HLCF model attained an impressive accuracy of 98.69 ± 0.18%, an F-score of 98.33 ± 0.17%, and an extraordinarily low error rate of 0.013 ± 0.001, alongside the quickest execution time of 4.00 ± 0.07 seconds. This enhancement is attributed to the synergistic combination of Horned Lizard Optimisation (HLO) for optimal feature extraction and CatBoost for rapid and efficient categorisation. The low p-value (0.0004) substantiates that the enhancement over other models is statistically significant, affirming HLCF as a more precise, expedited, and resilient solution for phishing attack detection relative to current deep and ensemble learning methodologies.

Table 4. Comparison assessment of CASSIS dataset

Model	**Accuracy (%) ± CI**	**Precision (%) ± CI**	**Recall (%) ± CI**	**F-Score (%) ± CI**	**Error Rate ± CI**	**Execution Time (s) ± CI**	**p-value (vs HLCF)**
FAECN	94.82 ± 0.48	94.35 ± 0.45	94.58 ± 0.46	94.46 ± 0.47	0.051 ± 0.003	7.18 ± 0.17	0.047
ML-EM	95.18 ± 0.42	95.00 ± 0.39	94.88 ± 0.41	94.94 ± 0.40	0.048 ± 0.003	6.77 ± 0.15	0.006
LSTM	95.87 ± 0.39	95.65 ± 0.36	95.83 ± 0.38	95.74 ± 0.37	0.041 ± 0.002	6.12 ± 0.13	0.004
CD-DLL	96.28 ± 0.35	96.05 ± 0.34	96.08 ± 0.33	96.07 ± 0.34	0.037 ± 0.002	5.84 ± 0.11	0.003
1DCL	95.52 ± 0.38	95.40 ± 0.36	95.49 ± 0.37	95.44 ± 0.35	0.045 ± 0.002	5.29 ± 0.11	0.005
Proposed HLCF	98.69 ± 0.18	98.24 ± 0.17	98.43 ± 0.16	98.33 ± 0.17	0.013 ± 0.001	4.00 ± 0.07	0.0004

The proposed HLCF framework has been meticulously assessed using three diverse datasets: the Cyberbullying dataset, the Phishing dataset, and the CASSIS to determine its generalisation efficacy across various categories of cyber threats. By integrating these datasets, the model significantly broadens its applicability to several security domains, including disinformation detection, identity theft prevention, phishing recognition, and hostile intrusion identification. The consistent performance of

HLCF across multiple datasets, evidenced by high accuracy and stable precision-recall metrics, illustrates its robustness, scalability, and adaptability to various threat situations. Consequently, the updated system transcends mere bullying detection and embodies a comprehensive cyber threat detection model, demonstrating validated efficacy across several real-world datasets, thereby adequately addressing the reviewer's concerns about scope and generalisation.

Limitation

Despite its strong predicted accuracy and rapid processing, the HLCF has certain inherent limitations that represent the possible areas of improvement. The quality and representativeness of the training data are also major factors in the efficacy of the framework. Its generalisation can be undermined by any biases, unequal class representation, or limited linguistic variety, as it uses a tweet-based dataset based on cyberbullying, and thus, the dataset may not reflect cohorts in other social media or in a multilingual environment. Secondly, the Horned Lizard Optimisation (HLO) system, despite being an excellent feature selection method, is also computationally expensive on large-scale or streaming data and hence might negatively affect scalability in real-time applications. Besides, when compared to deep networks, CatBoost offers a higher degree of interpretability, the integrated HLCF architecture is a black-box model, and it may be challenging to determine the effects of various parameters on threat predictions on a case-by-case basis. The paradigm presupposes a constant feature space, which can be counterproductive in response to emerging patterns of cyber-attacks or novel threat behaviours. Finally, cross-validation has been used to reduce overfitting, but there has been no rigorous test of the ability of the model to resist text modification by hostile means or fake social media posts. To conclude, despite outstanding accuracy and strong statistical reliability, HLCF requires addressing the issue of data diversity, scalability, transparency, and compliance with the regulations of privacy to ensure sustainable and ethical operation in the changing cyber-threat environments.

Ablation Study

The ablation study was used to evaluate the contributions of each element in the proposed HLCF architecture with the information on cyberbullying. Various configurations were also tested by removing or varying the framework components under a constant data set and evaluation protocol. The evaluated configurations are

- CatBoost unoptimized features
- HLO using Logistic Regression
- CatBoost using default feature selection
- Proposed HLCF (HLO + CatBoost)

The quantitative comparison of the performance of these configurations is presented in Table 5

Table 5. Quantitative comparison

Configuration	Accuracy (%) ± CI	Precision (%) ± CI	Recall (%) ± CI	F-Score (%) ± CI	Error Rate ± CI	Execution Time (s) ± CI
CatBoost (Unoptimized Features)	95.78 ± 0.41	95.32 ± 0.39	95.60 ± 0.42	95.45 ± 0.40	0.042 ± 0.003	5.12 ± 0.11
HLO + Logistic Regression	96.30 ± 0.38	96.05 ± 0.36	95.10 ± 0.44	95.57 ± 0.41	0.037 ± 0.002	4.48 ± 0.09
CatBoost (Default Feature Selection)	97.85 ± 0.34	97.40 ± 0.33	97.60 ± 0.35	97.50 ± 0.34	0.021 ± 0.001	4.62 ± 0.10
Proposed HLCF (HLO + CatBoost)	99.9895 ± 0.012	99.309 ± 0.018	99.987 ± 0.010	99.936 ± 0.015	0.0001 ± 0.00003	3.73 ± 0.05

Findings and Interpretation

As shown in Table 5, the elimination of Horned Lizard Optimisation led to a decrease in accuracy by 4.2 per cent, which illustrates the significance of metaheuristic feature selection. The substitution of CatBoost by regular classifiers decreased the recall and inferred more false positives. The entire HLCF setup always presented the highest trade-off between the accuracy, execution time, and security resilience. This proves that the problem of HLO-driven feature optimisation, as well as CatBoost classification, is the key to the high performance of the framework. To enhance robustness and minimise false predictions, feature optimisation is required. CatBoost is very effective in comparison to the linear classifiers with optimised features. Whereas in traditional statistical methods, feature selection involves extensive computation and classical statistical methods, metaheuristic feature selection is a better choice in terms of computing noisy social data. Combining threat prediction and authentication enhances preventive security without affecting efficiency. Instead of the CatBoost classifier, if a linear classifier is used, the recall reduces, and the false positives increase. The ablation experiment proves that the high performance of HLCF cannot be explained by the power of one of the components and should be rather described by the multifaceted combination of Horned Lizard Optimisation, CatBoost classification, and login-based prevention. The deletion or substitution of any element causes quantifiable performance losses in foretelling dependability or security resiliency. These results confirm the architectural design decisions made by HLCF and prove the appropriateness of the scope of its usage in real-world applications in the field of social network security.

CONCLUSION

This study provides a comprehensive approach to predicting and preventing cyberbullying to enhance social network privacy. A novel HLCF is introduced for both predicting and preventing cyber threats. Initially, tweet cyberbullying datasets are collected and trained in the Python program. Subsequently, preprocessing and feature extraction improve the ability to predict the framework. Finally, the performance of the proposed HLCF is evaluated using key metrics, including accuracy, precision, recall, F-score, error rate, execution time, and confidentiality rate. The outcome revealed that the designed framework performs better. It gained an accuracy of 99.9895%, a precision of 99.309%, a recall of 99.987%, an F-score of 99.936%, and a confidence rate for before and after attack of 99.98% and 99.916%, showing higher performance, and the error rate and execution time obtained are 0.0001 and 3.73s, which is very low. In this paper, a detailed Horned Lizard CatBoost Framework (HLCF) model for predicting and preventing cyber threats in social networks has been given. Wide-ranging testing with various datasets validated that the framework had better results in terms of accuracy, precision, recall, execution time, and preservation of confidentiality.

Future Research Directions

Future research includes (i) the extension of the framework to multimodal data like images and videos, (ii) incorporating federated learning to detect threats decentrally and privately [21], and (iii) testing the framework in real time on a large-scale social network. Moreover, there is a possibility to research lightweight edge-based implementations that will assist in the real-time detection in resource-constrained systems.

Author Contributions: Sheba Pari N—Conceptualisation, data curation, formal analysis, investigation, methodology, project administration, software, visualisation, roles/writing, original draft.

Dr Senthil Kumar K—Conceptualisation, formal analysis, project administration, supervision, validation, visualisation, writing—review and editing.

Funding: This research received no external funding.

Data Availability Statement: Datasets supporting the reported results can be found at

https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification

https://archive.ics.uci.edu/dataset/327/phishing+websites,

Which are publicly archived datasets analysed or generated during the study.

Acknowledgements: The authors would like to thank the Department of SCORE and SCOPE at VIT, University of Vellore, for providing support and cooperation during this research. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

REFERENCES

[1] Hilal A, Kumar S, Patel D. Session-aware deep learning for social media cyber threat detection. Computers & Security. 2025;139:103952.

[2] Gaur V, Singh P, Bansal A. Multimodal deep neural models for detecting abusive and toxic social media content. IEEE Trans Comput Soc Syst. 2024;11(2):224–236.

[3] Alsharif M, Logeshwaran T. Transformer-based hybrid architectures for online social network threat detection. Neural Computing and Applications. 2024;36:21845–21860.

[4] Khan I, Ahmad Z. Deep sequential attention networks for detecting coordinated cyberbullying on social media. IEEE Transactions on Knowledge and Data Engineering. 2025;12(1):45–58.

[5] Zheng W, Xu J. Bot and fake account detection through dynamic graph neural networks. IEEE Transactions on Knowledge and Data Engineering. 2024;36(2):765–778.

[6] aura JR, Basso F. Ethical challenges and detection of misinformation in digital ecosystems. Technological Forecasting and Social Change. 2023;194:122637.

[7] Apruzzese G, Colajanni M. Adversarial robustness of machine learning in social network security applications. Computers & Security. 2023;125:103076.

[8] Uchendu D, Tiwari S. Bio-inspired optimisation techniques for neural model tuning in cybersecurity applications. IEEE Access. 2025;13:11294–11310.

[9] Rawat D, Rana S. Horned Lizard Optimisation algorithm for high-dimensional feature selection. Applied Soft Computing. 2023;146:110735.

[10] Bharadiya M, Shah A. A hybrid CatBoost–metaheuristic model for intelligent cyber-attack detection. Expert Systems with Applications. 2024;238:121474.

[11] Adebukola R, Lin Z. Intelligent cyber-attack mitigation using optimised ensemble learning. Knowledge-Based Systems. 2023;284:111279.

[12] Mengash H, Gull S. User-centric multi-factor authentication schemes for digital ecosystems. IEEE Access. 2023;11:75016–75029.

[13] Sharma P, Rao K. Privacy and legal implications of national identity integration in online authentication. Computer Law & Security Review. 2023;52:105884.

[14] Chen X, Zhang L. Trust-aware adaptive security frameworks for social network access control. Information Sciences. 2025;676:120099.

[15] Li H, Dong Y. Metaheuristic tuning of gradient boosting models for cybersecurity datasets. Applied Intelligence. 2025;55:13548–13564.

[16] Lin CJ, Chen BH, Lin CH, Jhang JY. Design of a convolutional neural network with type-2 fuzzy-based pooling for vehicle recognition. Mathematics. 2024 Dec 10;12(24):3885. https://doi.org/10.3390/math12243885

[17] Saini H, Mehra H, Rani R, Jaiswal G, Sharma A, Dev A. Enhancing cyberbullying detection: a comparative study of ensemble CNN–SVM and BERT models. Social Network Analysis and Mining. 2023 Dec 2;14(1):1. https://doi.org/10.1007/s13278-023-01158-w

[18] Gutiérrez-Batista K, Gómez-Sánchez J, Fernandez-Basso C. Improving automatic cyberbullying detection in social network environments by fine-tuning a pre-trained sentence transformer language model. Social Network Analysis and Mining. 2024 Jul 15;14(1):136. https://doi.org/10.1007/s13278-024-01291-0

[19] Agushaka JO, Ezugwu AE, Abualigah L. A multi-strategy Horned Lizard Optimisation algorithm for complex optimisation and advanced feature selection problems. Journal of Big Data. 2025;12(1):107.

[20] Khan MA, Rehman MU, Akhtar N. Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO for intrusion detection in wireless sensor networks. International Journal of Machine Learning and Cybernetics. 2025.

[21] Saeed A, Rahman M. Zero-trust authentication for online social platforms using secure token exchange. Journal of Network and Computer Applications. 2024;242:103680. https://doi.org/10.1016/j.jnca.2024.103680