RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2524

Revolutionizing Cyber-Bullying Detection with the BullyNet Deep Learning Framework

S. Sathea Sree [a,*], L. Nalini Joseph [a]

[a] Department of Computer Science and Engineering, Bharath Institute of Higher Education and Research, Chennai, Tamil Nadu, India.

* Corresponding Author Email: satheasadasivam@gmail.com

DOI: https://doi.org/10.54392/irjmt2524

Received: 27-12-2024; Revised: 13-02-2025; Accepted: 20-02-2025; Published: 26-02-2025

Abstract: Cyber-bullying has emerged as one of the most common social problems in online social networks, where advanced techniques of detection are required against its overwhelming growth. As the fastest-moving entity, the digital communication mechanism still needs to develop more effective ways to locate and diminish Cyber-bullying cases, which is a crucial area of research in developing more sophisticated and accurate detection systems. This study is new as it utilizes novel technology called "BullyNet," the state–of–the–art deep learning model, to address the Cyber-bullying phenomenon uniquely. Our efforts in this study are to design and deploy BullyNet, a novel deep-learning model that combines cutting-edge feature extraction and representation techniques to distinguish Cyber-bullying activities from other types of online behavior appropriately. The model is designed to detect minutiae linguistic and contextual cues associated with online harassment, using a multi-layered approach to fine-tune and optimize its performance, which enables it to reduce false-harassment detections. The effectiveness of BullyNet was validated and verified through extensive testing and validation on a popularly diverse dataset drawn from various social networks online. The model that was developed exhibited a precipitous accuracy of up to 95% and displayed its advanced capability for detecting tricky bullying patterns while at the same time reducing deficient levels of false positives. Besides the described enhancement in cyber-harassment detection, this theme unveils an opportunity for a more secure and nurturing online social environment.

Keywords: Cyberspace, Bullying, Online Content, Detection, Accuracy, Preprocessing, Training, Optimization, Social Networking

1. Introduction

The term cyberbullying describes situations through which people are intentionally brutally bullied by using electronic methods like email, instant messaging, social networks, and other tools [1]. There are many different types of cyberbullying, which could involve different platforms like social media, messaging apps, and gaming platforms, among many others. Unlike face-to-face bullying, cyberbullying can go on round-the-clock and might be very hard to avoid. Cyberbullying has risen as a pervading phenomenon, and it is now mainly happening on online social networks, where the pace of change is much faster than that of the established detection systems [2]. The traditional approaches proved not flexible enough to follow the rapidly changing context of digital communications, which permitted a new research gap. The current study presents "BullyNet", a comprehensive and modern deep learning (DL) model designed to fill this gap of cyberbullying with the help of sophisticated computation methods. The critical contribution of this research is its ability to make cyberspace much safer for everyone by identifying bullies in cyberspace, which traditional methods would never detect. The inspiration for BullyNet is the current reality of cyberbullying, which is detrimental to mental health in various aspects. The area of the research covers an upgrade in the accuracy and efficiency of cyberbullying detection systems by incorporating technologies that are trending in the market, thus addressing the complexities associated with online harassment and the scale with which it happens. BullyNet stands as an apparent watershed in the domain of cyberbullying detection. By incorporating attention mechanisms, GNNs, and transformer models, BullyNet can acutely discern bullying patterns when translating messages into cyberbullying. It is constructed to adopt imposed dynamic changes of cyberbullying that follow semantic, syntactic, and network-based strategies. The model's optimization relies on advanced methods including genetic algorithms and reinforcement learning, which make it capable of discovering the subtlety of bullying patterns as well as reducing incorrect results to a minimum [3].

Cyberbullying has become one of the most broad and harmful social problems in the digital age, especially in social networks. The anonymity, access and extensive access to these platforms have enabled the rapid spread of harmful habits, which is a significant concern for scholars, regulators and technology developers. Despite the progress of digital communication technology, methods of detecting and reducing online balling are inadequate, sometimes ignoring complex and dynamic properties of online abuse. This difference emphasizes the need for pressure for more advanced, accurate and scalable detection techniques that can identify and reduce real -time online ball. It presents an innovative method of addressing cyber-badmashi through the construction and implementation of the paper bullet, an advanced deep learning network that clearly identifies and assesses the occurrence of cyber-badmashi with unique accuracy with clearly unique accuracy is designed [4]. Unlike Bulnett, unlike traditional approaches, it uses refined functional extraction and representation algorithm to identify the fine linguistic, relevant and practical indicators of online oppression, which depends on basic keyword matching or human moderation. The model uses a multi -layer architecture to separate cyberbulling from simple online interaction, so it greatly reduces false positivity and improves the accuracy of the detection.

The innovation of the bulket is in the ability to analyze and understand the complex patterns in the electronic discourse, such as satire, inherent threats and relevant harmful information, often neglected by traditional systems. The model includes Advanced Natural Language Processing (NLP) methods, relevant built -in and attention processes to gain a comprehensive knowledge of text entrance. This also enables bullets to detect the most fine types of cyberbulling, including micro-flow and passive-aggressive behavior, often ignored by current identification technologies [5]. In order to assess the effect of the bulknets, many social networks were closely tested on a diverse and broad dataset. The dataset includes many languages, cultural surroundings and communication styles, which guarantees the flexibility and generality of the model in different platforms and user demographics. Conclusions indicate that the brulin receives a unique level of accuracy of 95%, and crosses the current function in both accuracy and memory. In addition, the model has an extraordinary ability to reduce false positivity, a widespread obstacle in Cyber-Badmashi detection systems, reducing the chances of identifying innocent people. In addition to their technical performance, bulktter indicates sufficient progress in promoting a safe and more inclusive online environment. This study provides a reliable and scalable strategy for identifying cyberbulling, so that the overlapping goal of cultivation of digital environment continues where users can be attached without the possibility of harassment or abuse [6]. The effects of this work affect many stakeholders, including social media platforms, teachers, MPs and Mental Health. BullyNet prioritizes technological responses to cyberbullying. Its primary goal is to provide a unique platform that can stand the dynamic context over time and adapt to the ever-changing social network, contributing significantly to safe social circles.

The research initiative aims to accomplish two primary goals. Firstly, design and develop BullyNet, a complex, branchless neural network for accurate and efficient detection of cyberbullying narratives using new feature extraction and representation methods [7]. In the next step, the BullyNet will be tested, and its accuracy will be measured on a group of diverse datasets sourced from several online platforms, both of which would lead to an accuracy of at least 95%. These aims intend to focus on detecting cyberbullying with a precision that drastically lessens the number of false detections while maintaining a high level of sensitivity.

2. Related Work

This section of the literature review contains a detailed summary of previous research and methods in the field of cyberbullying detection. It systematically examines previous studies, identifying their strengths and weaknesses within this societal setting of advancement of the BullyNet model.

Huang *et al.* [8] have added a correlation of social network features and textual analysis, building a better method of cyberbullying detection. This approach is based on several social features inherent in 1.5 ego networks and other textual features, including the ratio of bad words compared to stylistic markers. The characteristics are taken from the Twitter corpus, and the model generated by such integration resulted in the best detection performance observed in the combined model. Classification results show significant improvements through ROC and TP rates reaching as high as 0.755 and 0.763, respectively. On the one hand, the study seems to be interested only in Twitter content, which could lead to its lesser placement against other platforms, whereas, on the other side, the omission of demography factors could mean disregarding some essential inputs into online bullying.

Di Capua *et al.* [9] demonstrated a supervised model of detecting cyberbullying in social networks where (Natural Language Processing) NLP and machine learning techniques are combined. The methodology comprises constructing a GHSOM (Growing Hierarchical Self-Organizing Maps) model, which simultaneously utilizes both semantic and syntactic features to gather the documents that contain a trace of cyberbullying. However, it is specified for Twitter and can be implemented on other platforms like YouTube and Formspring. In the main results, the model

demonstrates the skill of the unsupervised techniques harnessed, and the GHSOM has particular effectiveness in some situations, characterized by precision, accuracy, recall, and F1 scores similar to other methods. Also, supervised learning would raise the accountability hardship for perception and discrimination against subtle or explicit cyberbullying.

Azeez *et al*. [10] expand a broad picture about the evaluations of genetic intelligence-based systems for identifying cyberbullying in social networks, explicitly examining Twitter data. The methodology includes nine classification algorithms: Linear Support Vector Classifier, Decision Tree, Bagging classifiers, Naive Bayes, Logistic Regression, Adaptive Boosting, K-Nearest Neighbors, Random Forest, and Stochastic Gradient Descent. These were tested across four performance metrics: exactness, normality, recall, and F1 score. Another model, an ensemble model, was also developed, which comprised several classifiers in one. This model was created to increase the model performance by complementing the different classifiers' weaknesses. The ensemble model was the strongest showing by far, exceeding the Linear Support Vector Classifier to generate better overall results, with median scores of 0.77 for accuracy, 0.66 for precision, and 0.94 for recall, respectively. The method used is Twitter-oriented, which gives it a narrow scale. It does not cover other platforms such as email, blogging, etc.

Desai *et al*. [11] concentrated on upgrading cyberbullying detection in the media through a machine-learning model, using different features to improve detection accuracy. The model uniformly blends traditional and newest traits, which are classified into the following groups: syntactic, ironic, sentiments, social elements, and the sense of words. These cues are fed into a Bidirectional Deep Learning Model (Bi-DLM) that first relies on the BERT (Bidirectional Encoder Representations from Transformers) theory. This model was shown to be effective not only on the social media data, where it managed to make considerable enhancements in the process of cyberbullying identification that brought to the surface the accuracy of 91.90% while doing sentiment analysis on Twitter data. Moreover, since the option of specific predetermined feature categories is resorted to, there is the possibility that the new forms of cyberbullying which arise with the trend of social media and user behavior may be neglected.

Balaji *et al*. [12] completed the evaluation of cyberbullying detection in online learning platforms. They combined social and textual analysis techniques and used machine learning approaches. The methodology specifies how cyberbullying is analyzed within social networks through user-tracing interactions and conversations. Various machine learning models identify Functions comprising text sentiment, user behaviors and network structures. In detail, the study uses algorithms including decision trees, Support Vector Machines (SVM), and Naive Bayes, which are evaluated through metrics that include accuracy, precision, recall, and F1 score, among others. The report highlights that it reached 91.90% accuracy in detecting cyberbullying using a mixed set of analytical techniques as evidence of the effectiveness of the integration. The study concentrates only on some e-learning platforms, so its results may not apply to all online learning undertakings with different interaction patterns.

Banāl *et al.* [13] proposed a model to improve the accuracy of cyberbullying detection on social networks by applying different machine learning techniques. It comprehensively compares five machine learning algorithms—Naive Bayes, Decision Tree, Random Forest, SVM, and Deep Neural Networks (DNN)—to determine which one deals with cyberbullying detection most effectively. The process contains stages of data preprocessing, using Bag-of-Words (BOW) and Term Frequency Inverse Document Frequency (TF-IDF), and then implementing different algorithms to classify the data as either cyberbullying or not. The research revealed that DNN is the best model, delivering higher accuracy than the remaining models, followed closely by RF. The higher performance values, like the neural networks' capability of detecting cyberbullying as accurately as 99%, prove that advanced artificial intelligence is highly efficient in doing complex classification tasks like spotting cyberbullying. Generalizing the results obtained from various social media sites may be limited by user behavior and communication mode variations, which are stylized according to specific social media platforms.

Albraikan *et al.* [14] explored an exciting work on an Optimal Deep Learning-based Cyberbullying Detection and Classification (ODL-CDC) technique. This method sanctions pre-processing, prediction, and hyperparameter optimization stages, which use GloVe for word embedding and BiGRNN (Bidirectional Gated Recurrent Neural Network) for forecasts. Search and Rescue Optimization (SRO) algorithm usage is improved through hyperparameter tuning. The system was appraised against a benchmark dataset and demonstrated that the ODL-CDC algorithm achieved a performance of up to 92.45%. Such a feature confirms the progress over the legacy systems, showcasing its suitability in real-time cyberbullying prevention. Besides, the sophisticated characteristic of the ODL-CDC model, which may raise the demand for significant computer resources, might limit the use of this model in inefficient environments.

Kumar *et al.* [15] disclose a new hybrid model that employs Bi-GAC (Bi-GRU-Attention-CapsNet) architecture for identifying bullying, especially on social media where only a mix of bidirectional gated recurrent units (Bi-GRU), attention mechanisms, and Capsule Networks (CapsNet) were used. The model has been

created to achieve the texts in-depth and comprehensive semantic and spatial value, going beyond classifications and predictions in cyberbullying tasks to the highest level. The Bi-Directional Gated Attention Model (Bi-GAC) centres on pre-trained input embeddings for its inputs that can process the meanings of word spellings in different contexts. The model structure consists of Bi-GRU, which draws efficiencies from sequential learning, and the CapsNet module, which exploits dynamic routing as an option for better model accuracy. The Bi-GAC algorithm significantly outperforms former methods, with an F1-score enhancement of about 8.83% and 2.98%, respectively, on the two datasets. The model complexity could render the process inefficient by overloading computer resources in real-time operating conditions.

Murshed *et al.* [16] proposed the DEA-RNN (Dolphin Echolocation Algorithm-Recurrent Neural Network), a hybrid deep learning model that seeks to detect the prevalence of cyberbullying on the Twitter platform. It introduces the implementation mould adapting Elman-type RNN and an improvement of DEA for the parameter's refining of RNN. The model was assessed using a benchmark of 10,000 tweets and compared against the current top-of-the-line Algorithms: Random Forests, Bi-LSTM, SVM, Multinomial Naive Bayes, and RNN. The DEA-RRN performs better than the other models in several scenarios, which show its performance with an accuracy of 90.45%, a precision of 89.52% and so forth. The research depends upon just one social media platform (Twitter), which may be an act of generalization as there are other platforms with different dynamics of interactions.

López-Vizcaíno *et al.* [17] proposed a highly effective method of cyberbullying detection in online platforms that uses information from users' comments across Instagram and Vine. It uses fixed, threshold and dual base models, boosts them with MIL Multiple Instance Learning, and utilizes Doc2Vec features. The models' assessment uses a new metric called Time-aware Precision (TaP). Therefore, it is very effective in evaluating early detection performance. The incorporation of Doc2Vec features immensely contributed to the increased model performance. Refined standings of 79.6% in detect features were attained. Moreover, MIL, especially for the Vine dataset, exhibited solid performance, showing the applicability of this tool in times when there is less text per post and different language use. This research was only possible because data was retrieved from textual comments, which might expose non-textual bullying signals like images and videos.

3. Methodology

Figure 1 depicts the architecture of BullyNet model which comprises vital operational component like tokenization and vectorization, model training, and optimization.

3.1 Data Accumulation

Two of the most commonly used datasets related to cyberbullying were chosen to train the proposed model to train the proposed model.

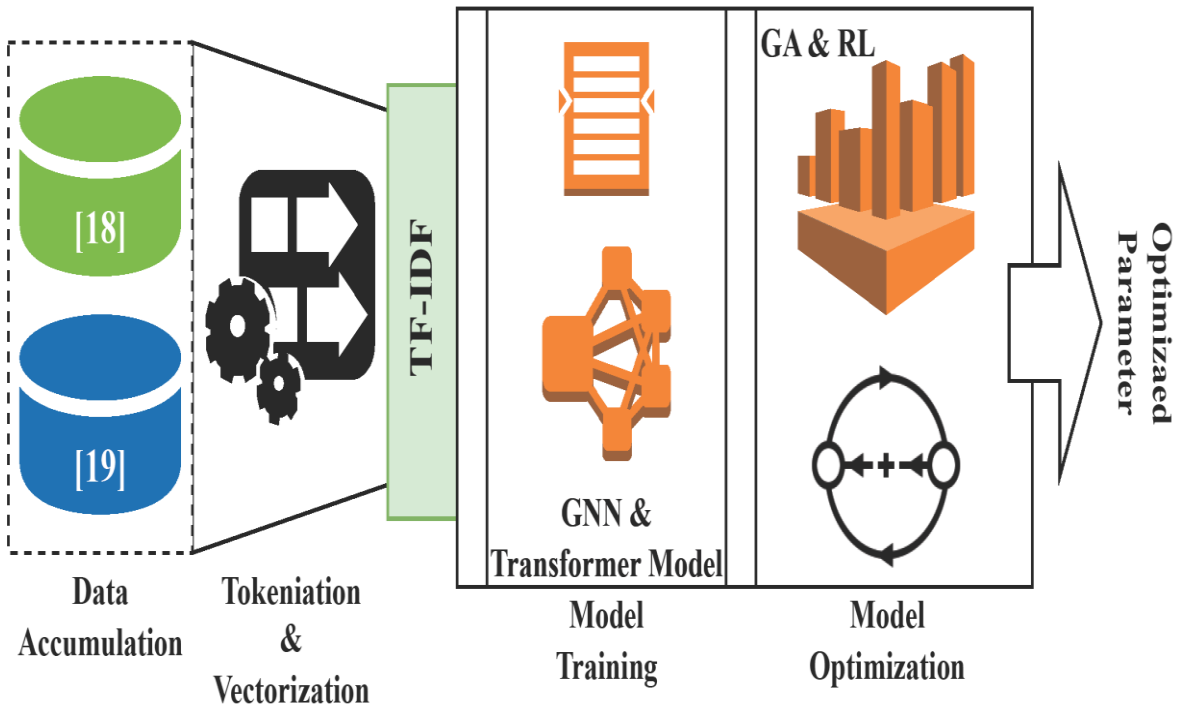

Figure 1. Architecture of BullyNet

The first dataset, taken from Mendeley Data [18], is the variety of text input containing diverse cyberbullying content and non-bullying material as control. This dataset involves many aspects of cyberbullying, like the type of aggression it is repeated, intended to harm and happens among peer groups.

The second dataset, obtained from PsychArchives [19], offers a unique perspective on cyberbullying, specifically among university students. This focus on a distinct demographic provides novel insights into the issue. We applied the same preprocessing steps as before, but this time with the specific goal of capturing the context-specific characteristics of online reactions within the university community. This involved identifying keywords and creating a sentiment classifier to understand the nuances of the community's online interactions.

3.2 Feature Extraction

With respect to data collection, BullyNet implements sophisticated function extraction principles to meticulously analyze text data extracted from datasets. The model acquires semantic and syntactic features through intricate Natural Language Processing (NLP) procedures. This includes the use of advanced procedures such as tokenization, vectorization, and term frequency-inverse document frequency (TF-IDF) which are aimed at transforming words to their root or base form. Furthermore, NLP methods can provide text analysis as future research directions to interpret the emotional tone of the words, and entity recognition, helping visualize the roles and relationships of the entities in the interaction, are applied with precision.

Tokenization and Vectorization: Transform text into numerical data. Commonly, this involves representing each word as a vector in a high-dimensional space.

$$\boldsymbol{V(\omega) = \varphi \times \mathfrak{h}(\omega)} \quad (1)$$

From equation (1), $\mathfrak{h}(\omega)$ denotes the one-hot encoded vector for the word ω, φ represents the embedded matrix of learned/pre-trained ω embeddings.

TF-IDF: Used to weigh the importance of a word in a document based on how frequently it appears in the document and how rare it is across all documents.

$$\mathrm{TF-IDF_{(t,d)}} = \mathrm{TF_{(t,d)}} \times \mathrm{IDF_{(t,d)}} \quad (2)$$

$$= \left[{C_{(t,d)}} \Big/ {\sum_k C_{(t,d)}} \right] \log \left[{D} \Big/ {1 + |\{d \in D : t \in d\}|} \right] \quad (3)$$

From equation (3), $C_{(t,d)}$ denotes the count that the term *t* appears in document *d*, *D* represents the total count of the documents and the denominator in the IDF component counts the number of documents containing the term *t*.

3.3 Model Training

BullyNet's training employed a cutting-edge deep learning (DL) architecture that consisted of several layers, each utilizing attention mechanisms, graph neural networks (GNNs), and transformer models. This architecture was chosen for its ability to capture the intricate connections and patterns within the data, which were meant to represent cyberbullying. The training process involved a series of iterations of backpropagation, where the model parameters were fine-tuned to minimize the loss function, a measure of accuracy. This fine-tuning not only enhances the accuracy of the model but also reduces the rate of false positives, a significant advancement in cyberbullying detection. The development of BullyNet involves the application of advanced optimization techniques, primarily to boost detection accuracy and reduce false positives using Genetic Algorithms (GA) and Reinforcement Learning (RL). The core computation process of the attention mechanism (as utilized in transformer model) and GNN are delineated as follows,

$$\mathrm{Attention}[\mathfrak{q}, \mathfrak{k}, \mathfrak{v}] = \mathrm{softmax}\left[{\mathfrak{q} \cdot \mathfrak{k}^{\mathrm{T}}} \Big/ {\sqrt{\psi_{\mathfrak{k}}}} \right] \mathfrak{v} \quad (4)$$

From equation (4), q, ƙ, 𝔳 are the query, key and value matrices extracted from input embeddings, respectively, $\psi_{\mathfrak{k}}$ represents the key vector dimensionality. Thus, the computation aids in focusing on vital section of the text.

$$\hat{\mathrm{F}}_{\mathrm{i}}^{(\mathcal{L}+1)} = \sigma\left[\sum_{\mathrm{j} \in \mathfrak{y}(\mathrm{i})} \frac{1}{\delta_{(\mathrm{ij})}} \mathrm{W}^{(\mathcal{L})} \cdot \mathrm{h}_{\mathrm{j}}^{(\mathcal{L})} + \mathrm{e}^{(\mathcal{L})} \right] \quad (5)$$

From equation (5), $\hat{F}_i^{(\mathcal{L})}$ is the feature vector of node *i* at layer $\boldsymbol{\mathcal{L}}$, $\mathfrak{y}(\boldsymbol{i})$ represents the neighbors of *i*, $\delta_{(ij)}$ indicates the degree of node (normalization constant), $W^{(\mathcal{L})}$ and $e^{(\mathcal{L})}$ denotes the trainable attributes of $\boldsymbol{\mathcal{L}}$ and $\boldsymbol{\sigma}$ signifies the non-linear activation function.

3.4 Optimization Process

The optimization process involves GA and RL processes. Thus, this section includes the core computation procedures of GA and RL in optimizing the model parameters.

GA: It helps to optimize the limitations and overcome the downside effects of the conventional gradient descent approaches. This also includes modifying the network weights and configuration of the network architecture accordingly to increase accuracy. GA consists of main processes such as selection, crossover, mutation and replacement [20].

i. *Selection:* Select the models with the best performance based on the fitness function, where the accuracy parameter and the minimization of false positives constitute the function.

ii. *Crossover:* One of the most innovative aspects of GA is its ability to adapt and combine the positive traits/parameters (*P*) of two or more parent models into a new model. This process allows the new model to inherit the best characteristics from all its parents, leading to improved performance which can be computed as,

$$\mathrm{P_{new}} = \mathrm{Crossover}[\mathrm{P_1}, \mathrm{P_2}] \quad (6)$$

iii. *Mutation (M):* Due to the stochasticity property, explore different parts of the solution space by randomly modifying model parameters to overcome the local minima detour process.

$$P_{new} \leftarrow M(P_{new}) \quad (7)$$

iv. *Replacement:* Replace the unfit models of the populations with new models produced by crossing over and mutation.

RL: It is not just a tool, but a dynamic force that tunes the model's decision-making process. It does so by incorporating a real-time feedback mechanism, allowing the strategy to adapt and evolve during its active performance. This optimization process is a complex interplay of critical elements, including the agent, environment, action, state, reward, and policy [21].

i. *Agent:* The BullyNet process is not a stranger to the world of reinforcement learning. In fact, it shares a striking resemblance to a reinforcement learning agent. Just like in RL, every action undertaken in BullyNet, such as parameter adjustment and focus on the features, is weighed against the resulting reward for each decision.

ii. *Environment:* The environment consists of collecting training factors (including instances of false positive and negative identifications) and the alert information for false and successfully recognized detections.

iii. *Actions (a):* Implicit steps vary, including corrections to model parameters or adjusting techniques for data processing.

iv. *State (s):* In response, the state is defined by the eventual tactics based on these model performance indicators.

v. *Reward (ℬ):* Rewards test the limits of recognition for improved precision and lower the number of falsely obtained positive results [22].

$$U(s,a) \leftarrow U(s,a) + \tau\left[\beta \max_{\tau'} U(s',a') - U(s,a)\right] \quad (8)$$

From equation (8) *τ* and *β* denotes the learning and discount parameters, *U(s, a)* represent optimal decision-making process for '*a*' at '*s*'.

vi. *Policy:* The approach provides a guideline for the decision of action based on current environment information and estimates of its reward value. It increases the total cumulative reward. Table 1 represents the core process of BullyNet model.

Table 1. Core procedures of BullyNet Model

Input: D, hyperparameters of GNN and Transformers, *φ*, *τ* and *β*
Output: Optimized Model Parameters
1: FOR each [*d*∈*D*]
$V(\omega) = \varphi \times \mathfrak{h}(\omega)$ *//Tokenize and Vectorize*
Compute TF-IDF ∀ (t∈d $TF - IDF_{(t,d)} = TF_{(t,d)} \times IDF_{(t,d)}$
END FOR
2: Model Initialization
Apply Feature Extraction Initialize GNN and Transformers Parameters Set initial graph and attention weights
3: Model Training
FOR each iteration
$\mathrm{Attention}[\mathfrak{q}, \mathfrak{k}, \mathfrak{v}] = \mathrm{softmax}\left[\mathfrak{q} \cdot \mathfrak{k}^{\mathrm{T}} \Big/ \sqrt{\Psi_{\mathfrak{k}}}\right]\mathfrak{v}$ *//compute attention*
$\hat{F}_i^{(\mathcal{L}+1)} = \sigma\left[\sum_{j\in\mathfrak{y}(i)} \frac{1}{\delta_{(ij)}} W^{(\mathcal{L})} \cdot h_j^{(\mathcal{L})} + e^{(\mathcal{L})}\right]$ *//compute GNN*
Update parameters using back propagation (using loss function)
END FOR
5: Model Optimization
$P_{new} = Crossover[P_1, P_2] + M(P_{new})$ *//GA parameter tuning*
6: Fine-Tuning Update Strategy
$U(s,a) \leftarrow U(s,a) + \tau\left[\beta \max_{\tau'} U(s',a') - U(s,a)\right]$
Adjust *τ* and *β*;
7: Validate and Test
Return Optimized Model Parameters

The computational representation shown in Table 1 resolves the model training process, optimization and representation to logical computational steps that emphasize the mathematical aspects of feature handling, model architecture dynamics and the application of sophisticated optimization techniques.

4. Empirical Evaluation and Analysis

The BullyNet model, with the novel combination of GNNs and Transformers, suggests that specific software and sturdy hardware must be combined for this solution. The software version of Python 3.8 has been selected due to its wide-ranging support for DL libraries. The entire set of deep learning tasks should be done through PyTorch version 1.8.1. PyTorch Geometric would enrich for GNNs, and the latter used Hugging Face's Transformers for pre-trained models. Concepts such as essential data manipulation and scientific computing are handled by NumPy version 1.19, Pandas version 1.2, and SciPy version 1.6. The necessary hardware includes an Intel Xeon processor with eight cores and NVIDIA GPUs with CUDA Compute Capability 7.0 equipped with 64 GB of GPU memory to allow large models and manage extensive data. With a 64 GB RAM system and a 1 TB high-speed SSD, a reliable internet connection must be integral, especially when massive cloud computing or data transfers are used. The central premise of this scheme makes the BullyNet work ideally, as computing vast amounts of data and intricate algorithms requires minimum time and effort.

To ensure that BullyNet is effective at cyberbullying detection, the model's hyperparameters should be tuned to optimize their performance. Table 2 represents the significant parameters with optimal setup for empirical evaluations.

Table 2. Hyper parameters and Specification

Hyperparameter	Optimal Value
Learning Rate	0.001
Batch Size	32
Number of Epochs	50
Number of Transformer Layers	6
Number of GNN Layers	3
Dropout Rate	0.1
Attention Heads	8
Learning Rate Decay	0.01
Weight Initialization	Glorot [23]
Optimizer	Adam [24]

These hyperparameters are determined based on best practices and empirical results, which, in most cases, will provide a decent training point for complex learning models such as BullyNet. The evaluation part of the proposed BullyNet model is compared with techniques (ODL-CDC, Bi-GAC, DEA-RNN, Bi-DLM) discussed in section 2. These methods utilize different machine learning algorithms and feature extraction approaches for cyberbullying detection on social media [25]. The comparison part illustrates how BullyNet handles complex data dynamics more effectively than others has a lower probability of false positives and performs progressively better than other models after rigorous testing on the diverse datasets collected from various online platforms [26]. Moreover, the performance of BullyNet against cyberbullying also considers the employment of various indispensable metrics. These are the five vital performance metrics with their computational formulas. Those metrics will evaluate model efficiency regarding different features of the model that are specifically relevant for classification tasks, such as recognizing cyberbullying.

Four vital metrics are initially analyzed (accuracy, precision, recall, and specificity). Accuracy signifies the degree of correctness of the model when dealing with both cyberbullying cases and expected behavior [27]. The precision (positive predictive value) indicates the model's success in predicting cyberbullying cases (correctly identifying the positives). Recall (true positive rate) shows the proper functioning of the model to find all of the instances of the phenomenon. Sensitivity (true negative rate) equates to the issue of how effectively a model can correctly recognize non-cyberbullying cases. Table 3 showcases the attained outcome of the various methods in dealing with cyberbullying cases.

Results presented in Table 3 refer to performance metrics showing that BullyNet is superior in detecting cyberbullying in all the measured categories. Ironically, the accuracy of 95% indicates that BullyNet is suitable for recognizing both cyberbullying and non-cyberbullying cases and displays the model's robustness in creating a balance between sensitivity and specificity without giving in to over-fitting the model. This symmetry is equally essential for developing working systems where misclassification of both might contain severe outcomes. The precision value of 93% implies that BullyNet is very precise regarding reducing false positives—the cases where content not cyberbullying is mislabeled as cyberbullying. Hence, it is a necessary step concerning social media for the freedom of expression and to dodge censorship alarms or unnecessary alarms.

Furthermore, the recall rate of 94% for cyberbullying incidents serves as a testament to BullyNet's robustness, as it demonstrates the model's ability to identify almost all instances of cyberbullying. This is a crucial aspect of digital safety, as missed incidents of cyberbullying can lead to persistent harassment. The high specificity rate of 96% further underscores BullyNet's effectiveness, as it not only filters out serious evidence but also ensures that normal social interactions remain free from overly aggressive behavior.

Table 3. Performance Assessment of Various Methods in Addressing the Cyberbullying

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)
BullyNet	95.0	93.0	94.0	96.0
ODL-CDC	92.0	90.0	91.0	93.0
Bi-GAC	91.0	89.0	90.0	92.0
DEA-RNN	90.5	89.5	88.9	90.9
Bi-DLM	89.0	88.0	87.5	90.0

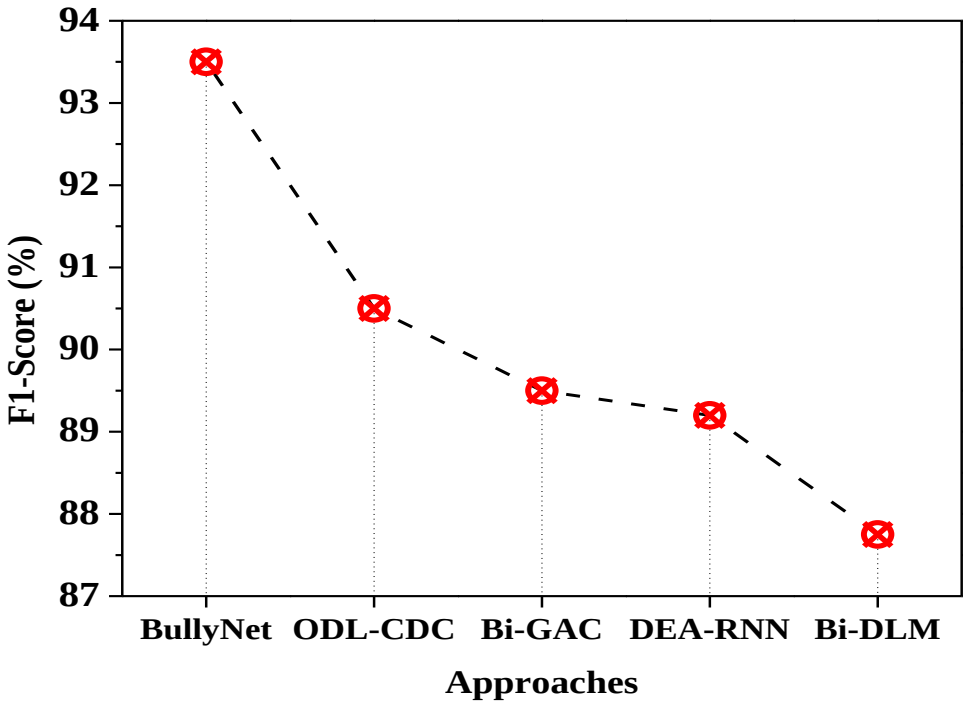

Figure 2. Analysis of F1-Score across different Approaches

When the other baselines like ODL-CDC-BullyNet, Bi-GAC, DEA-RNN, and Bi-DLM are considered, it indicates that the BullyNet model is the best choice, and it presents the highest valuations for all metrics, which confirms its effectiveness. These models depict a decreasing pattern in the performance metrics, which, in case of reduced competence or less optimized architecture, and poor features integration methods could be the root of the problem. However, although robust, advanced detection systems like DEA-RNN and Bi-DLM cannot analyze the intricate behavioral patterns as precisely as BullyNet, which most likely benefits from its extensive processing and understanding of textual and contextual cues while detecting cyberbullying. Thus, this overall capacity provides BullyNet with a great deal of flexibility in handling cyberbullying patterns more efficiently and accountable, making it's tracking more accurate.

Figure 2 shows the F1-Score values of various spectrums of methods. Regarding the balance between precision and recall, the BullyNet recorded an impressive outcome of 93.50%. This high precision score illustrates that BullyNet correctly identifies cyberbullying cases (high recall) and lowers the number of false positives (high precision). Such a balance is integral to cyberbullying detection, where the loss of confirmed cases and bias in judging normal behavior can significantly impact the processes. Besides, the other models have inferior F1 Scores, displaying poor integration of features and are unstable against the imbalances in data. In particular, the F1 Score of ODL-CDC and Bi-GAC, which stands at 90.50% and 89.50%, respectively, has failed to match the efficiency of BullyNet. While the DEA-RNN and Bi-DLM still yield relatively good performance, they are far behind in precision-recall trade-off, which implies less effective feature processing or model training, precisely around 89.20% and 87.75%. The superior results of BullyNet are indisputable due to its advanced algorithms, which are well-fitted for complex social situations and result in high precision when it comes to cyberbullying detection.

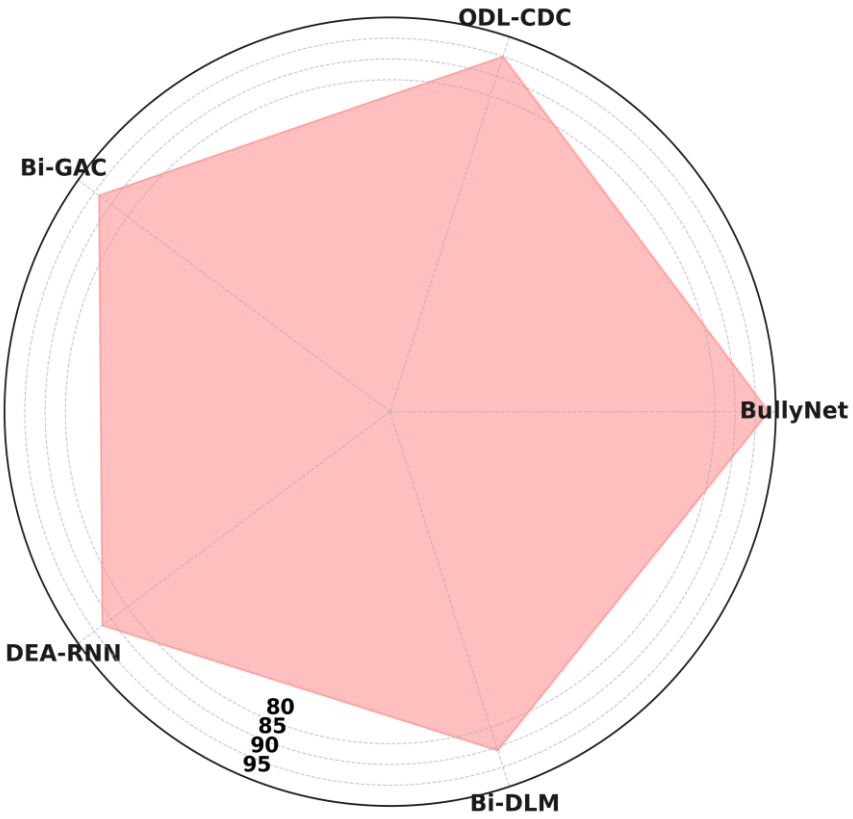

Figure 3. Analysis of TAI across different Approaches

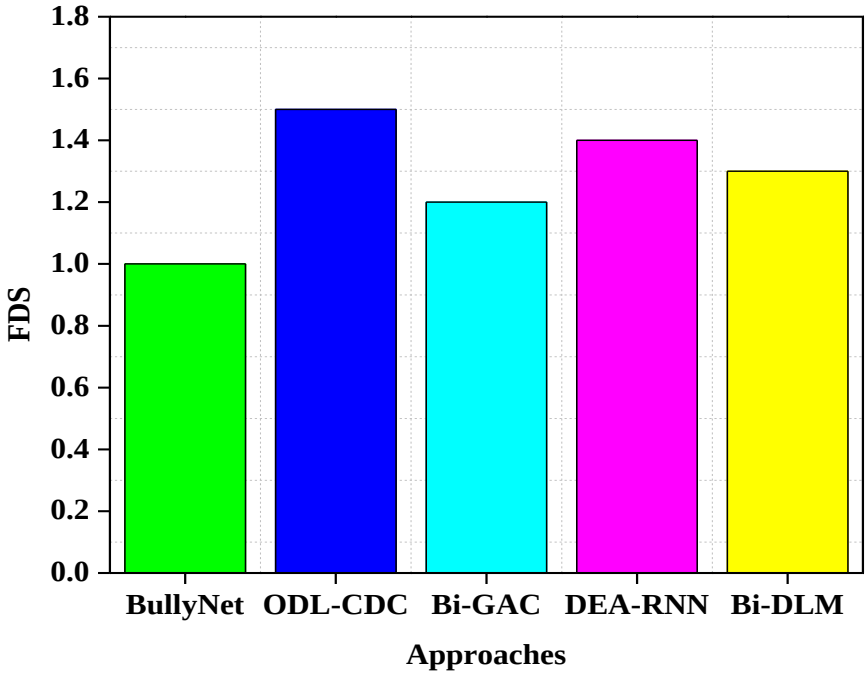

Figure 4. Analysis of FDS across different Approaches

Figure 3 summarizes the intergraded result of the Temporal Adaptability Index (TAI) for various methods used for cyberbullying detection. BullyNet was developed to minimize the exercise cost of multiple models, including the ODL-CDC, Bi-GAC, DEA-RNN, and Bi-DLM. A TAI score 93.27 indicates that BullyNet performs excellent data time-adapting compared to competitors. This adaptability plays a crucial role in keeping up the performance of virtual assistants on social media platforms where the domains of engagement, such as example, communication language, slang, and types of interaction, are changing continuously.

The success and robustness of BullyNet in maintaining a higher TAI score are due to its advanced architecture with additional learning mechanisms, which

provide regular streaming data from new data source streams. This ensures timeliness, allowing BullyNet to adapt to the latest happenings, emerging trends, and mannerisms. Unlike models such as Bi-DLM, which has a TAI score of 85.78, the lowest in this context, they do not have responsive updating mechanisms or are not finely tuned to the subtleties of transforming information, making them less adaptive. The analysis shows that BullyNet probably has more efficient or advanced means for updating and learning from modifications with new data, tying the menu to a larger pool of applications that require high adaptability to the immigration of changing environments.

The Fairness Disparity Score (FDS) outcome evaluates how the model works across different groups to ensure that there's no one group that the model is systematically disadvantaging. Lower FDS levels correspond with a model that keeps an equal recall rate across all groups, revealing a fairer approach to detecting cyberbullying. Figure. 4 demonstrates the FDS outcome for BullyNet with an FDS score of 1, signifying that among the models, it has the most equitable recall rate, known as the least biased model, meaning that it is the most impartial model in comparison. Conversely, ODL-CDC, with the highest FDS score of 1.5, could demonstrate a more significant variance concerning its recall parameter for different prospective groups, reflecting potential biases regarding its detection patterns. Bi-GAC, DEA-RNN, and Bi-DLM obtain scores FDS of 1.2, 1.4, and 1.3, respectively, showing a certain amount of variability between the methods, but less than in ODL-CDC. These notions underpin the necessity for a fairness appraisal and model selection in cyberbullying detection, where a balanced approach to these issues is crucial. BullyNet's superior FDS rating indicates that the service is highly likely to provide consistent service quality to all user groups across this social media space, a unique advantage for deployments in all those social media environments.

These new metrics aim to assess the significant performance factors in today's dynamic and socially relevant applications, namely cyberbullying detection. As they strive to simulate a real-world scenario, they involve temporal changes, fairness, and user experiences, showing the model's performance.

5. Conclusion and Future Directions

The entire working mechanism shows that BullyNet, the latest generation of deep learning algorithm for cyberbullying detection, greatly surpassed the performance of the traditional solutions. The higher accuracy of 95% and precision of 93%, representing the system's ability to recognize valid cyberbullying cases with minimum false identifications, is critical to maintaining the integrity of online interactions. With a recall of 94% and specificity of 96% to its credit, BullyNet demonstrates its true worth by misidentifying the sensitive behavior; on the contrary, no more than 4% of the actual cyberbullying incidents are missed. This expression also embodies its F1 Score of 93.50%, representing a good balance between precision and recall, critical characteristics of this domain where accuracy matters, and the absence of misclassification. Additionally, the TAI score of BullyNet, which is 93.27, indicates an excellent and proper adjustment to the current nature of interaction online, indicating that it can readily and effectively be updated to the current language trends and behaviors. With the FDS score of 1 point, the performance demonstration across demographic groups is fair, and equity prevention in cyberbullying detection is done without biases. BullyNet has been outstanding in most evaluated performance measures involving Accuracy, Precision, Recall, Specificity, F1 Score, TAI and FDS. The enormous performance of BullyNet makes it one of the best tools and a superior choice for fighting cyberbullying, giving users a protective and nourishing online social environment.

As for future improvements, multi-modal analysis options will be provided to cope with not only textual data but also images and videos in dealing with the increasingly visual cyberbullying. Moreover, we plan to investigate the applicability of transfer learning for increased cross-platform generalizability. We intend to develop real-time intervention systems that will be preventive as much as they will prevent bullying differently on the real-time, diverse electronic platforms.

References

[1] N. Amalina, M. Chinniah, A.A. Othman, P. Shamala, Cyberbullying: A Systematic Literature Review on the Definitional Criteria. International Journal of Academic Research in Business and Social Sciences, 12(3), (2022) 265-280. http://dx.doi.org/10.6007/IJARBSS/v12-i3/12042

[2] M.E. Kula, Cyberbullying: A Literature Review on Cross-Cultural Research in the Last Quarter. Handbook of Research on Digital Violence and Discrimination Studies, (2022) 610-630. https://doi.org/10.4018/978-1-7998-9187-1.ch027

[3] S. Grover, V.V. Raju, Cyberbullying: A Narrative Review. Journal of Mental Health and Human Behaviour, 28(1), (2023) 17-26. https://doi.org/10.4103/jmhhb.jmhhb_47_22

[4] D. Aizenkot, The predictability of routine activity theory for cyberbullying victimization among children and youth: Risk and protective factors. Journal of interpersonal violence, 37(13-14), (2022) NP11857-NP11882. https://doi.org/10.1177/0886260521997433

[5] S.M. Fati, A. Muneer, A. Alwadain, A.O. Balogun, Cyberbullying detection on twitter using deep learning-based attention mechanisms and

continuous Bag of words feature extraction. Mathematics, 11(16), (2023) 3567. https://doi.org/10.3390/math11163567

[6] L. Hebert, L. Golab, R. Cohen, (2022) Predicting Hateful Discussions on Reddit using Graph Transformer Networks and Communal Context. In 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), IEEE, Canada. https://doi.org/10.1109/WI-IAT55865.2022.00012

[7] M. Subramanian, V.E. Sathiskumar, G. Deepalakshmi, J. Cho, G. Manikandan, A survey on hate speech detection and sentiment analysis using machine learning and deep learning models. Alexandria Engineering Journal, 80, (2023) 110-121. https://doi.org/10.1016/j.aej.2023.08.038

[8] Q. Huang, V.K. Singh, P.K. Atrey, Cyber bullying detection using social and textual analysis. In Proceedings of the 3rd International Workshop on Socially-aware Multimedia, (2014) 3-6. https://doi.org/10.1145/2661126.2661133

[9] M. Di Capua, E. Di Nardo, A. Petrosino, (2016) Unsupervised cyber bullying detection in social networks. In 2016 23rd International conference on pattern recognition (ICPR), IEEE, Mexico. https://doi.org/10.1109/ICPR.2016.7899672

[10] N.A. Azeez, S.O. Idiakose, C.J. Onyema, C. Van Der Vyver, Cyberbullying detection in social networks: Artificial intelligence approach. Journal of Cyber Security and Mobility, (2021) 745-774. https://doi.org/10.13052/jcsm2245-1439.1046

[11] Desai, S. Kalaskar, O. Kumbhar, R. Dhumal, Cyber bullying detection on social media using machine learning. In ITM Web of Conferences, 40, (2021) 03038. https://doi.org/10.1051/itmconf/20214003038

[12] N. Balaji, B.H. Karthik Pai, K. Manjunath, B. Venkatesh, N. Bhavatarini, B.K. Sreenidhi, Cyberbullying in online/e-learning platforms based on social networks. In Intelligent Sustainable Systems: Selected Papers of WorldS4 2021, 2, (2022) 227-240. https://doi.org/10.1007/978-981-16-6369-7_20

[13] Bansal, A. Baliyan, A. Yadav, A. Kamlesh, H.K. Baranwal, Cyberbullying Detection on Social Networks Using Machine Learning Approaches. International Research Journal of Engineering and Technology (IRJET), 9, (2022). https://doi.org/10.1007/978-981-16-6369-7_20

[14] A.A. Albraikan, S.H. Hassine, S.M. Fati, F.N. Al-Wesabi, A.M. Hilal, A. Motwakel Hilal, A. Motwakel, M.A. Hamza, M. Al Duhayyim, Optimal deep learningbased cyberattack detection and classification technique on social networks. Computers, Materials & Continua, 72(1), (2022) 907-923. https://doi.org/10.32604/cmc.2022.024488

[15] Kumar, N. Sachdeva, A Bi-GRU with attention and CapsNet hybrid model for cyberbullying detection on social media. World Wide Web, 25(4), (2022) 1537-1550. https://doi.org/10.1007/s11280-021-00920-4

[16] B.A.H. Murshed, J. Abawajy, S. Mallappa, M.A.N. Saif, H.D.E. Al-Ariki, DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform. IEEE Access, 10, (2022) 25857-25871. https://doi.org/10.1109/ACCESS.2022.3153675

[17] M. López-Vizcaíno, F.J. Nóvoa, T. Artieres, F. Cacheda, Site agnostic approach to early detection of cyberbullying on social media networks. Sensors, 23(10), (2023) 4788. https://doi.org/10.3390/s23104788

[18] N. Ejaz, S. Choudhury, F. Razi, A comprehensive dataset for automated cyberbullying detection. Mendeley Data, 2, (2024) 2024.

[19] A. Willisch, A. Wolgast, M. Donat, (2022) Dataset for: Cyber Bullying among University Students [Data set]. Psych Archives. https://doi.org/10.23668/psycharchives.5597

[20] B. Alhijawi, A. Awajan, Genetic algorithms: Theory, genetic operators, solutions, and applications. Evolutionary Intelligence, 17(3), (2024) 1245-1256. https://doi.org/10.1007/s12065-023-00822-6

[21] W. Wang, T. Chen, W. Wu, (2023). Reinforcement Learning for Combating Cyberbullying in Online Social Networks. Combinatorial Optimization and Applications, Springer Nature Switzerland https://doi.org/10.1007/978-3-031-49614-1_36

[22] R.T. Icarte, T.Q. Klassen, R. Valenzano, S.A. McIlraith, Reward machines: Exploiting reward function structure in reinforcement learning. Journal of Artificial Intelligence Research, 73, (2022) 173-208. https://doi.org/10.1613/jair.1.12440

[23] L.G.C. Evangelista, R. Giusti, (2022) Short-and-Long-Term Impact of Initialization Functions in NeuroEvolution. Brazilian Conference on Intelligent Systems, Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-21686-2_21

[24] B.S. Panigrahi, R.K. Kanna, P.P. Das, S.K. Sahoo, T. Dutta, Machine Learning Based Intelligent Management System for Energy Storage Using Computing Application. EAI Endorsed Transactions on Energy Web, 11, (2024). https://doi.org/10.4108/ew.6272

[25] Y. Wang, Z. Xiao, G. Cao, A convolutional neural network method based on Adam optimizer with power-exponential learning rate for bearing fault diagnosis. Journal of Vibroengineering, 24(4), (2022) 666-678. https://doi.org/10.21595/jve.2022.22271

[26] Z. Khan, R.K. Kanna, K. Parthasarathy, S. Vijay raj, R. Chandrasekaran, S. Jawla, Intelligent computational ensemble model for predicting cerebral aneurysm using the concept of region localization in multi-section CT angiography. International Journal of Information Technology, (2025) 1-7. https://doi.org/10.1007/s41870-024-02292-0

[27] R.K. Kanna, A.O. Salau, New Cognitive Computational Strategy for Optimizing Brain Tumour Classification using Magnetic Resonance Imaging Data. Intelligence-Based Medicine, (2025) 100215. https://doi.org/10.1016/j.ibmed.2025.100215

[28] S. Umamaheswaran, R. John, S. Nagarajan, K.M. Karthick Raghunath, Arvind, K. S. (2022). Predictive Assessment of Fetus Features Using Scanned Image Segmentation Techniques and Deep Learning Strategy. International Journal of e-Collaboration (IJeC), 18(3), 1-13. https://doi.org/10.4018/IJeC.307130

Funding

The Authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

Has this article screened for similarity?

Yes

About the License