CRT.ORG

ISSN: 2320-2882



INTERNATIONAL JOURNAL OF CREATIVE **RESEARCH THOUGHTS (IJCRT)**

An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Cyberbullying Detection With Geo-Aware Transformers And Deep Learning-A **Comprehensive Survey**

¹Sushma A, ²Dr Deepa S.R ¹Assistant Professor, ²Professor and Head ¹Computer Science and Design ¹K.S Institute of Technology, Bengaluru, India

Abstract: This paper aims to develop an advanced system that detects and maps cyberbullying incidents on social media using AI technologies. The system will leverage Transformer techniques to identify relevant patterns and contextual signals from social media posts, accurately spotting instances of cyberbullying. Teachers can use the system to identify and support students affected by online harassment. Policymakers can benefit from precise data on cyberbullying trends and locations, enabling informed decision-making and the development of robust policies to combat online abuse. Social media platforms can utilize this technology to monitor and mitigate cyberbullying incidents, ensuring a safer digital environment for users.

Index Terms: Cyber bullying, BERT, FSSDL-CBDC, Deep learning, Machine learning, Transfer learning, **CNN**

Introduction

In the digital age, where social media, online communication, and technology play a crucial role in daily life, cyberbullying has become a major concern, especially among young people. Unlike traditional bullying, cyberbullying occurs through digital platforms such as social media, messaging apps, and online forums, making it more pervasive and difficult to control.

Cyberbullying can have severe psychological, emotional, and social consequences, leading to anxiety, depression, and even suicidal thoughts among victims. The anonymity provided by the internet often encourages perpetrators to engage in harmful behaviors without immediate consequences. Moreover, the widespread use of technology means that cyberbullying can occur anytime and anywhere, making it harder for victims to escape its effects. Given its long-term impact on mental health, self-esteem, and social interactions, addressing cyberbullying is crucial in the present generation. Efforts such as awareness campaigns, legal measures, parental guidance, and ethical digital education are essential to combat this issue. By fostering a culture of respect, empathy, and responsible online behaviour, we can create a safer digital environment for all.

I.LITERATURE REVIEW

Raju Kumar and Aruna Bhat [1] proposed study of Machine learning based models for detection, control and mitigation of cyber bullying in social media. The authors highlight the impact of online social media (OSM) on modern life, particularly among the younger generation. While OSM facilitates connectivity and idea-sharing, its rapid growth has also led to increased cybercrime, exposing users to online aggression and cyberbullying. The paper discusses cyberbullying, various machine and deep learning models used to address it, and the challenges in designing an effective prediction model.

This study reviewed machine learning (ML) and deep learning (DL) techniques for detecting and predicting cyberbullying and offensive content online. It explored different forms of cyberbullying, methods for dataset extraction, preprocessing, and classification. The study was divided into two sections based on data types and analyzed key factors to build effective classifiers. Additionally, it addressed challenges in cyberbullying detection and provided insights for developing improved models. The findings serve as a valuable resource for researchers aiming to advance cyberbullying detection systems.

Aljawharah Abdelwahab, Mohd Anul Haq, and Mohammed Alshehri [2] developed and proposed a system for Cyber Bullying Detection on social media using Machine Learning.

The authors present their experience in the use of Natural Language Processing (NLP) and machine learning (ML) techniques to detect cyberbullying on social media. Various ML models, including K- Nearest Neighbour (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Decision Trees (DT), and Random Forest (RF), were employed to analyze cyberbullying-related features. The study highlights the effectiveness of these models in predicting and addressing cyberbullying behavior, emphasizing their accuracy and impact on social networks. The results indicate that deep learning achieved the highest accuracy (0.96), followed by SVM (0.92) and KNN (0.90).

The study adopts a comprehensive methodology for detecting cyberbullying on social media platforms, specifically Twitter, utilizing Natural Language Processing (NLP), Machine Learning (ML), and Deep Learning (DL) techniques. The dataset comprises 47,692 tweets, each labelled as either cyberbullying or non-cyberbullying. The preprocessing phase involves cleaning the dataset by removing duplicates, emojis, links, special characters, and stop words. Additionally, stemming is applied to reduce words to their base forms, and hashtags are refined by removing the '#' symbol while preserving the core content. For feature extraction, the researchers employ the Count Vectorizer function from the Sklenar package, with a cap of 2,500 features. The dataset is then divided into training and testing sets using a 75:25 split, and features are scaled accordingly using a standard scaler.

The study evaluates two traditional ML models—K-Nearest Neighbour (KNN) and Support Vector Machine (SVM)—and compares them to a more advanced deep learning model. KNN utilizes the Euclidean distance metric to classify data points based on proximity, while SVM employs a hyperplane to separate data points in high- dimensional space. Both models yield strong results, with KNN achieving 90% accuracy and SVM slightly outperforming it at 92%. The deep learning model, however, stands out due to its complexity and effectiveness. The researchers design a six-layer neural architecture based on a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. This model includes an embedding layer for vectorizing input text, a convolutional layer to identify local features, a max pooling layer for dimensionality reduction, and dense layers for classification using ReLU and SoftMax activation functions. The model is compiled with the Adam optimizer and trained on a GPU-enabled system for enhanced performance.

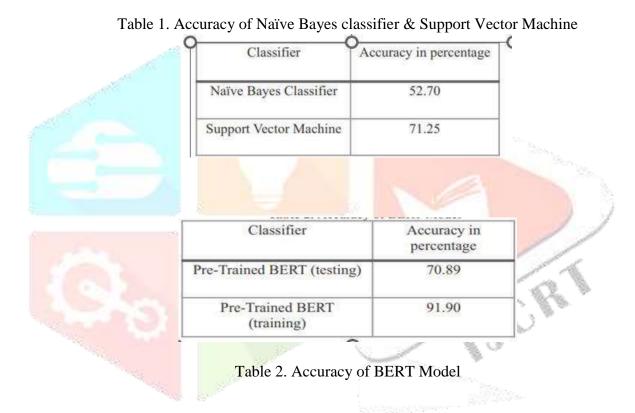
The findings of the study highlight the superior performance of deep learning models in the task of cyberbullying detection on social media. Among the three models tested, the CNN-LSTM deep learning framework achieved the highest accuracy of 96%, surpassing SVM (92%) and KNN (90%). This demonstrates the model's strong ability to recognize complex patterns in textual data that often signify bullying behaviour. The authors underline that while ML models like KNN and SVM are effective, deep learning provides a more robust and scalable solution, especially when dealing with large and nuanced datasets. The study concludes that incorporating advanced neural architectures, along with sufficient preprocessing and feature extraction, can significantly enhance the accuracy and reliability of cyberbullying detection systems. Looking forward, the authors suggest extending this research by incorporating larger,

multilingual datasets and exploring real-time deployment scenarios. They also propose integrating more sophisticated techniques like transfer learning and self-attention mechanisms to further improve detection performance and adaptability across different social platforms.

Aditya Desai, Shashank Kalaskar, Omkar Kumbhar , and Rashmi Dhumal [3] explores the importance of Cyber Bullying Detection on social media using Machine Learning in their design and development.

Cyberbullying on social media has led to serious mental health issues, particularly among young people, increasing low self-esteem and suicidal ideation. Existing machine learning models for cyberbullying detection lack essential features for accurate classification. This paper proposes an improved model using various features and a bidirectional deep learning approach with BERT to enhance cyberbullying detection.

Significance of using BERT:A semi-supervised cyberbullying detection approach using BERT and five key features achieved 91.90% accuracy, outperforming traditional models. Larger datasets can further improve accuracy. Future work includes developing an application and integrating additional models for enhanced detection.



The above Table 1 and 2 illustrates the comparison of BERT significantly outperforming Naïve Bayes, with a testing accuracy of 70.89% compared to 52.70%. While SVM slightly surpasses BERT in testing (71.25% vs. 70.89%), BERT shows much higher training accuracy (91.90%), indicating potential overfitting.

Neelakandan S, Sridevi M, Saravanan Chandrasekaran, Murugeswari K, Aditya Kumar Singh Pundir , Sridevi R, and T.Bheema Lingaiah [4] designed and developed a system using Deep Learning Approaches for Cyberbullying Detection and

Classification on Social Media using FSSDL- CBDC.

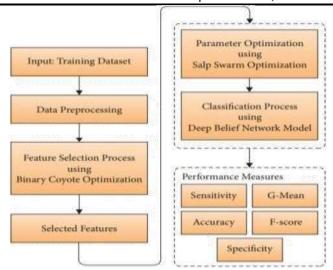


Figure 1. Working Process of FSSDL-CBDC Model

The above Figure 1 shows the working process of FSSDL-CBDC technique which integrates feature subset selection with deep learning for improved CB detection. It employs binary covote optimization (BCO) for feature selection and the salp swarm algorithm (SSA) with a deep belief network (DBN) for classification, forming the SSA-DBN model. Experimental results show that SSA-DBN achieves 99.983% accuracy, outperforming other approaches, highlighting the effectiveness of FSSDL-CBDC in CB detection.

This paper derives a conclusion that FSSDL-CBDC technique effectively detects and classifies cyberbullying on social media by integrating preprocessing, feature selection, and classification. The BCO-FSS approach optimizes feature selection, enhancing classification accuracy. The SSA-DBN model, which fine-tunes hyperparameters, outperforms the classic DBN model. Simulations confirm its superior performance compared to other methods. Future improvements may include outlier detection and feature reduction, benefiting fields like intrusion detection and network security.

Teoh Hwai Tengi and Kasturi Dewi Varathan [5] proposes a system on Cyberbullying Detection in Social Networks and provided a clear Comparison between Machine Learning and Transfer Learning Approaches

This research focuses on detecting cyberbullying using automated methods instead of manual approaches like reporting or blocking. Two techniques were used: Conventional Machine Learning (CML) and Transfer Learning. The study utilized AMiCA data, incorporating various features like textual, sentiment, emotional, and toxicity indicators. It was the first to use psycholinguistic tools like LIWC 2022 and Empath's lexicon for cyberbullying detection.

Among different language models tested, Distil BERT performed best, especially when combined with textual and toxicity features, achieving a 72.42% F-measure, outperforming other models. Transfer Learning proved to be the most effective, requiring less effort and yielding better results than traditional methods.

This research explores cyberbullying detection using both conventional machine learning and transfer learning approaches. Conventional methods, such as Logistic Regression and Linear SVC, benefited from feature engineering involving textual, sentiment, emotional, psycholinguistic, and toxicity features. Logistic Regression outperformed Linear SVC, especially when handling high-dimensional data.

Sreenivas Mekala, Y.Sai Nithin, B.Guru Raghav, T.Sai Kumar[6] proposed system integrates both hardware and software components to optimize street lighting and detect cyberbullying. For automatic streetlight control, an LDR sensor detects ambient light to distinguish between day and night, while IR sensors monitor vehicle movement to activate lights based on road activity.

In the cyberbullying detection framework, Natural Language Processing (NLP) is used to preprocess text using techniques like Bag-of-Words and TF-IDF. Then, SMOTE is applied to balance the dataset, and XGBoost is used for effective classification.

The BERT model further enhances detection by processing inputs in embedded form using token, position, and segment embedding. She The dataset consists of over 1000 tweets collected via Twitter API. Data cleaning includes conversion from .xlsx to .csv, removal of noise, and normalization to prepare for model training. The model is trained, validated, and fine-tuned iteratively to improve prediction accuracy.

The testing phase confirmed the effectiveness of the proposed model in detecting cyberbullying. Tweets containing bullying indicators were correctly categorized, as shown through the confusion matrix. When compared to traditional models like SVM (71.25%) and Naive Bayes (52.7%), the BERT-based model demonstrated significantly higher accuracy at 91.90%, indicating its superior performance in sentiment analysis and cyberbullying detection on Twitter data.

The study emphasizes the growing threat of cyberbullying due to increased social media usage, especially among teenagers. It explores the automated detection of bullying-related content using Bag-of-Words (BoW) and TF-IDF features, applying four machine learning methods, including SVM. The research highlights the potential of such systems to mitigate cyber harassment and expresses the intent to extend the approach using deep learning for analyzing Bengali literature in the future.

Pradeep Kumar Roy, Fenish Umeshbhai Mali[7] explores the relatively underexplored area of image-based cyberbullying detection, diverging from the more commonly researched text-based approaches.

It introduces a custom dataset comprising 3,000 manually annotated images, sourced from Google and public datasets like MMHS150K, with labels indicating bullying or non-bullying content. The figure 2 below shows the application of advanced artificial neural network CNN for image classification. The images were resized and standardized to match the input dimensions of pre-trained models such as VGG16 and InceptionV3. The modelling strategies included a custom CNN model with one to three layers and transfer learning approaches using VGG16 and InceptionV3. Various configurations were tested using SGD and Adam optimizers at different learning rates, along with dropout and early stopping to prevent overfitting. Performance was evaluated using accuracy, precision, recall, F1-score, confusion matrix, and AUC. Among all models, InceptionV3 achieved the best results, reaching 89% accuracy and an F1-score of 0.89, significantly outperforming traditional models and shallow CNNs.

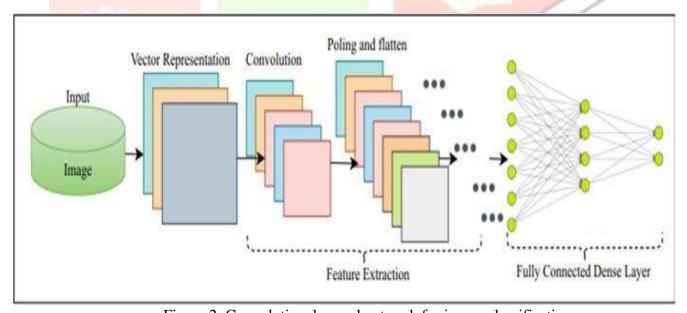


Figure 2: Convolutional neural network for image classification

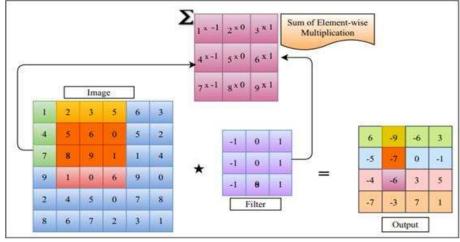


Figure 3: Output matrix formation with convolution operation

The above Figure 3 illustrates the convolution operation where a filter is applied to an image matrix to produce an output matrix through element-wise multiplication and summation.

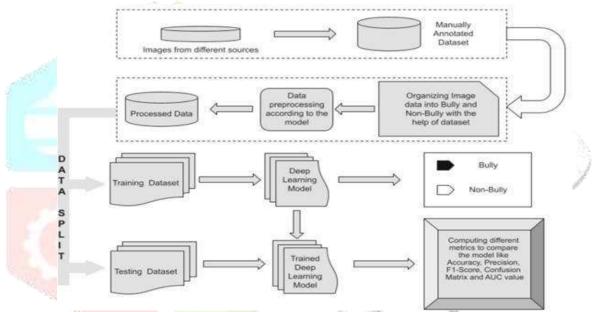


Figure 4: Proposed system design to detect bully posts

The above Figure 4 shows the methodology of detecting bully posts in the proposed system.

While the study makes a notable contribution by pioneering image-based cyberbullying detection, it also has a few limitations. The dataset size, although balanced, is relatively small and may not capture the full variability of real- world content. Manual labelling, even with three reviewers, introduces the potential for subjectivity. Moreover, the absence of real-time validation restricts the practical applicability of the model in dynamic social media contexts. The exclusive focus on image content, without considering accompanying text, narrows the model's detection capability since many harmful posts combine both modalities.

Lastly, while the study shows strong results with VGG16 and InceptionV3, it overlooks more advanced deep learning architectures like EfficientNet and Vision Transformers, which could potentially improve performance further. An integrated, multimodal approach with a larger dataset and broader model comparisons would be valuable future directions.

II. INFERENCES

This literature survey and study highlights the crucial role of advanced machine learning and deep learning techniques in detecting cyberbullying, given its severe psychological impact, particularly on children. The evaluation of various methods demonstrated that deep learning models, such as BERT and DistilBERT,

outperform traditional approaches like Logistic Regression and SVM, primarily due to their ability to learn contextual and semantic information from text without requiring extensive feature engineering. While conventional models benefit from integrating textual, sentiment, and toxicity features, transformer-based models excel in handling complex, imbalanced datasets with higher accuracy and efficiency.

Additionally, feature selection and hyperparameter tuning techniques, such as BCO-FSS and SSA-DBN, have shown promising results in optimizing classification performance. The effectiveness of these models highlights the growing need for automated, scalable cyberbullying detection systems that can operate in real time and adapt to evolving online threats.

The researchers explores cyberbullying detection using both conventional machine learning and transfer learning approaches. Conventional methods, such as Logistic Regression and Linear SVC, benefited from feature engineering involving textual, sentiment, emotional, psycholinguistic, and toxicity features. Logistic Regression outperformed Linear SVC, especially when handling high-dimensional data.

Transfer learning using fine-tuned transformers (DistilBert, DistilRoBerta, and Electra-small) proved more effective, with DistilBert achieving the highest F-measure (72.42%). Transfer learning eliminated the need for extensive feature engineering and worked well with smaller, imbalanced datasets. Overall, this study highlights the advantages of incorporating toxicity features in traditional models while demonstrating that fine- tuned transformer models significantly outperform conventional methods in cyberbullying detection.

The study highlights the significance of detecting cyberbullying due to its psychological impact, especially on children. The study evaluates different methods for cyberbullying detection, with deep learning achieving the highest accuracy (0.96), followed by SVM (0.92) and KNN (0.90). The findings emphasize the importance of using reliable methods to mitigate cyberbullying and reduce its spread across social media platforms.

"Cyberbullying Detection Using Machine Learning" is that while traditional machine learning models like SVM and Naive Bayes provide a foundational approach to detecting harmful content, the use of advanced deep learning models—specifically BERT—demonstrates a significant leap in performance. With an achieved accuracy of 91.90% using BERT on Twitter data, the study confirms that contextual understanding through deep learning offers far superior results in identifying cyberbullying. This underscores the potential of integrating sentiment analysis and transfer learning into real-time, automated detection systems to more effectively combat online abuse.

"Cyberbullying detection using deep transfer learning" is that transfer learning models significantly enhance the ability to detect image-based cyberbullying on social media platforms compared to traditional CNN approaches. Among the evaluated models, InceptionV3 consistently outperformed others, achieving an impressive 89% accuracy and a high F1-score, confirming its robustness in handling complex, imbalanced image data. The study reinforces the necessity of using large, annotated datasets and advanced deep learning architectures for effective cyberbullying detection and underscores the practical viability of scalable, automated systems to foster safer digital environments.

Advantages drawn on using BERT:A semi-supervised cyberbullying detection approach using BERT and five key features achieved 91.90% accuracy, outperforming traditional models. Larger datasets can further improve accuracy. Future work includes developing an application and integrating additional models for enhanced detection.

Significance of using the FSSDL-CBDC technique effectively detects and classifies cyberbullying on social media by integrating preprocessing, feature selection, and classification. The BCO-FSS approach optimizes feature selection, enhancing classification accuracy. The SSA-DBN model, which fine-tunes hyperparameters, outperforms the classic DBN model. Simulations confirm its superior performance compared to other methods.

Table 3: Comparative Review of Machine Learning and Deep Transfer Learning Approaches for Cyberbullying Detection

Paper Title	Author/Journal	Summary
Cyberbullying in online social media	Raju Kumar, Aruna Bhat (International Journal of Information Security)	This paper provides a comprehensive review of machine learning and deep learning techniques for detecting and predicting cyberbullying on social media, and highlights the key challenges faced by researchers in this field.
Cyber Bullying Detection on social media using Machine Learning.	Aljawharah Abdelwahab , Mohd Anul Haq, and Mohammed Alshehri	The findings of the study highlight the superior performance of deep learning models in the task of cyberbullying detection on social media. Among the three models tested, the CNN-LSTM deep learning framework achieved the highest accuracy of 96%, surpassing SVM (92%) and KNN (90%)
Cyber Bullying Detection on social media using Machine Learning	Aditya Desai, Shashank Kal <mark>as</mark> kar, Omkar Kumbhar, Ras <mark>hmi Dhum</mark> al, Student	The paper proposes a cyberbullying detection system using a BERT- based machine learning model that achieves 91.90% accuracy on a Twitter dataset, outperforming traditional ML models.
Deep Learning Approaches for Cyberbullying Detection and Classification on social media using FSSDL- CBDC.	Neelakandan S, Sridevi M, Saravanan Chandrasekaran, Murugeswari K, Aditya Kumar Singh Pundir , Sridevi R, and T.Bheema Lingaiah	The Paper proposes FSSDL-CBDC technique effectively detects and classifies cyberbullying on social media by integrating preprocessing, feature selection, and classification. The BCO-FSS approach optimizes feature selection, enhancing classification accuracy. The SSA-DBN model, which fine-tunes hyperparameters, outperforms the classic DBN model. Simulations confirm its superior performance compared to other methods.
Cyberbullying Detection in Social Networks: A Comparison Between Machine Learning and Transfer Learning Approaches	Teoh Hwai, Kasturi Dewi Varathan, Dewi Kasturi, Varathan (IEEE Access)	This paper develops an automatic system for cyberbullying detection using conventional machine learning and transfer learning approaches, with the goal of creating classification models that can effectively detect cyberbullying and non-cyberbullying posts on social media.
Cyber bullying detection using Machine learning	Sreenivas Mekala, Y.Sai Nithin, B.Guru Raghav, T.Sai Kumar	This paper provides a review on cyberbullying that testing phase confirmed the effectiveness of the proposed model in detecting cyberbullying. Tweets containing bullying indicators were correctly categorized, as shown through the confusion matrix. When compared to traditional models like SVM (71.25%) and Naive Bayes (52.7%), the BERT-based model demonstrated significantly higher accuracy at 91.90%, indicating its superior performance in sentiment analysis and cyberbullying detection on Twitter data.

Paper Title	Author/Journal	Summary
Cyberbullying detection using deep transfer learning	Pradeep Kumar Roy · Fenish Umeshbhai Mali	This paper study shows strong results with VGG16 and InceptionV3, it overlooks more advanced deep learning architectures like EfficientNet and Vision Transformers, which could potentially improve performance further. An integrated, multimodal approach with a larger dataset and broader model comparisons would be valuable future directions.

III. **FUTURE WORK**

Future research should focus on developing real- world applications incorporating these models for realtime cyberbullying detection. Enhancements such as outlier detection, feature reduction, and hyperparameter optimization can further improve model accuracy.

Additionally, integrating multiple deep learning models and exploring more extensive datasets will help enhance detection capabilities. Another promising avenue is expanding the scope beyond text-based detection by incorporating multimedia elements like images and videos, ensuring a more comprehensive cyberbullying prevention system. Future advancements should focus on integrating multimodal data, such as images and videos, and improving detection models to reduce false positives, ensuring a safer and more inclusive digital environment.

Develop and optimize a feature extraction module using Transformer techniques to identify and select relevant features from social media posts that are indicative of cyberbullying, incorporating textual and contextual signals.

Build and train two distinct deep learning models for the classification of cyberbullying and noncyberbullying content. Evaluate their performance in terms of accuracy, precision, recall, and F1-score to determine their effectiveness in identifying cyberbullying incidents.

Combine the outputs of the two deep learning models using an ensemble technique to improve overall detection accuracy and robustness. Assess the ensemble model's performance against individual models to demonstrate its efficacy in cyberbullying detection with geolocation context.

Incorporate geolocation data into the cyberbullying detection system to enable the mapping of incidents on a geo map. Analyze regional trends and patterns of cyberbullying, providing insights that can guide targeted intervention and prevention strategies.

REFERENCES IV.

[1] Raju Kumar1 · Aruna Bhat1 – "A study of machine learning-based models for detection, control, and mitigation of cyberbullying in online social media"

https://www.researchgate.net/publication/362566671 A study of machine learningbased models for detection control and mitigation of cyberbullying in online social media

- [2]Aljwharah Alabdulwahab1, Mohd Anul Haq2*, Mohammed Alshehri3 Department of Information Technology, College of Computer and Information Sciences, Majmmah University, Al-Majmaah, Saudia Arabia.1, 3 Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al-Majmaah,
- Arabhttps://www.researchgate.net/publication/375120231_Cyberbullying_Detection_using_Machine_Lear ning_and_Deep_Learning
- [3] Aditya Desai1, Shashank Kalaskar2, Omkar Kumbhar3, and Rashmi Dhumal4https://www.itmconferences.org/articles/itmconf/abs/2021/05/itmconf_icacc2021_03038/itmconf_icacc2021_03038.html
- [4] Neelakandan S,1 Sridevi M,2 Saravanan Chandrasekaran,3 Murugeswari K,4 Aditya Kumar Singh Pundir, 5 Sridevi R,6 and T.Bheema Lingaiah 7

https://onlinelibrary.wiley.com/doi/10.1155/2022/2 163458

- [5] TEOH HWAI TENG AND KASTURI DEWI VARATHAN IEEE ACCESS Volume11 Page 55533-55560.DOI10.1109/ACCESS.2023.3275130https://orcid.org/0000-0003-3421-4501
- Sreenivas Mekala. 2Y.Sai Nithin. 2B.Guru [6]1Dr. Raghav, 2T. Sai Kumar https://www.researchgate.net/publication/371760490
- [7] Pradeep Kumar Roy1 · Fenish Umeshbhai Mali1 https://orcid.org/0000-0001-5513-2834Volume11 Page 55533- 55560.DOI10.1109/ACCESS.2023.3275130 https://orcid.org/0000-0003-3421-4501
- [8] Yu X, Yang T, Lu J, Shen Y, Lu W, Zhu W, Bao Y, Li H, Zhou J (2021) Deep transfer learning: a novel glucose prediction framework for new subjects with type 2 diabetes. Complex Intell Syst.mplex Intell Syst. https://doi.org/10.1007/s40747-021-00360-7
- [9] Li S, Liu B, Li S, Zhu X, Yan Y, Zhang D (2021) A deep learning-based computer-aided diagnosis method of x-ray images for bone age assessment. Complex Intell Syst. https://doi.org/10. 1007/s40747-021-00376-z
- [10] Xue G, Liu S, Ma Y (2020) A hybrid deep learning-based fruit classification using attention model and convolution autoencoder. Complex Intell Syst. https://doi.org/10.1007/s40747-020-00192-x
- [11] "Cyberbullying Statistics, Facts and Trend. [online] https://frstsiteguide.com/cyberbullying-stats/ (2020)
- [12] S. Mahbub, E. Pardede, and A. S. M. Kayes, "Detection of harassment type of cyberbullying: A dictionary of approach words and its impact,' Secur. Commun. Netw., vol. 2021, pp. 1–12, Jun. 2021, doi: 10.1155/2021/5594175.https://onlinelibrary.wiley.com/doi/10.1155/2021/5594175
- [13] A. Bozyiğit, S. Utku, and E. Nasibov, "Cyberbullying detection: Utilizing social media features," **Expert** Appl., vol. 179, Oct. 2021, Art. no. 10.1016/j.eswa.2021.115001https://www.sciencedirect.com/science/article/abs/pii/S0957417421004425?vi a%3Dihub
 - [14] "Cyberbullying Statistics", 2019, [online] Available: https://enough.org/stats_cyberbullying