PUXplore: Multidisciplinary Journal of Engineering, Vol 2 , Issue 1, Jan - Mar 2026 || Online (ISSN 3108-2106)

Cyberbullying Detection Using Machine Learning:A Comprehensive Review

Ashwini Kumar Jha[1], Avinash Songa[2], Vishnu Ulliboyina[3], Pavan Kumar Chadaram[4], Rohith Purushottam Naidu[5]

[1] *Associate Professor, Department of Artificial Intelligence & Data Science, Parul University, India*

[2 to 5] *Student, Department of Computer Science & Engineering, Parul University, India*

E-mail: [1]ashwini.jha34918@paruluniversity.ac.in, [2]2303031240145@paruluniversity.ac.in, [3]2303031240154@paruluniversity.ac.in,[4]2303031240296@paruluniversity.ac.in, [5]2303031240575@paruluniversity.ac.in

ORCID: [1]0000-0002-6607-2168, [2]0009-0003-1206-2479

***Abstract* - The rapid growth of social media platforms and online communication technologies has increased the scale of digital interactions while simultaneously introducing serious challenges related to harmful online behavior such as cyberbullying. Cyberbullying involves the use of digital platforms to harass, threaten, or humiliate individuals, often causing significant psychological and emotional harm. With the massive volume of user-generated content shared across social networking sites, detecting such harmful behavior through manual moderation has become increasingly difficult. As a result, machine learning techniques have emerged as an effective approach for automatically identifying cyberbullying in online environments. This study applies a systematic literature review (SLR) approach to examine research published over the past decade (2015–2025) on cyberbullying detection using machine learning methods. The review analyzes existing studies to understand commonly used algorithms, datasets, feature extraction techniques, and evaluation strategies applied in cyberbullying detection systems. It also highlights the role of natural language processing and deep learning approaches in improving the identification of abusive online content. In addition, the study discusses key challenges associated with cyberbullying detection, including contextual language understanding, sarcasm interpretation, dataset imbalance, and multilingual communication. This comprehensive analysis aims to provide researchers and practitioners with a clearer understanding of current developments and potential future directions in automated cyberbullying detection for safer online environments.**

Keywords: Cyberbullying Detection, Machine Learning, Natural Language Processing, Online Harassment Detection, Social Media Analysis, Text Classification, Deep Learning, Automated Content Moderation

I. INTRODUCTION

Online communication has become an integral part of modern society as digital platforms continue to connect millions of users across the world. Social media networks, discussion forums, and messaging applications allow people to exchange ideas, express opinions, and interact instantly regardless of geographical distance[1]. Although these platforms offer many advantages for communication and information sharing, they have also created new opportunities for harmful online behavior. One of the most serious problems that has emerged within digital communities is cyberbullying, which has become a growing concern for researchers, educators, and policymakers[2].Cyberbullying refers to the act of using digital technologies such as social media platforms, messaging services, and online forums to intentionally harass, threaten, or humiliate individuals[5]. Unlike traditional forms of bullying, cyberbullying can occur continuously and may spread rapidly through online networks, reaching a large audience within a short period of time. Victims of cyberbullying often experience severe emotional and psychological consequences including anxiety, depression, social withdrawal, and reduced self-confidence[6]. In many cases, young users and adolescents are particularly vulnerable because they spend a significant amount of time engaging with social media platforms.The rapid increase in online interactions has also resulted in an enormous amount of user-generated content being produced every day. Millions of posts, comments, and messages are uploaded to social networking platforms every minute, making it extremely difficult for human moderators to manually monitor and filter harmful content[9]. As the scale of online communication continues to grow, traditional moderation approaches become inefficient and insufficient for effectively identifying abusive behavior. This situation highlights the need for automated solutions capable of detecting cyberbullying content in large-scale online environments.In recent years, machine learning techniques have gained significant attention as an effective approach for addressing the problem of cyberbullying detection. Machine learning models are capable of analyzing large volumes of textual data and identifying patterns associated with abusive language or aggressive behavior. By learning from previously labeled datasets, these models can automatically classify online messages as bullying or non-bullying content. Various algorithms such as Naïve Bayes, Support Vector Machines, Decision Trees, Random Forest, and deep learning models have been explored to improve the accuracy of cyberbullying detection systems[29].Another important component in cyberbullying detection is Natural Language Processing (NLP), which enables computers to analyze and interpret human language present in online messages[13,14]. NLP techniques are widely used to extract meaningful

features from textual data, including sentiment, context, and semantic relationships between words. Methods such as word embeddings, sentiment analysis, and contextual language models have significantly improved the ability of machine learning systems to understand online communication.Despite the progress made in this research area, detecting cyberbullying remains a complex challenge[34]. Online communication often includes slang expressions, sarcasm, abbreviations, emojis, and cultural references that can be difficult for automated systems to interpret correctly. In addition, harmful behavior may sometimes appear indirectly through subtle language or contextual cues rather than explicit abusive words. These challenges make it necessary to develop more advanced models that can understand the context and intent behind online messages.Due to the increasing awareness of online harassment and its potential impact on mental health, research related to cyberbullying detection has grown rapidly over the past decade[30]. Numerous studies have proposed different machine learning frameworks, datasets, and feature extraction techniques for identifying harmful online interactions. However, the effectiveness of these approaches varies depending on the dataset, algorithm, and contextual understanding of the model.

The purpose of this review paper is to examine and summarize existing research on cyberbullying detection using machine learning techniques. The study analyzes different detection methods, commonly used datasets, and evaluation techniques proposed in previous research. It also discusses the key challenges faced in developing accurate cyberbullying detection systems and highlights potential directions for future research aimed at improving automated moderation technologies.

II. PROCEDURE AND ANALYSIS

2.1 Data Collection

The data collection process for this review focused on identifying relevant research studies related to cyberbullying detection and the application of machine learning techniques in online safety systems. Various forms of academic publications including journal articles, conference papers, technical reports, and review papers were collected from well-known digital research databases such as IEEE Xplore, SpringerLink, ScienceDirect, ACM Digital Library, and Google Scholar. These sources were selected because they provide high-quality peer-reviewed research in the fields of computer science, artificial intelligence, and cybersecurity.

2.2 Inclusion and Exclusion Criteria

After the initial collection of studies from the selected databases, a structured screening process was conducted to ensure the relevance, quality, and reliability of the literature included in this review.Studies were included if they explicitly discussed cyberbullying detection, machine learning models for identifying harmful online behavior, or related concepts such as natural language processing, sentiment analysis, abusive language detection, and automated content moderation in social media platforms. Research proposing detection frameworks, classification models, performance evaluations, or domain-specific applications of cyberbullying detection in online environments such as social networking platforms, discussion forums, messaging systems, and educational platforms was considered eligible. Only peer-reviewed journal articles, conference papers, and review studies published within the selected time frame were included to maintain academic credibility and relevance. Additionally, the selected studies were required to be written in English and available in full-text form to allow comprehensive evaluation.Conversely, duplicate records identified across databases were removed to avoid redundancy. Articles published outside the selected time range or studies focusing mainly on general social media analytics without addressing cyberbullying detection were excluded. Non-peer-reviewed materials including blogs, editorials, opinion articles, and informal technical reports lacking sufficient methodological detail were also omitted. Furthermore, studies that mentioned cyberbullying only briefly without presenting detailed detection methods, algorithmic analysis, or experimental validation were excluded during both abstract and full-text screening. This systematic filtering process ensured that the final set of selected studies remained closely aligned with the objectives of the review while maintaining methodological consistency and research quality.

2.3 Analysis Methodology

The selected studies were analyzed using a qualitative review methodology aimed at systematically synthesizing existing knowledge on cyberbullying detection using machine learning techniques. Each paper was carefully examined to extract and evaluate its key contributions, including proposed detection models, feature extraction methods, classification algorithms, and practical implementation strategies. Particular attention was given to commonly used approaches such as Natural Language Processing (NLP), sentiment analysis, text representation techniques, and machine learning

algorithms including Support Vector Machines, Naïve Bayes, Decision Trees, Random Forest, and deep learning models[18,19]. The analysis also considered how contextual language understanding, behavioral patterns, and linguistic features were incorporated into cyberbullying detection frameworks across different online environments[36,39].To maintain consistency in the review process, a thematic analysis approach was adopted. The reviewed studies were grouped into categories based on data sources (such as social media platforms, discussion forums, and messaging applications), machine learning techniques applied, and the type of datasets used for model training and evaluation[29]. For each study, relevant information was extracted regarding research objectives, methodology, dataset characteristics, experimental design, validation techniques, and reported performance metrics. This structured comparison enabled the identification of common methodological patterns, widely adopted algorithms, and frequently reported challenges such as dataset imbalance, difficulty in detecting sarcasm or implicit harassment, and limitations in multilingual text analysis[19].Furthermore, the analysis examined both the strengths and limitations of the cyberbullying detection approaches discussed in the literature. Reported advantages such as improved automated moderation, enhanced text classification accuracy, and the ability to process large volumes of online data were contrasted with practical challenges including model generalization issues, computational complexity, and the difficulty of interpreting contextual meaning in online communication[39]. By synthesizing findings across multiple studies, this methodology enabled the identification of emerging research trends, technological limitations, and potential directions for improving automated cyberbullying detection systems. Overall, the qualitative analysis provided a comprehensive understanding of how machine learning techniques are applied to identify harmful online behavior across various digital platforms, thereby supporting the objectives of this review.

2.4 The Knowledge Gap

Although cyberbullying detection has been widely discussed in academic research and technology-driven moderation systems, several significant research gaps still remain. A large portion of existing studies primarily focus on developing classification models and theoretical detection frameworks, while comparatively fewer works provide comprehensive insights into real-world deployment scenarios on large-scale social media platforms[34]. In particular, practical challenges associated with implementing automated cyberbullying detection systems in dynamic online environments are not sufficiently explored. Social media platforms operate with massive volumes of user-generated content, diverse communication styles, and rapidly changing language patterns, which makes the practical deployment of detection systems more complex than what is typically observed in controlled experimental settings[39].Another important limitation involves dataset availability and diversity. Many studies rely on relatively small or platform-specific datasets that may not accurately represent the wide range of communication styles present in global online communities[38]. As a result, machine learning models trained on limited datasets may struggle to generalize effectively when applied to different platforms, languages, or cultural contexts. Furthermore, the imbalance between bullying and non-bullying examples within many datasets can significantly affect model performance and lead to biased predictions.Performance evaluation and scalability also represent important gaps in the current literature. While numerous studies report improvements in detection accuracy using machine learning and deep learning models, fewer works provide detailed analysis of computational requirements, processing efficiency, or scalability when applied to high-volume real-time social media streams[36]. As automated detection systems must process millions of posts and comments continuously, understanding their impact on system performance and resource consumption requires deeper investigation.Finally, there is a lack of standardized evaluation frameworks for comparing cyberbullying detection models across different studies. Although many researchers report metrics such as accuracy, precision, recall, and F1-score, variations in datasets and experimental settings make it difficult to perform consistent comparisons between proposed models. Addressing these research gaps through larger datasets, improved contextual language models, standardized evaluation practices, and real-world platform testing will be essential for advancing the effectiveness and practical deployment of automated cyberbullying detection systems in modern digital environments[41].

III. LITERATURE REVIEW

3.1 Traditional Machine Learning Methods

Traditional machine learning techniques have been widely used for detecting cyberbullying in social media platforms. These methods rely on extracting meaningful features from textual data and using classification algorithms to identify harmful or abusive messages[29]. Early research focused on algorithms such as Support Vector Machines (SVM), Naïve

Bayes (NB), Decision Trees (DT), Random Forest (RF), K-Nearest Neighbors (KNN), and Logistic Regression (LR).These algorithms typically use text preprocessing techniques such as tokenization, stop-word removal, stemming, and lemmatization to clean the input data. Feature extraction methods such as Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and n-grams are commonly applied to convert textual content into numerical representations that machine learning models can process[31].Several studies have shown that SVM and Random Forest classifiers achieve high accuracy in detecting cyberbullying due to their ability to handle high-dimensional textual features. Naïve Bayes classifiers are also widely used due to their simplicity and efficiency in large-scale text classification problems[38]. These traditional machine learning methods form the foundation for many cyberbullying detection systems used in research today.

3.2 Deep Learning Approaches

In recent years, deep learning techniques have gained significant attention for cyberbullying detection due to their ability to automatically learn complex patterns from large textual datasets[31]. Unlike traditional machine learning models that rely heavily on manual feature extraction, deep learning models can learn semantic and contextual relationships directly from raw text data.Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN), particularly Long Short-Term Memory (LSTM) networks, are widely used for detecting cyberbullying in social media posts[37]. CNN models are effective in extracting important local features from text, while LSTM networks are capable of capturing long-term dependencies and contextual information within sentences.More recently, transformer-based models such as Bidirectional Encoder Representations from Transformers (BERT) and RoBERTa have been used to improve the performance of cyberbullying detection systems. These models utilize pre-trained language representations to understand complex linguistic patterns, sarcasm, and offensive language commonly used in online communication[39]. As a result, deep learning approaches generally achieve higher accuracy compared to traditional machine learning methods in detecting cyberbullying on social media platforms.Furthermore, researchers have explored multimodal cyberbullying detection, where deep learning models analyze not only textual content but also images, videos, and emojis shared on social media[33].This approach is particularly useful because cyberbullying may occur through memes, images, or multimedia posts rather than just plain text.Multimodal deep learning models can therefore provide a more comprehensive understanding of harmful online behavior.

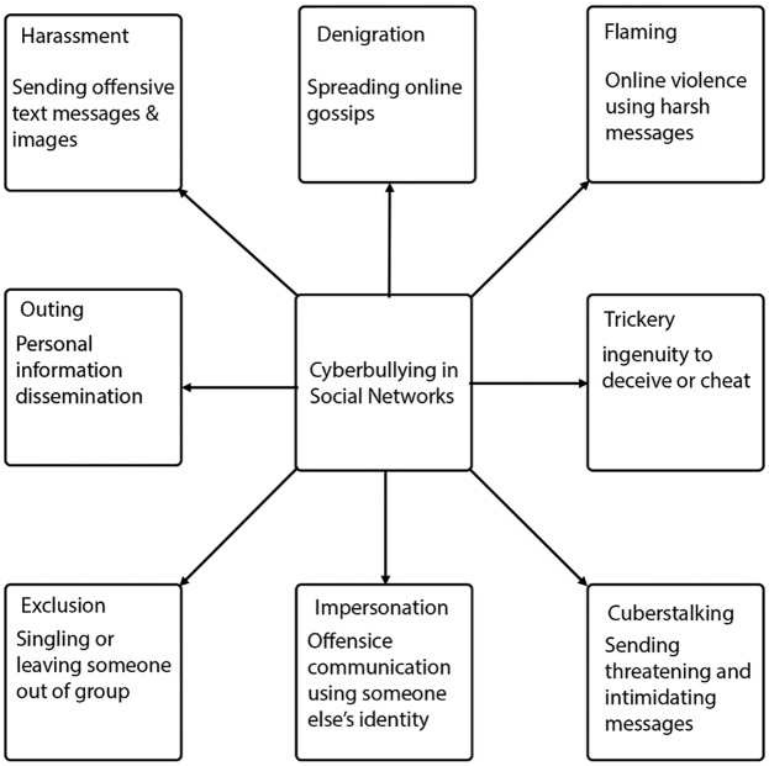

Fig 1: Various ways of Cyberbullying on social media platforms(adapted from [21])

3.3 Datasets used in Cyberbullying Detection

Datasets play a crucial role in the development and evaluation of cyberbullying detection systems. Machine learning and

deep learning models require large amounts of labeled textual data to learn patterns associated with abusive or harmful online behavior[30]. Most datasets used in cyberbullying research are collected from popular social media platforms such as Twitter, Facebook, Instagram, YouTube, and online discussion forums.Several publicly available datasets have been widely used in cyberbullying detection studies. One of the commonly used datasets is the Formspring dataset, which contains question–answer interactions collected from the Formspring social networking platform. This dataset includes labeled examples of bullying and non-bullying messages and has been frequently used for training and evaluating cyberbullying detection models[35].Another widely used dataset is the Twitter cyberbullying dataset, which consists of tweets labeled as abusive, offensive, or non-offensive. Due to the large volume of data available on Twitter, researchers often collect datasets using Twitter APIs and annotate the messages manually or through crowdsourcing platforms. Twitter datasets are particularly useful because they contain informal language, slang, and abbreviations that reflect real-world online communication.The Wikipedia talk page dataset is another important resource used in cyberbullying and hate speech detection research. This dataset contains comments from Wikipedia discussion pages that have been labeled for toxicity, personal attacks, and harassment[37]. It has been widely used to train machine learning models for detecting toxic and abusive online conversations.In addition to these datasets, several researchers also use Kaggle cyberbullying datasets, which contain labeled comments collected from different social media sources. These datasets often include multiple categories such as insult, threat, identity hate, and severe toxicity. However, many cyberbullying datasets suffer from the problem of class imbalance, where non-bullying messages significantly outnumber bullying content. This imbalance can negatively affect the performance of machine learning models and requires techniques such as resampling or data augmentation to address the issue[38].Overall, the availability of diverse and well-annotated datasets plays a critical role in improving the accuracy and robustness of cyberbullying detection systems. As social media platforms continue to evolve, researchers are increasingly focusing on creating larger and more representative datasets that capture different forms of online harassment and abusive behavior.

3.4 Cyberbullying Detection Framework

Cyberbullying detection systems are designed to automatically identify harmful or abusive content posted on social media platforms. These systems typically follow a structured pipeline consisting of several stages, including data collection, preprocessing, feature extraction, model training, and classification[29]. The overall process is illustrated in Fig 2.The first stage in the cyberbullying detection framework is data collection, where textual data such as posts, comments, and messages are gathered from social media platforms including Twitter, Facebook, Instagram, and online discussion forums. These datasets contain both bullying and non-bullying messages, which are used to train and evaluate machine learning models.

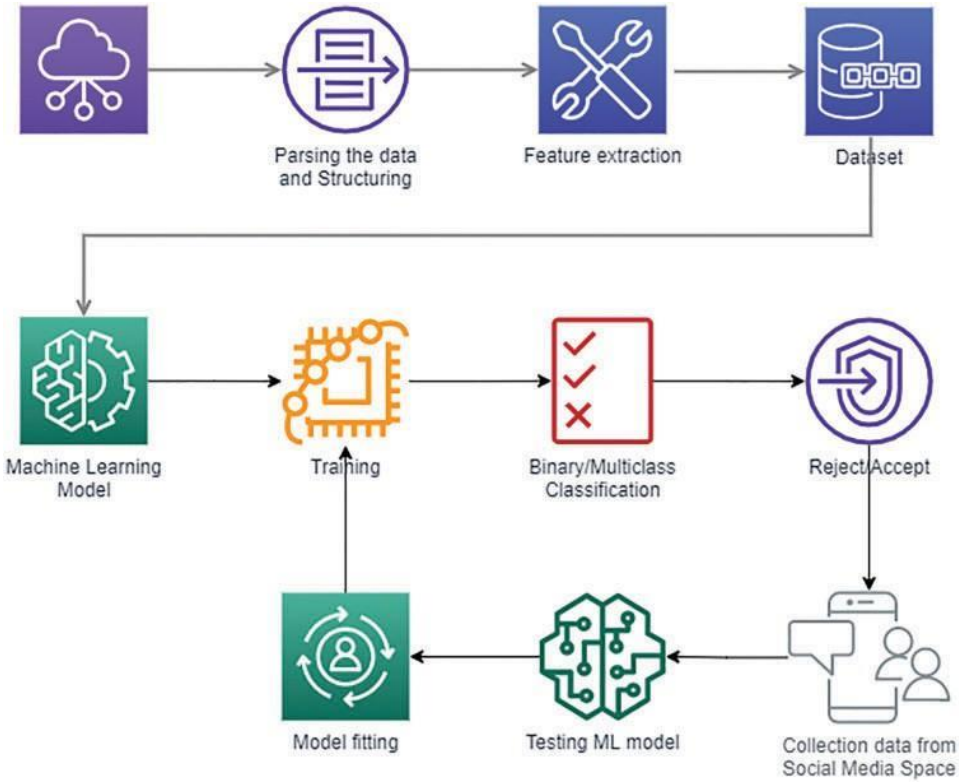

Fig 2: Machine Learning Pipeline for Cyberbullying Detection (adapted from [30])

After preprocessing, feature extraction techniques are used to convert textual data into numerical representations that machine learning models can understand. Common feature extraction methods include Bag-of-Words (BoW), Term Frequency–Inverse Document Frequency (TF-IDF), and word embeddings such as Word2Vec and GloVe[32]. These features capture important linguistic patterns that help distinguish abusive language from normal communication.The extracted features are then used to train machine learning or deep learning models. Algorithms such as Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks are commonly used to classify social media messages into bullying or non-bullying categories.

3.5 Machine Learning Algorithms Used in Cyberbullying Detection

Various machine learning algorithms have been applied to detect cyberbullying in social media data. These algorithms analyze textual features extracted from online posts and classify messages as bullying or non-bullying[29]. Machine learning models are trained using labeled datasets that contain examples of abusive and non-abusive messages collected from platforms such as Twitter, Facebook, YouTube, and online discussion forums by removing noise such as punctuation, URLs, emojis, and stop words. Table 1 summarizes several research studies that have used different machine learning algorithms for cyberbullying detection. The table highlights the diversity of approaches used in previous studies and shows how different algorithms have been applied to analyze social media data.

3.5.1 Support Vector Machines (SVM)

Support Vector Machines are one of the most commonly used algorithms in cyberbullying detection research. SVM works by identifying an optimal decision boundary that separates bullying and non-bullying messages in a high-dimensional feature space. Due to their effectiveness in handling text classification problems, SVM models have been widely used in many studies analyzing Twitter and online forum datasets[35]. Researchers often combine SVM with feature extraction techniques such as Bag-of-Words (BoW) and TF-IDF to improve classification performance.

3.5.2 Naive Bayes (NB)

Naïve Bayes is another widely used algorithm for cyberbullying detection, especially in text classification tasks. It is

based on Bayes' theorem and assumes that the features used for classification are independent of each other. Despite this simplifying assumption, Naïve Bayes has shown strong performance in detecting abusive language in social media messages. Its simplicity, low computational cost, and ability to handle large datasets make it a popular choice for cyberbullying detection systems[38].

3.5.3 Decision Tree and Random Forest

Decision Tree algorithms classify data by creating a tree-like structure of decision rules based on input features. In this approach, the dataset is repeatedly split into smaller subsets according to feature values, forming a hierarchical structure consisting of nodes and branches. Each internal node represents a decision based on a feature, while the leaf nodes represent the final classification outcome. In cyberbullying detection, Decision Trees analyze textual features extracted from social media messages to determine whether the content contains bullying behavior[40].One of the key advantages of Decision Tree models is their interpretability. The decision-making process can be easily visualized and understood, allowing researchers to analyze which textual features contribute most to cyberbullying classification. However, Decision Trees can sometimes suffer from overfitting, especially when the model becomes too complex and closely fits the training data rather than generalizing well to new data.Random Forest models are widely used in cyberbullying detection research because they provide strong classification performance and can handle large and complex datasets effectively. In addition, Random Forest algorithms are capable of identifying important features within textual data, which helps researchers understand patterns associated with abusive or offensive language on social media platforms[35].

3.5.4 K-Nearest Neighbours and Logistic Regression

K-Nearest Neighbors (KNN) is a simple and widely used classification algorithm that assigns a label to a new data instance based on the majority class of its nearest neighbors in the feature space. In cyberbullying detection systems, KNN compares a new social media message with previously labeled messages and determines whether it belongs to the bullying or non-bullying category based on similarity measures. The value of K represents the number of nearest neighbors considered during classification. Distance metrics such as Euclidean distance or cosine similarity are commonly used to identify the closest neighbors[38].

Logistic Regression is another commonly used machine learning algorithm for binary classification problems such as cyberbullying detection. Unlike KNN, Logistic Regression is a parametric model that estimates the probability that a given input belongs to a particular class. It uses a logistic function (sigmoid function) to map the predicted values between 0 and 1, representing the probability of a message being classified as bullying or non-bullying.Logistic Regression is widely used in text classification tasks because it is computationally efficient, easy to implement, and performs well with high-dimensional feature spaces[40].

Study	SVM	NB	RF	DT	KNN	LR	RB	Ensemble	Other
Agustín et al.[30]	C	×	×	×	×	×	×	×	×
Perera et al. [31]	C	×	×	×	×	×	×	×	×
Zinoviyeva et al[32].	C	×	C	×	×	C	×	×	×
Sarna et al.[33]	C	C	×	C	C	×	×	×	×
Thun et al. [34]	C	C	C	C	C	C	×	×	×
Lopez-Vizcaino et al.[35]	C	×	C	×	×	C	×	×	×
Mohammed Ali Al-garadi[36]	C	C	C	×	×	×	×	×	×
Gencoglu[37]	×	×	×	×	C	×	×	×	×
Meng [38]	×	×	×	×	×	×	×	C	C
Ahmed et al.[39]	C	C	C	×	×	C	×	C	×
Balakrishnan[40]	×	C	C	C	×	×	×	×	×
Silva [41]	C	C	C	C	C	×	×	×	×

Table 1**:** Summary of machine learning algorithms tested in cyberbullying literature (compiled from[30-41])

3.5.5 Ensemble Learning Methods

Ensemble learning approaches combine multiple machine learning models to improve prediction accuracy and robustness. Techniques such as bagging, boosting, and stacking are commonly used to combine classifiers like SVM, Decision Trees, and Logistic Regression. In cyberbullying detection research, ensemble models often achieve better performance because they leverage the strengths of multiple algorithms to detect abusive language patterns more effectively[39,40].In cyberbullying detection research, ensemble learning methods have demonstrated improved performance compared to individual classifiers because they capture diverse patterns in textual data. By combining algorithms such as Support Vector Machines, Decision Trees, and Logistic Regression, ensemble models are able to detect abusive language and harmful online behavior more effectively across different social media platforms[38-41].

IV. Methodology

This study proposes a framework for cyberbullying detection on social media platforms using machine learning techniques. The methodology follows several stages including data collection, data preprocessing, feature analysis, and decision-making using machine learning models[29,38]. The overall architecture of the proposed system is illustrated in Fig. 3.The proposed methodology begins with the collection of textual data from various social media platforms where users interact through comments, posts, and messages. These datasets may contain both normal and abusive language, which helps in training machine learning models to distinguish between cyberbullying and non-cyberbullying content. The collected data typically includes different forms of online harassment such as flaming, harassment, impersonation, and offensive communication.After data collection, a preprocessing stage is performed to clean and normalize the text data. Social media text often contains noise such as special characters, emojis, links, abbreviations, and spelling variations.The processed features are then provided to a decision-making system that includes machine learning classifiers and fuzzy logic mechanisms. Machine learning algorithms such as Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF), and Logistic Regression (LR) are commonly used to classify text into cyberbullying and non-cyberbullying categories[38-41]. This methodology helps in identifying harmful online interactions and contributes to the development of automated cyberbullying detection systems for safer social media environments.

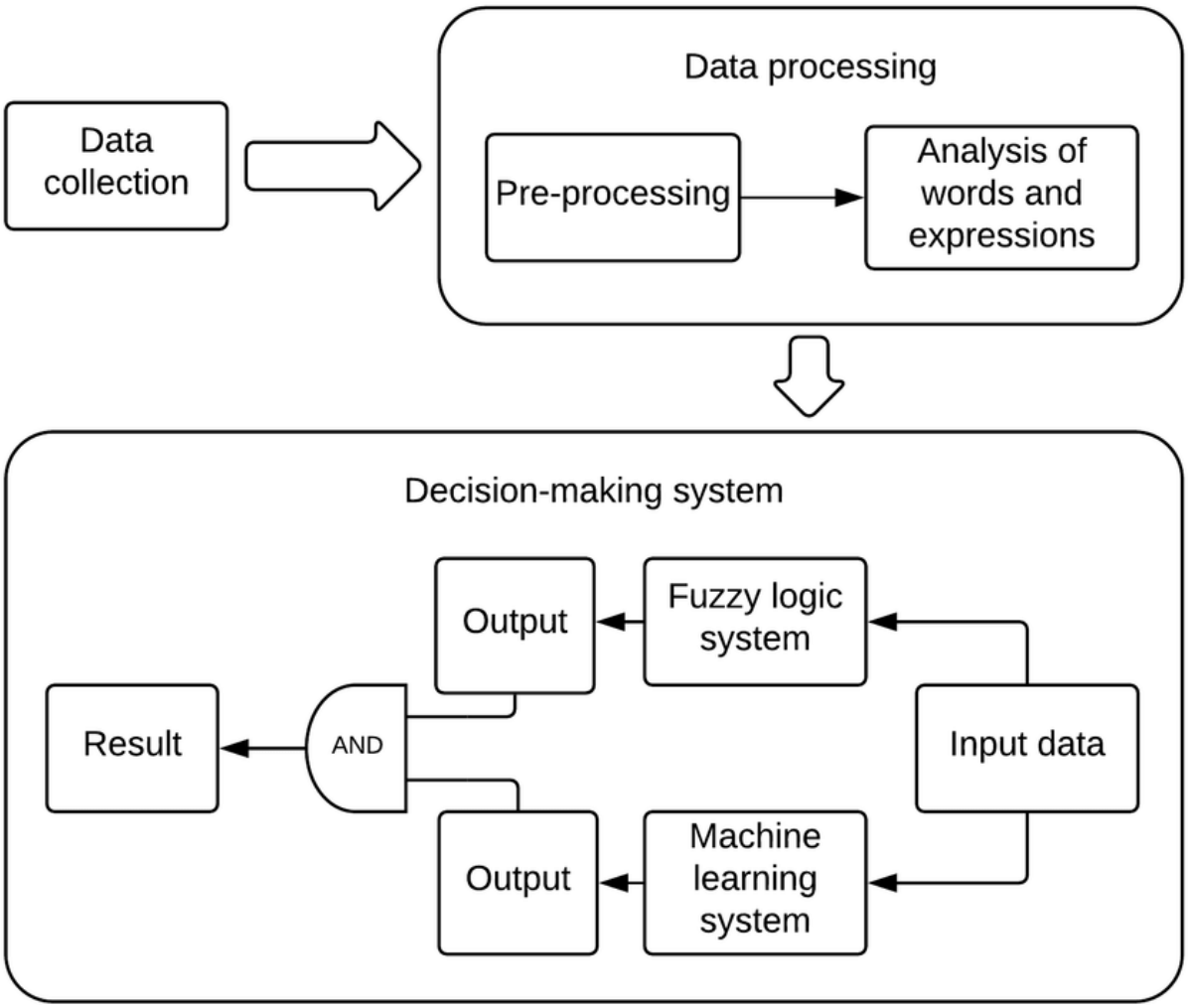

Fig 3: Architecture of Cyberbullying Detection System (adapted from[41])

4.1 Data Collection

The first step in the cyberbullying detection framework is the collection of textual data from social media platforms such as Twitter, Facebook, Instagram, and online discussion forums. These platforms generate a large amount of user-generated content in the form of comments, posts, and messages. Such data may contain both normal communication and abusive language[35].The collected datasets typically include labeled examples of cyberbullying and non-cyberbullying content, which are required for training machine learning models. These datasets help the system learn patterns associated with harmful online behavior such as harassment, flaming, impersonation, and offensive communication[29].

4.2 Data Preprocessing and Feature Analysis

Social media data is often unstructured and noisy. Therefore, preprocessing techniques are applied to clean and normalize the collected text data. This stage includes operations such as removing stop words, eliminating punctuation marks, converting text to lowercase, and tokenizing sentences into individual words[31].After preprocessing, feature extraction techniques are applied to transform textual data into numerical representations that machine learning models can process. Common techniques include Bag-of-Words, Term Frequency–Inverse Document Frequency (TF-IDF), and n-gram models[37]. These features help identify patterns in language usage and enable the system to distinguish between abusive and non-abusive messages.

4.3 Decision Making and Model Classification

After feature extraction, the processed data is provided to the decision-making system. In this stage, machine learning algorithms are used to classify the messages into cyberbullying or non-cyberbullying categories. Common classifiers used in cyberbullying detection research include Support Vector Machines (SVM), Naïve Bayes (NB), Random Forest (RF), K-Nearest Neighbors (KNN), and Logistic Regression (LR)[38-41].In addition, fuzzy logic techniques may be used to handle uncertainty and ambiguous language patterns in online conversations. The outputs generated by the machine learning model and fuzzy logic system are combined to produce the final classification result. The effectiveness of the model is evaluated using performance metrics such as accuracy, precision, recall, and F1-score[39].

V. Future Research Directions

Although cyberbullying detection using machine learning has gained considerable attention in recent years, several areas still require deeper investigation to improve the effectiveness, scalability, and practical applicability of automated detection systems[19]. Future research should focus on improving the ability of cyberbullying detection models to handle large-scale social media data, diverse communication styles, and continuously evolving language patterns[33]. As online platforms increasingly host millions of active users generating vast amounts of content every day, there is a need for efficient and scalable detection mechanisms that can analyze text data without introducing significant computational overhead or delays[34]. Developing optimized machine learning architectures and real-time content moderation systems capable of processing large volumes of online interactions remains an important research direction.In addition to scalability, intelligent automation represents another critical area for advancement. While machine learning and artificial intelligence have already been widely explored for identifying abusive language and harmful online behavior, their integration into real-world moderation systems still requires further validation. Future studies should investigate more robust classification models capable of understanding contextual meaning, reducing false positives and false negatives, and improving overall detection accuracy. Incorporating behavioral analysis, contextual language understanding, and adaptive learning mechanisms into cyberbullying detection systems can significantly enhance the ability to identify harmful interactions before they escalate.Integration and interoperability challenges further highlight the need for standardized frameworks and evaluation practices for cyberbullying detection systems[38]. Online communication environments are highly diverse and include social media platforms, online forums, messaging applications, and digital learning environments[10]. Future research should focus on developing interoperable detection frameworks and evaluation methods that allow cyberbullying detection models to function effectively across different platforms. Additionally, standardized datasets and benchmarking metrics are necessary for accurately comparing the performance of detection algorithms proposed by different studies.

Cross-platform and cross-cultural research is also essential to address the global nature of online communication[11]. Language usage, cultural expressions, and communication styles vary significantly across different regions and communities. Long-term studies, large-scale dataset analysis, and comparative evaluations across multiple online environments would provide valuable insights into improving the reliability, fairness, and effectiveness of cyberbullying detection technologies[36]. By addressing these research directions, future advancements can transform cyberbullying detection systems from experimental research prototypes into scalable and reliable tools capable of supporting safer digital communities.

5.1 Standardization and Unified Cyberbullying Detection Frameworks

Current literature reveals a noticeable lack of universally accepted standards and implementation models for cyberbullying detection systems across diverse online platforms.Although many researchers and organizations propose their own machine learning approaches for detecting abusive language, inconsistencies in dataset selection, feature extraction techniques, and evaluation metrics often create confusion for researchers and practitioners. Future research should prioritize the development of unified detection frameworks and reference models that clearly define preprocessing methods, classification approaches, and performance evaluation criteria. Establishing standardized implementation guidelines would improve interoperability between social media platforms, data processing systems, and automated moderation tools[31].Moreover, standardized benchmarking datasets and evaluation mechanisms could help researchers measure the effectiveness of detection algorithms in a more systematic manner. Such harmonization would not only simplify the development of cyberbullying detection systems but also promote collaborative research and consistency across different online communication environments.

5.2 Advanced AI-Driven Cyberbullying Detection Models

While machine learning techniques have been widely explored for detecting abusive language and harmful online interactions,the development of more advanced artificial intelligence models for contextual understanding remains an

evolving area[19]. Future research should focus on developing adaptive detection models capable of continuously learning from language patterns, user behavior, and contextual communication signals.Machine learning algorithms can be used to generate dynamic risk scores that help identify potentially harmful interactions in real time[33]. However, several challenges must be addressed. Long-term empirical studies are also necessary to validate the reliability and robustness of AI-driven cyberbullying detection systems across different social media environments.

5.3 Cyberbullying Detection in Cross Platform Environments

The rapid expansion of global online communication has introduced additional challenges for cyberbullying detection due to linguistic diversity and platform heterogeneity[21].Social media users communicate using different languages, dialects, and informal expressions, which makes it difficult for detection systems to accurately interpret abusive behavior[29].Traditional text classification models trained on single-language datasets may not perform effectively when applied to multilingual online environments[34].Future research should therefore explore detection methods capable of analyzing multilingual text and adapting to cultural variations in communication patterns.Additionally, decentralized detection frameworks may be required to process harmful interactions directly within different platform infrastructures without relying entirely on centralized moderation systems. Investigating multilingual analysis techniques and cross-platform detection models could significantly enhance the effectiveness of cyberbullying detection systems in global digital communities.

VI. CONCLUSION

This review examined existing research on cyberbullying detection using machine learning techniques and summarized the key approaches, datasets, and challenges reported in previous studies. The analysis highlights the growing importance of automated systems in identifying harmful online behavior as social media platforms continue to expand. By analyzing different machine learning models and natural language processing methods used in prior work, this study provides a structured overview of the current state of cyberbullying detection research.Research activity in this area has increased significantly over the past decade as concerns about online harassment and digital safety have become more prominent. However, despite the increasing number of proposed detection models, several limitations remain. Many studies mainly focus on improving classification accuracy under controlled experimental conditions, while fewer investigations examine how these models perform in real-world social media environments where language usage is dynamic and unpredictable.Detecting cyberbullying remains challenging because online communication often includes sarcasm, indirect insults, slang expressions, and contextual meanings that are difficult for automated systems to interpret accurately. These factors can reduce the reliability of existing detection systems and highlight the need for improved language understanding models. In addition, the limited availability of large and diverse datasets continues to restrict the ability of machine learning algorithms to generalize across different platforms and communication styles.

Further research is required to improve contextual analysis, develop multilingual detection systems, and design models capable of operating efficiently on large-scale social media data. The development of standardized datasets and evaluation methods would also help researchers compare different detection techniques more effectively. Overall, this review aims to support ongoing efforts to improve cyberbullying detection technologies.

REFERENCES

1. Edosomwan, S.; Prakasan, S.K.; Kouame, D.; Watson, J.; Seymour, T. The history of social media and its impact on business. J. Appl. Manag. Entrep. 2011, 16, 79–91.
2. Bauman, S. Cyberbullying: What Counselors Need to Know; John Wiley & Sons: Hoboken, NJ, USA, 2014.
3. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M. Detecting and Monitoring Hate Speech in Twitter. Sensors 2019, 19, 4654. [CrossRef]
4. Miller, K. Cyberbullying and its consequences: How cyberbullying is contorting the minds of victims and bullies alike, and the law's limited available redress. S. Cal. Interdisc. Law J. 2016, 26, 379.
5. Price, M.; Dalgleish, J. Cyberbullying: Experiences, impacts and coping strategies as described by australian young people. Youth Stud. Aust. 2010, 29, 51.
6. Smith, P.K. Cyberbullying and Cyber Aggression. In Handbook of School Violence and School Safety; Informa UK Limited: Colchester, UK, 2015.

7. Sampasa-Kanyinga, H.; Roumeliotis, P.; Xu, H. Associations between Cyberbullying and School Bullying Victimization and Suicidal Ideation, Plans and Attempts among Canadian Schoolchildren. PLoS ONE 2014, 9, e102145. [CrossRef]
8. Davidson, T.; Warmsley, D.; Macy, M.; Weber, I. Automated Hate Speech Detection and the Problem of Offensive Language. arXiv 2017, arXiv:1703.04009.
9. Mc Guckin, C.; Corcoran, L. (Eds.) Cyberbullying: Where Are We Now? A Cross-National Understanding; MDPI: Wuhan, China, 2017.
10. Vaillancourt, T.; Faris, R.; Mishna, F. Cyberbullying in Children and Youth: Implications for Health and Clinical Practice. Can. J. Psychiatry 2016, 62, 368–373. [CrossRef]
11. Görzig, A.; Ólafsson, K. What Makes a Bully a Cyberbully? Unravelling the Characteristics of Cyberbullies across Twenty-Five European Countries. J. Child. Media 2013, 7, 9–27. [CrossRef]
12. Salton, G.; Buckley, C. Term-wsevening approaches in automatic text retrieval. Inf. Process. Manag. 1988, 24, 513–523.
13. Liu, Q.; Wang, J.; Zhang, D.; Yang, Y.; Wang, N. Text Features Extraction based on TF-IDF Associating Semantic. In Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC), Chengdu, China, 7–10 December 2018; Institute of Electrical and Electronics Engineers (IEEE): Piscataway, NJ, USA, 2018; pp. 2338–2343.
14. Pennington, J.; Socher, R.; Manning, C. Glove: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014.
15. Goldberg, Y.; Levy, O. word2vec Explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv 2014, arXiv:1402.3722.
16. Li, J.; Huang, G.; Fan, C.; Sun, Z.; Zhu, H. Key word extraction for short text via word2vec, doc2vec, and textrank. Turk. J. Electr. Eng. Comput. Sci. 2019, 27, 1794–1805. [CrossRef]
17. Jiang, C.; Zhang, H.; Ren, Y.; Han, Z.; Chen, K.-C.; Hanzo, L. Machine Learning Paradigms for Next-Generation Wireless Networks. IEEE Wirel. Commun. 2016, 24, 98–105. [CrossRef]
18. Medhat, W.; Hassan, A.; Korashy, H. Sentiment analysis algorithms and applications: A survey. Ain Shams Eng. J. 2014, 5, 1093–1113. [CrossRef]
19. Al-Garadi, M.A.; Hussain, M.R.; Khan, N.; Murtaza, G.; Nweke, H.F.; Ali, I.; Mujtaba, G.; Chiroma, H.; Khattak, H.A.; Gani, A. Predicting Cyberbullying on Social Media in the Big Data Era Using Machine Learning Algorithms: Review of Literature and Open Challenges. IEEE Access 2019, 7, 70701–70718. [CrossRef]
20. Maalouf, M. Logistic regression in data analysis: An overview. Int. J. Data Anal. Tech. Strat. 2011, 3, 281–299. [CrossRef]
21. R. M. Kowalski, S. P. Limber and A. McCord, "A developmental approach to cyberbullying: Prevalence and protective factors," Aggression and Violent Behavior, vol. 45, pp. 20–32, 2019.
22. Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. Applied Logistic Regression; Wiley: Hoboken, NJ, USA, 2013; Volume 398
23. G. Allsopp, J. Rosenthal, J. Blythe and J. S. Taggar, "Defining and measuring denigration of general practice in medical education," Education for Primary Care, vol. 31, no. 4, pp. 205–209, 2020.
24. Chavan, V.S.; Shylaja, S.S. Machine learning approach for detection of cyber-aggressive comments by peers on social media networks. In Proceedings of the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Kochi, India, 10–13 August 2015; Institute of Electrical and Electronics Engineers (IEEE): New York, NY, USA, 2015; pp. 2354–2358.
25. E. Villar-Rodriguez, J. D. Ser, S. Gil-Lopez, M. N. Bilbao and S. Salcedo-Sanz, "A Meta-heuristic learning approach for the non-intrusive detection of impersonation attacks in social networks,"International Journal of Bio-Inspired Computation, vol. 10, no. 2, pp. 109–118, 2017.
26. A. Cassiman, "Spiders on the world wide web: Cyber trickery and gender fraud among youth in an Accra zongo," Social Anthropology, vol. 27, no. 3, pp. 486–500, 2019.
27. K. Williams, C. Cheung and W. Choi, "Cyberostracism: Effects of Being Ignored over the Internet," Journal of Personality and Social Psychology, vol. 79, no. 5, pp. 748–762, 2000.
28. D. Álvarez-García, J. C. Núñez, A. Barreiro-Collazo and T. García, "Validation of the cybervictimization questionnaire (CYVIC) for adolescents," Computers in Human Behavior, vol. 70, pp. 270–281, 2017.
29. A. Sanchez-Medina, I. Galvan-Sanchez and M. Fernandez-Monro, "Applying artificial intelligence to explore sexual cyberbullying behaviour," Heliyon, vol. 6, no. 1, pp. 1–9, 2020.
30. A. Perera and P. Fernandol, "Accurate cyberbullying detection and prevention on social media," Procedia Computer Science, vol. 181, pp. 605–611, 2021.
31. E. Zinoviyeva, W. Karl Hardle and S. Lessmann, "Antisocial online behavior detection using deep learning. Decision Support Systems, vol. 138, no. 1, pp. 1–9, 2020.
32. R. Zhao and K. Mao, "Cyberbullying detection based on semantic-enhanced marginalized denoising autoencoder," IEEE Transactions on Affective Computing, vol. 8, no. 3, pp. 328–329, 2016.
33. L. Thun, P. The and C. Cheng, "CyberAid: Are your children safe from cyberbullying?," Journal of King Saud University-Computer and Information Sciences, vol. 34, no. 7, pp. 4099–4108, 2022.
34. M. Lopez-Vizcaino, F. Novoa, V. Carneiro and F. Cacheida, "Early detection of cyberbullying on social media networks," Future Generation Computer Systems," vol. 118, pp. 219–229, 2021.
35. M. Al-garadi, K. Varatham and S. Ravana, "Cybercrime detection in online communications: The experimental case of cyberbullying detection in the twitter network," Computers in Human Behavior, vol. 63, pp. 433–443, 2016.
36. O. Gencoglu, "Cyberbullying detection with fairness constraints," IEEE Internet Computing, vol. 25, no. 1, pp. 20–29, 2020.
37. Z. Meng, S. Tian and L. Yu, "Regional bullying text recognition based on two-branch parallel neural networks," Automatic Control and Computer Sciences, vol. 54, no. 4, pp. 323–334, 2020.
38. J. Hani, M. Nashaat, M. Ahmed, Z. Emad, E. Amer et al., "Social media cyberbullying detection using machine learning," International Journal of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 703–707, 2019.
39. T. Ahmed, S. Ivan, M. Kabir, H. Mahmud and K. Hasan, "Performance analysis of transformer-based architectures and their ensembles to detect trait-based cyberbullying,"Social Network Analysis and Mining, vol. 12, no. 1, pp. 1–17, 2022.
40. V. Balakrishnan, Sh. Khan and H. Arabnia, "Improving Cyberbullying Detection using Twitter Users' Psychological Features and Machine Learning," Computers and Security, vol. 90, no. 1, pp. 1–11, 2019.
41. Y. Silva, D. Hall and C. Rich, "BullyBlocker: Toward an interdisciplinary approach to identify cyberbullying," Social Network Analysis and Mining, vol. 8, no. 1, 1–15, 2018.
42. A. K. Jha, "Sensing and Supervising through IoT," International Journal of Computer Applications, vol. 152, no. 9, pp. 7–9, 2016.

ISSN: 0975-8887. DOI: 10.5120/ijca2016911723.
43. A. K. Jha, M. P. Patel, and T. D. Pawar, “Fog offloading: Review, research opportunity and challenges,” in Proc. 2019 Int. Conf. Smart Syst. Invent. Technol. (ICSSIT), 2019, pp. 1224–1227. DOI: 10.1109/ICSSIT46314.2019.8987905.
44. A. K. Jha, M. P. Patel, and T. D. Pawar, “A proposed model of computation offloading in fog environment,” Sambodhi (UGC Care Journal), vol. 43, no. 03(IV), pp. 1–6, Nov.–Dec. 2020. ISSN: 2249-6661.
45. A. K. Jha and T. Pawar, “Computation Offloading for Smart Healthcare Applications,” in IoT Applications for Healthcare Systems. Cham: Springer, 2022, pp. 121–136. DOI: 10.1007/978-3-030-91096-9_7.
46. A. K. Jha, M. P. Patel, and T. D. Pawar, “Computation offloading using K-nearest neighbour time critical optimisation algorithm in fog computing,” International Journal of Wireless and Mobile Computing, vol. 23, no. 3–4, pp. 281–292, 2022. ISSN: 1741-1084 (Print), 1741-1092 (Online). DOI: 10.1504/IJWMC.2022.127593.
47. A. K. Jha, M. P. Patel, and T. D. Pawar, “Extended hybrid cluster algorithm for computation offloading in fog computing,” International Journal on Technical and Physical Problems of Engineering (IJTPE), issue 51, vol. 14, no. 2, pp. 176–182, Jun. 2022.
48. M. Patel, A. Mehta, A. K. Jha, A. Patel, and A. Nayak, “A deep reinforcement prediction model for live VM migration in fog,” International Journal on Technical and Physical Problems of Engineering (IJTPE), issue 58, vol. 16, no. 1, pp. 277–283, Mar. 2024.
49. V. Soni and A. Jha, “IoT Botnet Attacks Detection Using Deep Learning Approaches: A Review,” IET Conference Proceedings, vol. 2025, no. 7, pp. 253–260, 2025.
50. R. Shankar, I. Kumar, M. Kashyap, A. K. Jha, and B. P. Chaudhary, “A Review on NOMA scheme for emerging 6G wireless networks: State of the Art, Key Schemes, Future scope and Security Issues,” Radioelectronics and Communications Systems, vol. 68, no. 5, pp. 271–284, 2025. DOI: 10.3103/S0735272725010017.
51. M. S. Shaikh, A. K. Jha, B. R. Soni, R. N. K. Patel, and D. P. M., “Flying Edge Intelligence: UAV-Driven Edge Computing for Autonomous Precision Farming,” in Proc. 2025 Int. Conf. Emerging Technol. Eng. Appl. (ICETEA), 2025, pp. 1–6. DOI: 10.1109/ICETEA64585.2025.11099749.
52. A. K. Jha, A. Khatri, K. Kanda, A. Haider, and R. Shah, “A review of security, privacy, and authentication mechanisms in social media web applications,” PUXplore Multidisciplinary Journal of Engineering, vol. 2, no. 1, Mar. 2026, doi: 10.62373/3n8bxz70.
53. Ali, S. I., Kale, G. P., Shaikh, M. S., Ponnusamy, S., & Chouhan, P. S. (2024). AI Applications and Digital Twin Technology Have the Ability to Completely Transform the Future. In S. Ponnusamy, M. Assaf, J. Antari, S. Singh, & S. Kalyanaraman (Eds.), Harnessing AI and Digital Twin Technologies in Businesses (pp. 26-39). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-3234-4.ch003
54. Karunambiga, K., Ali, S. I., Esteban, A. P., & Pascual, M. (2023). Marketing policy in service enterprises using deep learning models. International Journal of Intelligent Systems and Applications in Engineering, 12(7s), 239-243.
55. Ali, S. I., Ravuri, H. K., Lakshmi, V. T., Ramya, A., Lavanya, K., & Bahade, S. (2024). The role of nanomaterials in the development of high-performance batteries. Nanotechnology Perceptions, 20(S11), 1125–1140. https://www.nano-ntp.com
56. Ali, S. I., Salunke, B. A., Salunke, S., Chouhan, P. S., & Shahane, S. (2026). Decentralized Smart Grids With AI and Blockchain: Enabling Peer-to-Peer Energy Trading and Energy Equity. In E. Babulak (Ed.), Advancing Energy Production and Distribution With Blockchain and AI (pp. 163-200). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-6996-9.ch006
57. Ali, S. I. (2026). Reinforcement Learning for Autonomous Optimization in Intelligent Engineering. In E. Babulak (Ed.), AI-Driven Approaches for Fully Automated Smart Engineering (pp. 313-344). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-4839-1.ch011
58. Ali, S. I., Kalaivaani, P. T., Ambigaipriya, S., & Rafeeq, M. D. (2023). Evaluation of AI model performance. In Toward artificial general intelligence: Deep learning, neural networks, generative AI (p. 125). Walter de Gruyter GmbH & Co. KG.
59. Ali, S. I., Jadhav, J., Arunkumar, R., & Kanagavalli, N. (2022). A smart resource utilization algorithm for high-speed 5G communication networks based on cloud servers. ICTACT Journal on Communication Technology, 13(??), 2800. https://doi.org/10.21917/ijct.2022.0414
60. Ali, S. I. (2026). Algorithmic Justice: Navigating AI's Role in Cybersecurity and Legal Transformation. In J. Luftman & A. Tomer (Eds.), Moral and Legal Aspects of Artificial Intelligence: Machine Bias and Rule of Law (pp. 229-264). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-3114-0.ch007
61. Ali, S. I., Dubey, A., Salunke, S., Salunke, B. A., & Chopkar, P. N. (2026). Advancing Energy Production and Distribution With Blockchain and AI: Towards Intelligent, Secure, and Sustainable Power Ecosystems. In E. Babulak (Ed.), Advancing Energy Production and Distribution With Blockchain and AI (pp. 83-112). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-6996-9.ch004
62. Agal, S., Raulji, K. & Odedra, N.D. A machine learning approach to risk based asset allocation in portfolio optimization. Sci Rep 15, 42263 (2025). https://doi.org/10.1038/s41598-025-26337-x
63. Sanjay Agal, Krishna Raulji, Nikunj Bhavsar and Pooja Bhatt. “Spatiotemporal Graph Networks for Relational Reasoning in Campus Infrastructure Management”. International Journal of Advanced Computer Science and Applications (ijacsa) 16.10 (2025). http://dx.doi.org/10.14569/IJACSA.2025.016108