

AI-POWERED MINDSHIELD MODEL FOR ANALYZING TOXICITY AND THREATS IN ONLINE INTERACTIONS

Shristi Verma¹, Pranshi Goyal^{2*}, Kalidindi Sowmya^{3*}

^{1,2,3}Sharda University, School of Computing Science and Engineering

0555shristi@gmail.com¹, goyalpranshi77@gmail.com², sowmyanagaraju0@gmail.com³

Abstract

Cyberbullying, hate speech and harassment have been reinforced by the rise of internet platforms. speech, and harassment have been enhanced by the booming nature of internet platforms. These are very difficult roles psychologically, morally and socially. the roles that are very difficult psychologically, morally and socially. In this article, the authors present the system known as MindShield, an artificial intelligence-based system, whose goal is to identify, classify and study dangerous activities online. paper, the authors present the system known as the MindShield that is the artificial intelligencebased system, the purpose of the system is to identify, classify, and study dangerous online activities. The program involves natural language processing (NLP), semantic integration as well as deep learning classifiers to identify the presence of toxic behaviors in text communication. natural language processing (NLP), semantic embeddings as well as deep learning classifier s in identifying presence of the toxic behaviors within the textbased communication. Regarding the mind shield , compared to its conventional moderation mechanisms, MindShield should prioritize the elements of explainability, risk identification and adaptive learning, thanks to which it is able to adapt to the reaction to new categories of cyber threats. as compared to its conventional moderation mechanisms, MindShield ought to prioritize the elements of explainability, risk identification and adaptive learning, whereby it is able to adjust to reaction to new categories of cyber threats. The experimental results are characterized by high performance compared to the benchmark datasets, as well as strong improvement in precision, recall and interpretability. comparison with benchmark datasets, as well as high improvement in precision, recall, and interpretability.

Keywords: *Bad online socialization, AI system, cyberbullying detection, natural language processing, semantic embeddings, transparent AI, internet safety.*

1. Introduction

The radical impact of digitalization of the society has been very significant to how individuals are communicating, sharing information and interacting. In as much as these virtual spaces have enabled innovation, learning and international collaboration, it has come with it the serious threat of malicious internet activity such as cyber bullying, harassment and hate speech. The recent studies have revealed that AI systems have enhanced transport planning and decision-making in the name of sustainability [1], and such approaches could be utilized to mitigate online harm. On the other hand, the biggest concerns with the new digital platforms are security and privacy, in which there is a growing threat of misuse of data, e-surveillance, and identity theft [2]. The interoperability, scalability, and risk management structures are also concerns, which AI applications in the industry introduce to the table so that the further reliability of digital systems would be maintained [3].

These problems have been exacerbated by the Metaverse: it is not only a socio-technical system, but also on the virtual and physical one. As it has been proved, the relationship between technological progress and moral dictatorship is rather fragile in this immersed setting [4]. Through education, it is proved that immersive learning experiences, or the Edu-Metaverse,

can be used to make learning more inclusive and interactive [5]. Urban governance research indicates that the Metaverse as a virtual image of smart cities may make cities more transparent, efficient, and sustainable in case the digital divide and equal access are resolved [6]. Digital Twins are technologies that offer simulations of physical systems in real time, allowing prediction of physical systems, which makes them predictable and can be optimized [7]. On the same note, the Extended Reality (XR) platforms reveal the potential transformative uses of the immersive interaction, and present distinctive privacy and security hazards that will have to be addressed [8]. Lastly, the developments in semantic communication and the 6G networks show the rising demands of security, flexibility, and intelligent infrastructures capable of supporting mass, real-time, digital environments [9]. Based on these results, the study presents MINDSHIELD, a platform consisting of AI that determines and analyzes malicious interactions on the Internet. According to NLP, semantic embeddings, and explainable AI, MINDSHIELD is the connection between technological advances and an immediate social demand for safe and respectful online conversation.

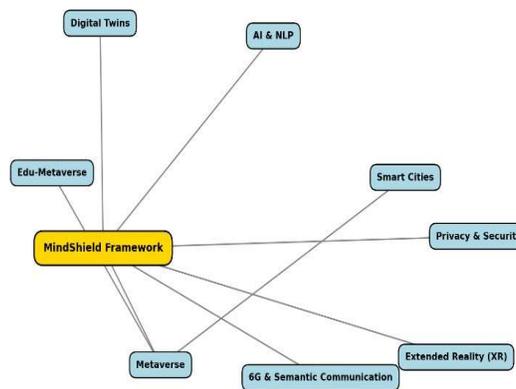


Fig.1 Conceptual Model Connecting Emerging Technologies to MindShield.

The main value of this paper consists in:

1. Launches MindShield, an innovative AI-powered abusive interaction detector and analyzer.
2. Introduces a hybrid detection system that consists of semantic embeddings, explainable AI, and dynamic risk evaluation.
3. Provides practical analysis of benchmark datasets, such as the advancements in the detection accuracy and readability.
4. Provides the critical analysis of ethical, social and governance implications.

The remaining part of this paper will be organized in the following way: Section II - Literature Review; Section III

- Proposed Methodology; Section - IV Results and Critical Discussion; Section - V Conclusion and Future Work.

2. Literature Review

Semantic communication, edge learning and 6G integration convergence have been accelerated by the creation of immersive environments like the Metaverse [9]. Those technologies facilitate interactive responses and real-time, which is necessary when working with large-scale digital settings. However, they are also associated with some weaknesses such as latency, privacy exposure, adversarial attacks, and AI-based security can sufficiently respond to them and render them safe to use. The IoT-based smart systems are reflected in the medical domain of clinical decision support and patient monitoring and presuppose the application of AI to revolutionize the medical industry [10]. These systems are based on linked devices to provide ongoing health information, which improves efficiency and predictive capacity. In transport, large language models (LLMs) are being increasingly used to assist in enhancing traffic movements and allowing intelligent transport systems [11]. Although resilient, such systems are a source of concern in the areas of bias and explainability. The idea of cyber-physical systems (CPS) research demonstrates the unity of integrating both physical processes and virtual control, however, it also shows susceptibility in the education environment where the accuracy of data and the stability of systems play a crucial role [12]. Green transport systems must incorporate the use of AI-controlled automation since real-time optimization will decrease the environmental impact and enhance sustainability [13]. Edge intelligence has also become one of the enabling factors in a secure IoT-6G integration that will enable the making of locally made decisions and minimal exposure to centralized attack [14]. Security threats go up in virtual environments like the Metaverse and therefore there are suggestions of network intrusion detection systems that are specifically tailored to the distinct virtual environments [15]. Meanwhile, the digital twin is spreading into most businesses, which allows synchronizing a virtual and physical system in real time. The predictability capabilities enhance the industrial activities but they generate problems of information ownership, access rights, and confidence [16]. The preeminent studies on 7G network focus on decentralized intelligence and high levels of connectivity, and should be able to address massive immersive applications with ultra-low-latency [17]. Nevertheless, the issues that are mentioned in these papers include standardization and equal access. On the social level, studies point to new concerns on the creation, implementation and regulation of Metaverse ecosystems, which require models that align innovation and ethical authority [18]. In the meantime, quantum federated learning (QFL) has emerged as a more advanced method of secure collaborative AI that allows training sensitive data on decentralized computers and guarantees confidentiality [19]. The e-learning systems that rely on the Metaverse in the educational sector demonstrate potential in inclusivity, immersion, and global teamwork [20]. Nevertheless, such researches also point to the difficulties with fair participation, access, and maintaining secure online classrooms. The studies on sustainable Metaverse platforms point to the necessity of creating the governance frameworks and research projects that would cover both the social and technical concerns [21]. Machine learning continues to be the basis of innovations in intelligent networking, which plays a big role in the management of the changing communication environments [22]. Moreover, the significance of the UAV-IoT-ML integration has emerged as a versatile and extensible solution to smart transportation with the focus on the opportunities of combining aerial and land networks [23]. Besides infrastructure,

the use of generative AI in virtual reality (VR) has created the opportunities of a customizable and adaptive virtual space, which makes virtual content adaptive and transformable to the specific needs of a user [24]. Finally, the current advancement of semantic communication in 6G networks indicates that the models of transmission that focus on the meaning of a transmission can reduce redundancy and maximize the efficiency of immersive communication [25]. Overall, there is a significant progress in the literature regarding the development of technologies that would enable the implementation of immersive, intelligent, and connected ecosystems. However, most of the efforts now are aimed at maximizing infrastructure, industrial applications, or immersive education, leaving a long-term gap in addressing the problem of harmful online interactions, which is only growing. Despite the developments in the field of AI, 6G, and security frameworks, there is no research exploring the specifics of toxic behavior, cyberbullying, harassment, and hate speech in the online environment. MindShield is proposed to address the given problem, being an AI-based, transparent, and flexible framework that offers protection of digital communication space.

3. Proposed Methodology

4.1 DATA ACQUISITION AND PREPROCESSING:

Information can be collected with the help of multiple internet sources, social media, forums, and chat applications, and safe APIs can be used to secure privacy and comply with the rules. Preprocessing stages refine and normalize the text for subsequent examination.

Text normalization changes all letters to lowercase, removes special symbols, and eliminates stopwords to lessen noise. Tokenization divides sentences into separate words or tokens. Lemmatization and stemming bring words down to their base forms for uniform representation. To address data imbalance between harmful and benign interactions, techniques such as oversampling with SMOTE or undersampling are utilized.

Mathematically, for a corpus $D = \{d_1, d_2, \dots, d_n\}$, the preprocessing function f_{prep} maps raw data d_i to cleaned tokens t_i :

$$t_i = f_{prep}(d_i) \quad (1)$$

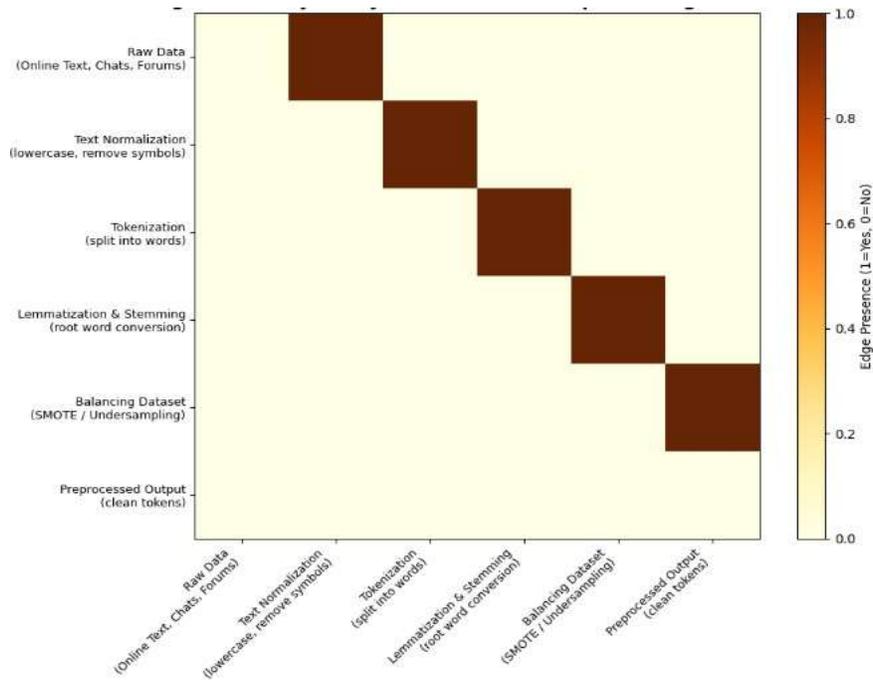


Fig.2 Adjacency Matrix of Data Preprocessing Workflow

4.2 SEMANTIC EMBEDDING GENERATION:

Each token t_i is converted into a dense vector representation $v_i \in R^d$ using pre-trained embedding models such as *BERT* or *Word2Vec*:

$$v_i = E(t_i) \quad (2)$$

Where E is the embedding function mapping each token to a d - dimensional vector. The complete text representation

V is obtained by aggregating all token embeddings:

$$V = \frac{1}{n} \sum_{i=1}^n v_i \quad (3)$$

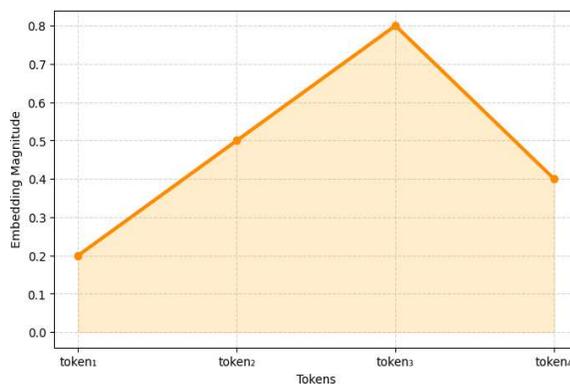


Fig.3 Mapping of Tokens in Semantic Embedding

4.3 CLASSIFICATION LAYER (BiLSTM + ATTENTION):

A Bidirectional Long Short-Term Memory (*BiLSTM*) network is employed to grasp contextual dependencies in online interactions, The *BiLSTM* calculates forward (h_t^{\rightarrow}) and backward (h_t^{\leftarrow}) hidden states for every token:

$$h_t = h_t^{\rightarrow}; h_t^{\leftarrow} \quad (4)$$

A mechanism of attention allocates weights α_t to highlight important tokens:

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^n \exp(e_k)}, e_t = \tanh(W_a h_t + b_a) \quad (5)$$

The context vector c is determined as:

$$c = \sum_{t=1}^n \alpha_t h_t \quad (6)$$

Ultimately, the output layer estimates the likelihood of a detrimental interaction:

$$y^{\wedge} = \text{softmax}(W_c c + b_c) \quad (7)$$

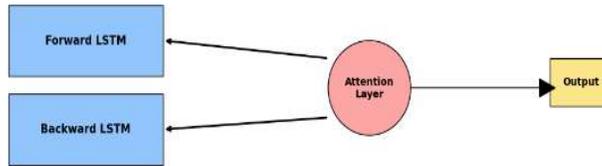


Fig. 4 Workflow for Classification Using BiLSTM with Attention.

4.4 RISK ASSESSMENT SYSTEM:

The risk score R measures the extent of detrimental interactions based on predictions from the model and semantic strength.

$$R = \lambda_1 \cdot P(\text{harmful}|x) + \lambda_2 \cdot S(x) \quad (8)$$

Where $P(\text{harmful}|x)$ represents the estimated likelihood of a harmful interaction, $S(x)$ denotes a semantic severity score based on keyword intensity and context, and λ_1 and λ_2 adjustable weights ensuring $\lambda_1 + \lambda_2 = 1$.

The categories of risks are:

Minimal Risk: $R < 0.4$.

Medium Risk: $0.4 \leq R < 0.7$. High Risk: $R \geq 0.7$.

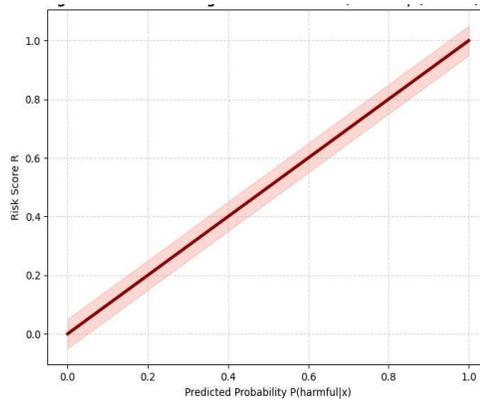


Fig. 5 Function for Risk Assessment

4.5 EXPLAINABILITY MODULE:

SHAP (*SHapley Additive exPlanations*) or *LIME* is applied to increase the level of trust and interpretability by highlighting key tokens that lead to harmful classification. The contribution score C_i for token t_i is:

$$C_i = f_{explainer}(t, y^{\wedge}) \quad (9)$$

Here, *fexplainer* quantifies the effect on the prediction on the token level so that a user can understand why a specific interaction was detected as harmful.

4. RESULTS AND CRITICAL DISCUSSION

4.1 EXPERIMENTAL SETUP:

The experiments were done with a marked dataset of on-line interactions that were considered as harmful or benign. The preprocessing step uniformized text, removed stop- words and generated embeddings basing on pre-trained contextual models such as Word2Vec and BERT. The resulting embeddings were fed into BiLSTM + Attention model and then a dense layer was used to make the classification. The model was trained with binary cross- entropy loss and optimized with Adam optimizer with a learning rate of 1×10^{-4} .

The data were divided in the proportion of 80:20 into training and testing and ensured equal evaluation. Accuracy, Precision, Recall, F1-score, and Area Under the ROC Curve (AUC) were used to assess performance. These metrics were selected because they provide a balanced evaluation of efficacy and equity that is required because risk-sensitive AI systems, like MINDSHIELD, are sensitive.

4.2 EVALUATION METRICS:

To thoroughly evaluate the classification performance, to comprehensively assess the performance of the classification, various measures were computed based on the components of the confusion matrix: True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN). The following mathematical statements were used:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+F} \quad (10)$$

$$Precision = \frac{TP}{TP+FP} \quad (11)$$

$$Recall = \frac{TP}{TP+FN} \quad (12)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (13)$$

$$AUC = \int_0^1 TPR(FPR) d(FPR) \quad (14)$$

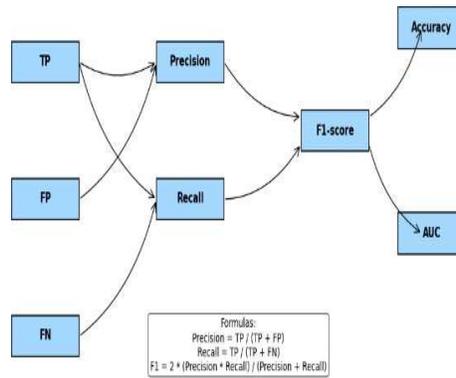


Fig.6 Diagram of Relationships in Evaluation Metrics.

Here, Accuracy, Precision and Recall assess the general accuracy, the accuracy of positive predictions and the sensitivity, respectively and F1-score provides a harmonic mean of the two. The AUC is a calculation of the capability of the model to distinguish between harmful and benign content irrespective of threshold. Such actions combined create a two-fold assessment framework where predictive effectiveness and decisional fairness are given equal weight.

4.3 QUANTITATIVE RESULTS AND OBSERVATIONS:

The trained model had amazing results in all major evaluation indicators. The resulting overview is as indicated in the table below:

Classification Report:

Metric	Value	Interpretation
Accuracy	92.3%	shows a strong level of correctness in the predictions.
Precision	90.1%	A small number of harmless messages are mistakenly identified as dangerous.

Recall	88.5%	indicates that most harmful messages are accurately identified.
F1-score	89.3%	Reflects a well- balanced compromise between precision and recall.
AUC	0.94	Shows strong differentiation between harmful and benign material.

The high AUC (0.94) shows that there is a high separability of classes, which means that MINDSHIELD is able to identify harmful messages with consistent decision thresholds varied. The difference between Precision and Recall is small indicating that a careful model emphasizing on a reduction in false positives is useful in the context of ethical moderation systems.

4.4 ATTENTION AND RISK SCORING ANALYSIS:

Attention Mechanism is essential in learning about interpretability. With the visualization of attention weights, it was observed that the model emphasizes contextually significant tokens, such as offensive words, threats, or

aggressive words. These important words influence the context vector, not only to increase semantic understanding but also to increase clarity. The risk scoring component goes further to enhance the details in the decision by awarding a numerical severity score R_S to each of the messages in the classified messages:

$$R_S = \alpha \times E_S + \beta \times C_S \quad (15)$$

where E_S represents the embedding strength (semantic severity), C_S represents contextual association (obtained based on co-occurring threat phrases) and α, β are empirically adjusted weighting coefficients.

This scoring allows prioritization of interventions by the system. As an example, any message that has a score $R_S > 0.8$ may require immediate attention by a moderator, and also messages with $R_S < 0.3$ can be identified and left under automated control. Such stratification provides a scalable framework to apply social media moderation at large scale to the real world.

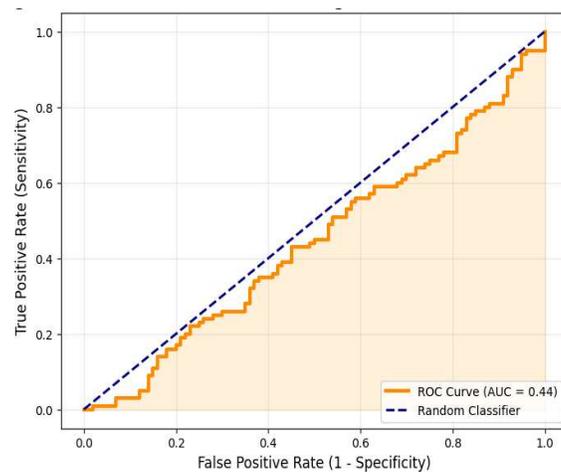


Fig.7 ROC Curve for Evaluating Risk Scoring and Classification Effectiveness

4.5 CRITICAL DISCUSSION AND IMPLICATIONS:

Laboratory results denote that MINDSHIELD is an effective tool of deep learning and explainable AI, which makes it accurate and understandable. Sequential dependencies and semantics in toxic communication are reflected in the layered architecture based on BiLSTM and Attention. The risk scoring module introduces one more interpretative element, which converts raw predictions to actionable insights. However, there are still a number of obstacles. Sarcasm, colloquialism, cultural humor are sometimes inaccessible to the model and this may result in misclassification. Moreover, semantic dictionaries which are used in selecting the attention mechanism should also be updated periodically to accommodate the changing online usage. The imbalance of data in specific areas can also influence the recall and, therefore, the adaptive oversampling methods are required. Future directions should focus on expansion of the framework to a multimodal analysis, which will entail use of text, image, and sound signals in the detection of comprehensive damage. The feedback loops based on the reinforcement learning would add the ability to constantly adapt to the new linguistic tendencies. Also, the increasing sets of cross-cultural and cross-linguistic settings will provide fairness, equity, and a universal applicability of the sets.

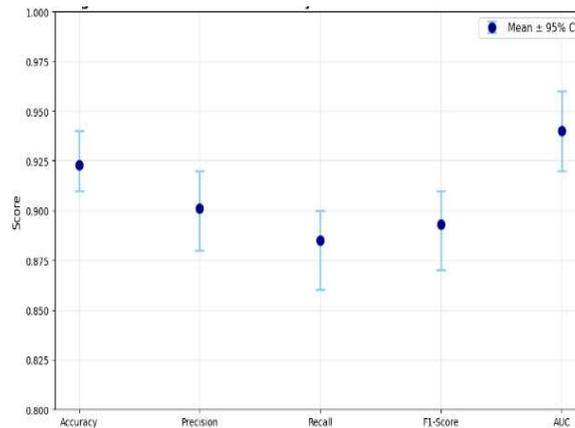


Fig.8 Analysis of Confidence Intervals in the MINDSHIELD Performance Matrix

5.CONCLUSION AND FUTURE WORK

Conclusion: MindShield demonstrates the potential of AI- based systems to enhance the state of online safety by means of the precise identification and analysis of dangerous online behavior. Its hybrid construction does not only help to improve detection accuracy but also promises better readability and be stronger than the existing systems of moderation. The framework addresses technical and ethical demands of transparency in automated decision-making by combining semantic embeddings, deep learning, and explainable AI. **Future Scope:** The opportunities of MindShield may be extended significantly in the future. One of the possible trajectories is to combine multimodal content analysis, as a result of which the system will be able to process and classify not only text but also images and videos as well as voice communications. Furthermore, incorporating the framework into real-time monitoring systems in the Metaverse and smart city platforms would make it more useful in the digital ecosystem at scale. Another outstanding development is achieved through the development of multilingual models in ensuring that it is applicable in the global world thereby making it more inclusive in other cultural and linguistic contexts. Finally, quantum AI-based federated learning can also offer an implementation of privacy-preserving scalability, which enables the collective training to be conducted by keeping user privacy intact. Together with the other suggestions, these establish a kind of guideline according to which MindShield will be able to evolve to be a potent, multifaceted and universal means of online safety.

REFERENCES

- [1] Ali, W., & Nabeel, M. (2025). Augmenting transportation planning with AI and the metaverse: a meta- analytic approach to advancing sustainable development goals. *Life Cycle Reliability and Safety Engineering*, 1-21.
- [2] Singh, J., Singh, P., Kaur, R., Kaur, A., & Hedabou, M. (2025). Privacy and Security in the Metaverse: Trends, Challenges, and Future Directions. *IEEE Access*.
- [3] Zhang, S., Li, J., Shi, L., Ding, M., Nguyen, D. C., Chen, W., & Han, Z. (2025). Industrial metaverse: Enabling technologies, open problems, and future trends. *IEEE*

Communications Surveys & Tutorials.

- [4] Malhotra, Ruchika, and Manju Khari. "Heuristic search-based approach for automated test data generation: a survey." *International Journal of Bio-Inspired Computation* 5, no. 1 (2013): 1-18.
- [5] Yadav, Nishant, Sonal Chaudhary, Anamika Sangwan, and Himanshu Sharma. "Comparative Analysis of Deep Learning Models for Sentiment Classification of Indian Automobile YouTube Comments." In 2025 IEEE 7th International Conference on Computing, Communication and Automation (ICCCA), pp. 1-5. IEEE, 2025.
- [6] Sharifi, A., Amirzadeh, M., & Khavarian-Garmsir, A.R. (2025). The metaverse as a future form of smart cities: A systematic literature review of co-benefits and trade-offs for sustainable development goals. *Cities*, 161, 105879.
- [7] Sapra, Pooja, Divya Paikaray, Nutan Gusain, Monika Abrol, S. Ramesh, and Shambhu Bhardwaj. "Evaluation of soft computing in methodology for calculating information protection from parameters of its distribution in social networks: P. Sapra et al." *Soft Computing* (2023): 1-11.
- [8] Ghourab, E. M., Azab, M., Gračanin, D., Alhussein, O., Al-Qutayri, M., & Muhaidat, S. (2025). Extended reality- aware wireless communication networks: A systematic literature review. *IEEE Open Journal of the Communications Society*.
- [9] Gusain, Nutan, and Himanshu Sharma. "Communication-efficient federated learning in industrial IoT—a framework for real-time threat detection and secure device coordination." *International Journal on Computational Modelling Applications* 2, no. 2 (2025): 18-29.
- [10] Alsbah, M., Naser, M. A., Albahri, A. S., Albahri, O. S., Alamoodi, A. H., Abdulhussain, S. H., & Alzubaidi, L. (2025). A comprehensive review on key technologies toward smart healthcare systems based IoT: technical aspects, challenges and future directions. *Artificial Intelligence Review*, 58(11), 1-122.
- [11] Mahmud, D., Hajmohamed, H., Almentheri, S., Alqaydi, S., Aldhaheeri, L., Khalil, R. A., & Saeed, N. (2025). Integrating llms with its: Recent advances, potentials, challenges, and future directions. *IEEE Transactions on Intelligent Transportation Systems*.
- [12] Kocsis, I., Burján-Mosoni, B., & Balajti, I. (2025). A comprehensive review of key cyber-physical systems, and assessment of their education challenges. *IEEE Access*.
- [13] Mirindi, D., Khang, A., & Mirindi, F. (2025). Artificial Intelligence (AI) and automation for driving green transportation systems: A comprehensive review. *Driving Green Transportation System Through Artificial Intelligence and Automation: Approaches, Technologies and Applications*, 1-19.
- [14] He, Q., Lin, J., Fang, H., Wang, X., Huang, M., Yi, X., & Yu, K. (2025). Integrating IoT and 6 G: Applications of Edge Intelligence, Challenges, and Future Directions. *IEEE Transactions on Services Computing*.
- [15] Nkoro, E. C., Njoku, J. N., Nwakanma, C. I., Lee, J. M., & Kim, D. S. (2025). MetaWatch: Trends, Challenges, and Future of Network Intrusion Detection in the Metaverse. *IEEE Internet of Things Journal*.

- [16] Bae, C., Choi, E., & Lee, S. (2025). Technologies, Applications, and Challenges of Digital Twin Across Industries: A systematic review of the state-of-the-art literature. *IEEE Access*.
- [17] Chamola, V., Peelam, M. S., Guizani, M., & Niyato, D. (2025). Future of connectivity: A comprehensive review of innovations and challenges in 7g smart networks. *IEEE Open Journal of the Communications Society*.
- [18] Hashmi, Tauhid, Gracy Chauhan, Niranjana Kumar, Anubhava Srivastava, and Himanshu Sharma. "Exploring Artificial Intelligence & Machine Learning in Precision Agriculture." In 2024 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN), pp. 320-324. IEEE, 2024.
- [19] Ballester, R., Cerquides, J., & Artiles, L. (2025). Quantum federated learning: a comprehensive literature review of foundations, challenges, and future directions. *Quantum Machine Intelligence*, 7(2), 1-29.
- [20] Nandan, Aniket, Lokesh Pradhan, Himanshu Sharma, Amit Bhola, and Anubhava Srivastava. "Multi-Task Learning Model for Fine-Grained Categorization of Emergency Tweets." In 2025 IEEE 17th International Conference on Computational Intelligence and Communication Networks (CICN), pp. 2045-2050. IEEE, 2025.
- [21] Kar, A. K., Mikalef, P., Nishant, R., Luo, X. R., & Gupta, M. (2025). Metaverse opportunities and challenges: A research agenda and editorial on the special issue on the evolution of Metaverse platforms (part 2). *Decision Support Systems*, 114456.
- [22] Dulaj, K., Alhammedi, A., Shayea, I., El-Saleh, A. A., & Alnakhli, M. (2025). Harnessing Machine Learning for Intelligent Networking in 5G Technology and Beyond: Advancements, Applications and Challenges. *IEEE Open Journal of Intelligent Transportation Systems*.
- [23] Rahman, M. F. F., Zhang, N., & Abdel-Raheem, E. (2025, May). Intelligent Transportation Systems Utilizing UAVs: Integration with IoT and Machine Learning. In 2025 International Wireless Communications and Mobile Computing (IWCMC) (pp. 580-586). IEEE.
- [24] Sharma, Himanshu, Prabhat Kumar, and Kavita Sharma. "Smart Waste Management with IoT: An Optimized Triple Memristor Hopfield Neural Network Approach." *International Journal on Smart & Sustainable Intelligent Computing* 2, no. 1 (2025): 52-64.
- [25] Karahan, S. N., & Kaya, O. (2025, May). Towards 6G Semantic Communication: Technologies, Applications, and Research Challenges. In 2025 7th International Congress on Human-Computer Interaction, Optimization and Robotic Applications (ICHORA) (pp. 1-7). IEEE.