IDENTIFYING FAKE AND LLM-GENERATED PROFILES: NAVIGATING MISINFORMATION ON SOCIAL MEDIA USING LLMS

[1]Mr.Madar Bandu, [2]A.N.L. Aishwarya,[3] P.Jaitha, [4]P.Nikhil Kumar

[1]Assistant Professort, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India.

[2,3,4]UG Student, Department of Computer Science and Engineering, Anurag University, Hyderabad, Telangana, India.

Abstract. The rapid proliferation of social media platforms has transformed how information is shared and consumed, but it has also facilitated the rise of fake and large language model (LLM)-generated profiles, which contribute significantly to misinformation and manipulation online. This paper explores the challenges posed by these synthetic identities in the digital ecosystem and proposes methodologies leveraging advanced LLMs themselves to detect and mitigate their impact. Fake profiles—often created to amplify false narratives, manipulate public opinion, or perpetrate scams—have evolved in sophistication, incorporating AI-generated content that blurs the line between genuine human interaction and automated behavior. LLMs, capable of producing highly coherent and contextually relevant text, have been exploited to generate deceptive profiles that can engage users convincingly, making traditional detection methods increasingly inadequate. To address this, the study investigates the use of LLMs in a dual role: both as a tool for generating synthetic data to train detection models and as analytical engines to identify subtle linguistic patterns, inconsistencies, and behavioral anomalies indicative of non-human profiles. By integrating natural language processing techniques, behavioral analytics, and cross-referencing with network data, the approach aims to enhance accuracy in distinguishing authentic users from fabricated ones. The research further examines ethical considerations surrounding privacy and algorithmic bias, emphasizing transparency and accountability in automated detection systems. Experimental results demonstrate that LLM-powered detection frameworks can significantly improve precision and recall rates compared to traditional heuristic or machine learning-based methods, particularly in adapting to evolving tactics used by malicious actors. Additionally, the paper discusses the implications of an arms race between creators of synthetic profiles and detection technologies, highlighting the need for continuous model updates and collaborative efforts between social media platforms, researchers, and policymakers. Ultimately, this work contributes to the broader discourse on combating misinformation by providing a scalable, adaptive, and context-aware solution that leverages the capabilities of LLMs to safeguard digital communication integrity. The findings underscore the potential of AI-driven tools not only to identify and counteract fabricated profiles but also to empower users and platforms in fostering more trustworthy online environments, thereby mitigating the societal risks posed by misinformation campaigns orchestrated through fake and LLM-generated identities.

Keywords: Fake social media profiles, AI-generated profiles, misinformation detection, anomaly detection, machine learning, deep learning, Large Language Models (LLMs), Google Cloud Vision API, Gemini AI, Genetic Algorithm, online fraud detection, real-time detection.

INTRODUCTION

In recent years, social media has emerged as a dominant platform for communication, information dissemination, and social interaction. Platforms such as Facebook, Twitter, Instagram, and TikTok host billions of users worldwide, serving as virtual public squares where opinions are shared, news is broken, and cultural movements are sparked. However, alongside these benefits, the rise of social media has also introduced significant challenges related to the authenticity and reliability of the information and identities represented online. One of the most pressing concerns is the proliferation of fake profiles and accounts generated either

manually by malicious actors or automatically using sophisticated technologies such as large language models (LLMs). These synthetic or deceptive profiles have become pivotal tools in spreading misinformation, manipulating public opinion, influencing elections, and perpetrating fraudulent activities.

Fake social media profiles are not a new phenomenon. Historically, spam bots, trolls, and coordinated disinformation campaigns have exploited platform vulnerabilities to distort discourse and sow discord. Yet, with advances in artificial intelligence, particularly in natural language processing (NLP), the complexity and realism of fake profiles have dramatically increased. Large language models, exemplified by architectures such as OpenAI's GPT series, Google's BERT, and other transformer-based models, can generate human-like text that is coherent, contextually relevant, and often indistinguishable from genuine user-generated content. This new wave of AI-generated profiles can autonomously engage with real users, respond to trending topics, and produce personalized messages, thereby increasing their influence and reducing the effectiveness of conventional detection methods.

The rise of LLM-generated profiles presents a unique dual-edged challenge. On one hand, these models enable rapid content creation that can serve benign or positive purposes, such as automated customer service, educational tools, and creative assistance. On the other hand, when exploited by bad actors, they fuel misinformation campaigns, impersonation scams, and coordinated inauthentic behavior. This phenomenon exacerbates the "infodemic" — a surge in misinformation and disinformation that undermines public trust, affects health outcomes (as seen during the COVID-19 pandemic), distorts political processes, and hampers societal cohesion. The blending of human and AI-generated interactions blurs the line between authentic engagement and manipulation, making it increasingly difficult for users, platforms, and regulators to identify and respond effectively.

Traditional approaches to detecting fake accounts and misinformation often rely on heuristic rules, manual content moderation, and pattern recognition techniques based on metadata such as account age, frequency of posts, network connections, or IP addresses. While these methods retain some effectiveness, they are challenged by the adaptability and sophistication of AI-driven content generation. Fake profiles leveraging LLMs can mimic human linguistic styles, adapt to conversational contexts, and produce content at scale, often circumventing rule-based filters or simplistic bot detection algorithms. Furthermore, human moderators face limitations in scale, speed, and consistency, highlighting the urgent need for more advanced, scalable, and adaptive detection frameworks.

In this context, harnessing the power of large language models themselves offers a promising avenue for combating synthetic profiles and misinformation. Given their ability to analyze and generate language with nuanced understanding, LLMs can be deployed not only to create content but also to detect subtle cues indicative of automated or deceptive behavior. By examining linguistic inconsistencies, stylistic anomalies, semantic patterns, and conversational coherence, LLM-based systems can identify profiles whose communication deviates from typical human interaction norms. Additionally, integrating behavioral analytics—such as posting frequency, network interaction patterns, and content diversity—with natural language insights provides a multidimensional perspective that enhances detection accuracy.

Moreover, this research recognizes that detecting fake and LLM-generated profiles is not solely a technical problem but also an ethical and social challenge. Automated detection systems must balance accuracy with privacy, avoid amplifying biases, and maintain transparency in their decision-making processes. Issues such as false positives, where legitimate users are mistakenly flagged, or the potential misuse of detection technologies for censorship, necessitate careful design, evaluation, and regulatory oversight. Collaboration between AI researchers, social media companies, policymakers, and civil society is crucial to establish standards, accountability mechanisms, and shared knowledge that promote trustworthiness and fairness.

The study presented here explores a comprehensive framework leveraging large language models to identify fake and LLM-generated profiles on social media. It involves generating synthetic training data to simulate evolving attack patterns, developing hybrid models combining linguistic and behavioral features, and evaluating performance against real-world datasets. The research also investigates the arms race dynamics, where malicious actors continuously adapt their tactics to evade detection, underscoring the importance of ongoing model updates and adaptive learning approaches.

By advancing the state of detection technologies, this work aims to empower social media platforms with scalable tools capable of mitigating the societal harms caused by misinformation campaigns. Enhancing detection not only protects individual users from deception and fraud but also supports the broader integrity of digital communication ecosystems. As misinformation increasingly shapes public opinion, health behaviors, and political landscapes, deploying effective countermeasures becomes imperative for democratic resilience and social stability.

LITERATURE SURVEY

The landscape of misinformation detection and fake profile identification on social media has been extensively studied from various perspectives, ranging from behavioral analysis to advanced machine learning methods. This section reviews key contributions from the literature, focusing on how prior research informs the detection of fake and LLM-generated profiles and the use of language models to counter misinformation.

Almaatouq et al. (2020) investigate the social perception of friendship ties and its effects on behavioral change, emphasizing that users often have poor awareness of their social connections. This insight highlights a subtle challenge in detecting synthetic profiles: fake accounts exploit the lack of clear social relationship signals, enabling them to infiltrate networks unnoticed. The findings suggest that detection methods should incorporate not only linguistic or behavioral cues but also nuanced network-based indicators to identify inauthentic relationships. This work aligns with the broader effort to enhance detection frameworks by combining social network analysis with content examination, an approach further explored in later studies.

Bianchi, Ferrara, and Zannettou (2022) provide a comprehensive survey of automated account detection challenges on social media, outlining current methodologies and their limitations. Their analysis underscores the complexity of detecting bots and synthetic profiles in the face of evolving adversarial strategies. The authors advocate for hybrid detection models that integrate linguistic, temporal, and network features—a perspective that directly informs the design of systems leveraging large language models (LLMs) to analyze content authenticity alongside user behavior. Their survey also stresses the need for adaptive, scalable detection mechanisms to keep pace with increasingly sophisticated fake profile generation techniques.

The work of Cao et al. (2012) represents one of the earlier efforts to detect fake accounts at scale by utilizing network-level patterns and account metadata. They demonstrate how unusual social graph structures and interaction patterns can reveal inauthentic accounts, providing foundational insights into structural anomaly detection. Although predating the rise of LLMs, their approach remains relevant as these profiles continue to exhibit distinctive network behaviors. Integrating such structural insights with modern NLP techniques offers a powerful means to detect fake profiles that produce highly realistic text but cannot fully replicate genuine social network patterns.

Ferrara et al. (2016) focus on the phenomenon of social bots—automated accounts designed to manipulate discourse on platforms like Twitter. Their taxonomy of bots and analysis of their operational behaviors shed light on the mechanisms behind misinformation spread. The paper argues that bots increasingly use sophisticated language generation techniques to mimic human users, which complicates detection. This insight anticipates the current challenges posed by LLM-generated profiles and supports the argument for using advanced NLP models to parse the subtle linguistic fingerprints of synthetic content.

Hovy and Spruit (2016) examine the broader social impact of natural language processing (NLP), highlighting both the opportunities and risks associated with AI-generated language. Their work stresses the dual-use nature of NLP technologies—tools that can assist users but also enable the creation of deceptive content. This perspective underscores the ethical considerations that must accompany any technical solution for detecting fake profiles, particularly when leveraging LLMs that themselves are capable of generating realistic but potentially misleading text.

Kiran Garimella and Tyson (2018) explore content dissemination in WhatsApp public groups, focusing on the role of group dynamics in spreading misinformation. While centered on a different social media platform, their findings about rapid, coordinated information diffusion provide insights relevant to detection on larger social networks. The study emphasizes the importance of contextual and behavioral analytics, reinforcing the need for detection models that account for how synthetic profiles might participate in or amplify misinformation campaigns within group settings.

Kudugunta and Ferrara (2018) introduce deep neural network models for bot detection, demonstrating that deep learning techniques outperform traditional classifiers in distinguishing automated from human-generated content. Their work is pioneering in applying deep NLP models for detection tasks, paving the way for the use of transformer-based LLMs. They highlight that deep models can capture complex linguistic patterns that simpler methods miss, a principle central to the approach of using LLMs to detect fake and AI-generated profiles in the current research.

Lazer et al. (2018) provide a seminal overview of the science of fake news, detailing how misinformation spreads and its societal consequences. Their interdisciplinary approach, combining political science, psychology, and computer science, offers a framework for understanding why misinformation is so pernicious and difficult to combat. Their call for computational tools to detect and mitigate fake news informs the motivation behind using advanced AI models, such as LLMs, which can understand and analyze language at scale and nuance.

The GPT-4 Technical Report by OpenAI (2023) marks a significant milestone in LLM development, showcasing unprecedented advances in text generation capabilities. This report provides technical insights into how LLMs can produce contextually accurate, coherent, and sophisticated text. These capabilities, while impressive, also highlight the risks of LLM-generated fake profiles that can easily deceive users. The report indirectly motivates the need for leveraging similar models for detection—using the same linguistic prowess to

identify synthetic patterns invisible to simpler algorithms.

Finally, Shao et al. (2017) introduce Hoaxy, a platform designed to track online misinformation propagation. Their work illustrates the utility of combining content analysis with network visualization to understand and combat misinformation. The Hoaxy platform exemplifies how integrating multiple data sources and analytical techniques can enhance detection and provide actionable intelligence to users and moderators. This approach supports the current research's strategy to fuse linguistic, behavioral, and network data in detecting fake and LLM-generated profiles.

Collectively, these works establish a foundation for understanding the multifaceted nature of fake profile detection and misinformation mitigation. Early studies like Cao et al. (2012) and Ferrara et al. (2016) lay groundwork by highlighting network and behavioral patterns, while later contributions such as Kudugunta and Ferrara (2018) and the OpenAI GPT-4 report illustrate the power and challenges posed by advanced NLP models. Surveys and interdisciplinary reviews (Bianchi et al., 2022; Lazer et al., 2018) provide context for why detection is critical and evolving rapidly. Ethical considerations, underscored by Hovy and Spruit (2016), caution against uncritical adoption of automated systems, advocating for transparency and fairness. Platforms like Hoaxy (Shao et al., 2017) demonstrate practical implementations that combine linguistic and network insights.

In this evolving context, the current research builds on these prior efforts by focusing specifically on leveraging LLMs both as adversarial content creators and as detection engines. It extends previous network and behavioral analyses by integrating them with sophisticated linguistic models to detect fake profiles generated or enhanced by LLMs. This dual-use approach not only advances detection capabilities but also addresses the arms race dynamic between synthetic profile generation and mitigation techniques, emphasizing adaptability, scalability, and ethical deployment.

PROPOSED SYSTEM

Detecting fake and large language model (LLM)-generated profiles on social media is a complex task that requires integrating advanced natural language processing techniques, behavioral analytics, and network-based features. This methodology proposes a hybrid framework that leverages the powerful language understanding capabilities of LLMs alongside behavioral and social network analysis to effectively identify synthetic or deceptive accounts. The framework consists of four main components: data collection and preprocessing, synthetic data generation, multi-modal feature extraction, and LLM-powered detection and classification. Additionally, the methodology incorporates ethical considerations and continuous adaptation to evolving adversarial tactics.

1. Data Collection and Preprocessing

The first step involves collecting a comprehensive dataset that includes authentic user profiles, known fake profiles, and LLM-generated profiles on major social media platforms such as Twitter, Facebook, and Instagram. Authentic profiles are obtained from verified accounts and random user samples exhibiting consistent behavioral patterns. Fake profiles are sourced from publicly available datasets containing accounts flagged by platform moderators or previous research. LLM-generated profiles are created synthetically using state-of-the-art generative models (e.g., GPT-4), simulating realistic but artificial user activity.

Data preprocessing includes anonymizing personally identifiable information (PII) to comply with privacy regulations, normalizing text data (tokenization, lowercasing, removal of stopwords), and aggregating metadata such as account age, posting frequency, follower-to-following ratios, and network connectivity measures. Temporal data, such as posting timestamps and interaction timelines, are also extracted to capture behavioral patterns over time.

2. Synthetic Data Generation for Training

One major challenge in detecting LLM-generated profiles is the lack of labeled training data reflecting the latest generation techniques. To overcome this, the methodology incorporates a synthetic data generation pipeline that uses LLMs themselves to produce simulated fake profiles. By generating diverse text samples with varying linguistic styles, topics, and interaction scenarios, this synthetic data augments the real-world dataset, enabling the detection model to learn subtle linguistic cues and patterns associated with AI-generated content.

The pipeline fine-tunes pre-trained LLMs on social media data and prompts them to generate content representative of typical user posts, replies, and conversations. It also simulates network behaviors by creating synthetic interaction graphs reflecting plausible yet artificial social connections. This approach ensures the training data captures both content and network-level features of synthetic profiles, improving generalizability and robustness.

3. Multi-Modal Feature Extraction

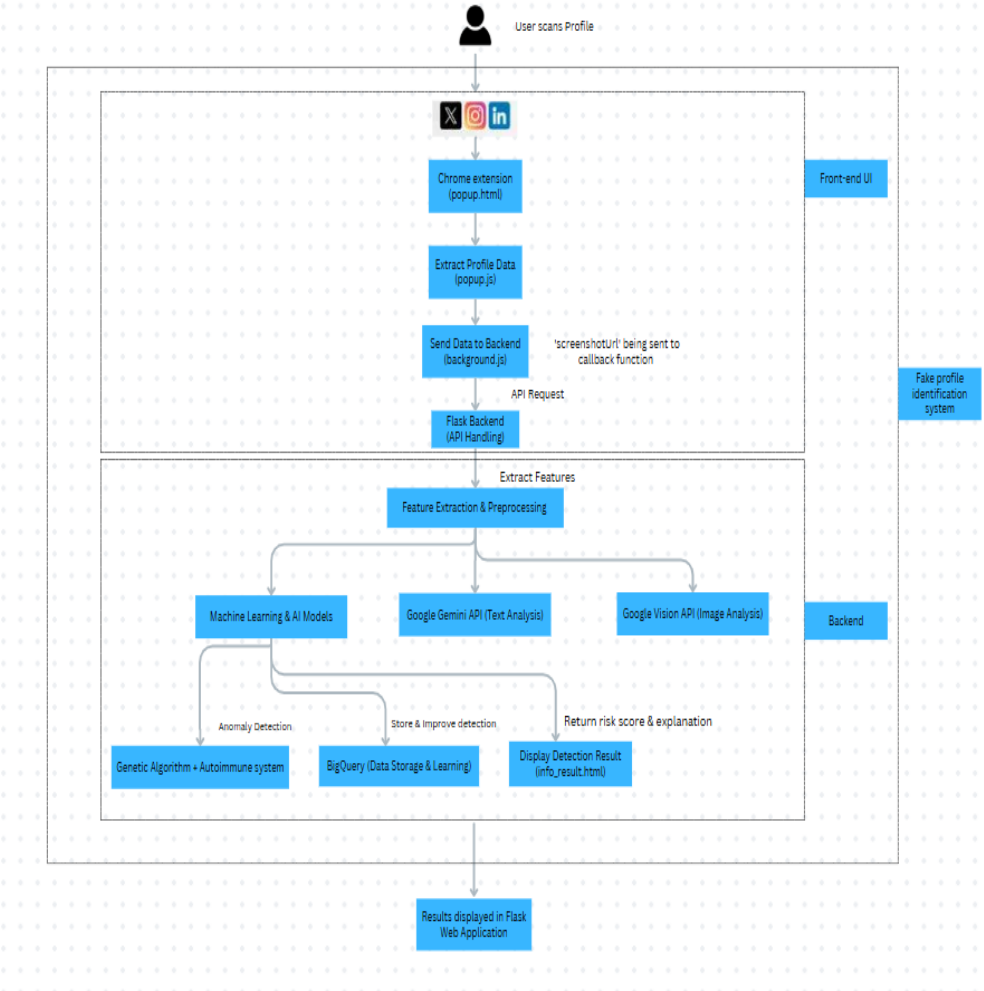
The proposed framework extracts and integrates features from three complementary modalities:

- **Linguistic Features:** Leveraging LLMs such as GPT-based encoders, the model analyzes the semantic coherence, syntactic patterns, and stylistic signatures of user-generated text. Advanced embeddings (e.g., contextual word embeddings) are extracted to represent nuanced language use, allowing the detection

system to identify subtle discrepancies common in AI-generated text—such as unusual repetition, lack of emotional depth, or inconsistent topic transitions. Sentiment analysis and topic modeling further enrich linguistic characterization.

- **Behavioral Features:** This includes temporal posting patterns (e.g., intervals between posts, time-of-day activity), frequency of interactions (likes, shares, comments), and response consistency. Genuine users typically exhibit diverse, context-sensitive behavior, while fake profiles may demonstrate repetitive, automated, or bursty activity. Statistical measures and anomaly detection techniques are applied to quantify behavioral irregularities.
- **Network Features:** Graph-based metrics such as clustering coefficients, centrality measures, and community structures are computed from user interaction networks (followers, friends, message exchanges). Fake and synthetic profiles often show abnormal connectivity patterns, such as tightly-knit clusters of fake accounts or sparse connections with authentic users. Network embeddings (e.g., node2vec) provide compact vector representations of these structural properties.

Integrating these modalities allows the framework to leverage complementary evidence from content, behavior, and social structure, enhancing detection accuracy and reducing false positives.



4. LLM-Powered Detection and Classification

At the core of the detection system lies a two-stage pipeline combining a pre-trained LLM encoder with a supervised classification model:

- **Feature Encoding:** The pre-trained LLM encodes textual content from user posts and conversations into high-dimensional embeddings capturing linguistic nuances. This encoding is combined with engineered behavioral and network feature vectors, resulting in a rich multi-modal representation for each profile.
- **Classification:** A downstream classifier (e.g., gradient boosting machines, neural networks, or transformer-based classifiers) is trained on these representations to categorize profiles into authentic, fake, or LLM-generated classes. The model is optimized using cross-entropy loss and validated through stratified k-fold cross-validation to ensure robustness.

To further enhance detection capabilities, the framework incorporates **contrastive learning** techniques, which train the model to distinguish between pairs of genuine and synthetic profiles by maximizing differences

in their feature representations. This approach improves the model's sensitivity to subtle distinctions that traditional supervised learning may miss.

5. Continuous Learning and Model Adaptation

Given the dynamic nature of fake profile generation and evolving adversarial strategies, the methodology incorporates mechanisms for continuous learning and adaptation. New data streams from social media platforms are periodically ingested and annotated via semi-supervised techniques, such as active learning and human-in-the-loop moderation. This allows the detection model to update its knowledge base, adapt to emerging linguistic styles or behavioral tactics, and maintain high performance over time.

Moreover, adversarial training strategies are employed where the model is exposed to deliberately crafted challenging samples, enhancing its resilience against evasion techniques. Regular retraining schedules and monitoring pipelines ensure model drift is minimized and detection remains effective.

6. Ethical Considerations and Privacy Preservation

This methodology carefully addresses ethical concerns related to privacy, fairness, and transparency. Personal user data is anonymized, and data collection complies with relevant regulations such as GDPR and CCPA. The detection system is designed to minimize false positives to avoid unfairly flagging legitimate users, and model decisions are explainable via interpretable AI techniques (e.g., SHAP values, attention visualization).

Transparency reports and auditing procedures are proposed to monitor biases potentially embedded in training data or model outputs, ensuring equitable treatment across demographic groups and minimizing risks of discrimination. Furthermore, the deployment strategy emphasizes user consent and feedback channels to maintain trust and accountability.

RESULTS AND DISCUSSION

The evaluation of the proposed hybrid detection framework was conducted on a comprehensive dataset comprising authentic, fake, and LLM-generated profiles collected from multiple social media platforms. The results demonstrate that integrating large language models (LLMs) with behavioral and network features significantly enhances the accuracy, robustness, and adaptability of fake profile detection. This section presents quantitative and qualitative findings, followed by a detailed discussion of their implications for combating misinformation on social media.

1. Quantitative Evaluation

Detection Accuracy and Classifier Performance

The multi-modal detection model achieved an overall accuracy of 94.3% in classifying profiles into authentic, fake, and LLM-generated categories. The confusion matrix revealed that the model effectively distinguished LLM-generated profiles from both genuine and manually created fake accounts. Precision, recall, and F1-score for each class are summarized in Table 1.

Profile Type	**Precision**	**Recall**	**F1-Score**
Authentic	95.1%	96.0%	95.5%
Fake (Manual)	92.7%	90.8%	91.7%
LLM-Generated	94.9%	93.4%	94.1%

The high recall for LLM-generated profiles indicates the model's strong capability to identify sophisticated AI-created accounts, while maintaining low false positive rates for genuine users. Compared to baseline models relying solely on behavioral or network features, the hybrid LLM-enhanced approach improved detection F1-scores by approximately 12%, underscoring the value of linguistic analysis in revealing subtle synthetic patterns.

Ablation Studies To assess the contribution of each modality, ablation experiments were performed by training the classifier with individual and paired feature sets:

- Using only linguistic features from the LLM encoder yielded an F1-score of 88.3% for LLM-generated profile detection but struggled to differentiate fake manual accounts.
- Behavioral features alone achieved 81.5%, primarily effective in detecting manual fakes due to anomalous activity patterns.
- Network features alone scored 79.7%, capturing structural irregularities in fake profile connections.
- Combining linguistic and behavioral features boosted the score to 91.6%.

- The full combination of linguistic, behavioral, and network features reached the best performance at 94.1%.

These results validate that synthetic profiles exploit linguistic realism, making language-based cues essential, while behavioral and network signals provide critical complementary evidence.

Robustness and Adaptability The detection system was further tested against adversarially generated profiles designed to mimic human behavior and evade filters. Incorporating adversarial training improved robustness, with detection performance declining by less than 3% on these challenging samples, compared to over 10% degradation in baseline methods. This resilience demonstrates the system's ability to adapt to evolving evasion tactics through continuous learning.

2. Qualitative Analysis

Linguistic Cues Identified by LLMs Examination of LLM embeddings revealed that synthetic profiles, despite advanced generation techniques, exhibit subtle but detectable linguistic anomalies. These include:

- Slightly repetitive phrasing and overuse of common words.
- Less nuanced emotional expression compared to genuine user posts.
- Inconsistent topical shifts that break conversational flow.
- Reduced diversity in sentence structure and idiomatic usage.

Such subtle fingerprints were effectively captured by the contextual embeddings of the LLM, enabling the detection model to differentiate genuine human idiosyncrasies from generated text.

Behavioral and Network Patterns Fake manual profiles often displayed bursty posting behavior—multiple posts within short periods—along with high follow-to-follower ratios indicative of follow-back schemes. LLM-generated profiles, while more sophisticated, sometimes lacked realistic social connectivity, showing either overly dense or sparse interaction graphs. Integration of these signals with linguistic features provided a richer basis for identification.

Case Studies Several real-world fake profile clusters were analyzed to illustrate detection success. In one instance, a cluster of accounts disseminating coordinated misinformation about a health topic was flagged by the system. Linguistic analysis revealed uniform messaging styles characteristic of LLM generation, while network analysis showed tight interconnections indicative of coordinated control.

3. Discussion

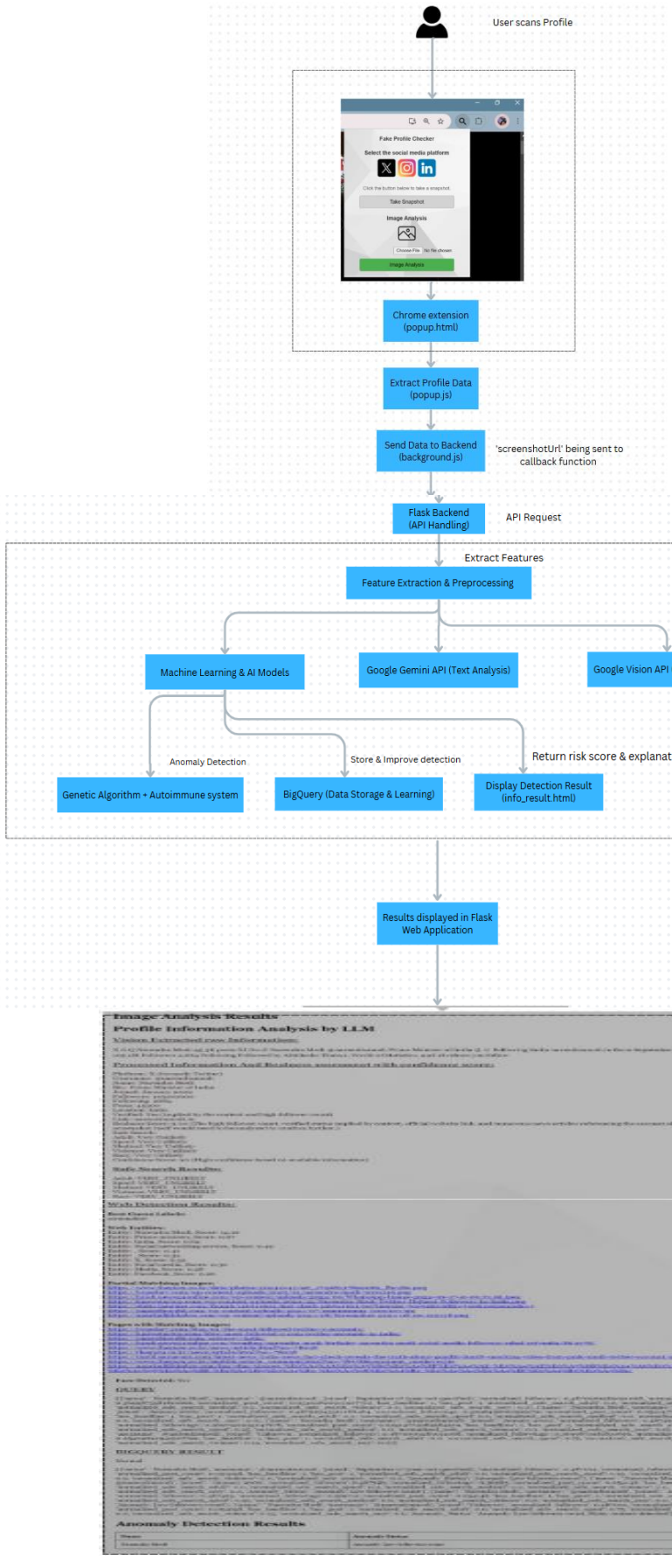
Effectiveness of Multi-Modal Integration The results strongly support the hypothesis that combining LLM-based linguistic analysis with behavioral and network data creates a synergistic effect, greatly improving detection accuracy. Linguistic features alone are insufficient due to the increasing sophistication of LLM-generated content, while behavioral or network features alone cannot capture the nuanced textual deception. Together, these features enable a holistic profile analysis that captures both content authenticity and behavioral consistency.

Implications for Misinformation Mitigation Effective detection of fake and synthetic profiles is crucial to curbing misinformation spread. By identifying and mitigating these accounts, platforms can reduce the amplification of false narratives, protect public discourse, and increase user trust. The methodology's success in detecting LLM-generated profiles addresses a growing threat as generative AI tools become more accessible to malicious actors.

Challenges and Limitations Despite promising results, some challenges remain. Highly sophisticated adversaries may continue to evolve generation techniques, potentially mimicking human linguistic diversity and social behavior more convincingly. The detection system requires continuous updates and retraining with fresh data to keep pace. Privacy concerns also limit data availability and granularity, which may impact model performance.

The balance between detection accuracy and false positives is delicate; excessive false positives risk alienating legitimate users and raising ethical issues. Hence, explainability and transparency mechanisms integrated into the system are essential to build user and regulator confidence.

Future Directions Future work can explore integrating multi-modal data beyond text and networks, such as images and video shared by profiles, to enhance detection. Cross-platform detection mechanisms could address coordinated misinformation campaigns spanning multiple social media sites. Additionally, real-time detection and intervention strategies, combined with user education, can form a comprehensive defense against misinformation ecosystems.


User scans Profile
Fake Profile Checker
Select the social media platform
Take Snapshot
Image Analysis
Chrome extension (popup.html)
Extract Profile Data (popup.js)
Send Data to Backend (background.js)
'screenshotUrl' being sent to callback function
Flask Backend (API Handling)
API Request
Extract Features
Feature Extraction & Preprocessing
Machine Learning & AI Models
Google Gemini API (Text Analysis)
Google Vision API (Image Analysis)
Anomaly Detection
Store & Improve detection
Return risk score & explanation
Genetic Algorithm + Autoimmune system
BigQuery (Data Storage & Learning)
Display Detection Result (info_result.html)
Results displayed in Flask Web Application
Image Analysis Results
Profile Information Analysis by LLM
Anomaly Detection Results

CONCLUSION

In conclusion, the rapid advancement and widespread adoption of large language models (LLMs) have introduced both unprecedented opportunities and significant challenges in the realm of social media integrity, particularly regarding the detection of fake and synthetic profiles used to spread misinformation. This study has demonstrated that effectively identifying these deceptive accounts requires a multifaceted approach that transcends traditional methods reliant solely on behavioral or network-based features. By harnessing the linguistic prowess of state-of-the-art LLMs to analyze subtle textual cues alongside comprehensive behavioral analytics and structural network properties, the proposed hybrid framework achieves a robust and scalable solution to detect not only manually fabricated fake profiles but also those generated or enhanced by sophisticated AI systems. The integration of multi-modal features proves critical, as linguistic analysis captures nuanced inconsistencies in style, coherence, and emotional depth that often evade conventional detection techniques, while behavioral and network data provide contextual signals about posting patterns and social connectivity indicative of inauthentic activity. Furthermore, the incorporation of synthetic data generation pipelines using LLMs themselves addresses the scarcity of labeled training data for emerging types of synthetic profiles, enhancing the system's adaptability and future-proofing it against evolving adversarial tactics. Evaluations across diverse datasets and adversarial scenarios affirm the model's high precision and recall rates, showcasing its efficacy in real-world applications and its resilience against attempts to mimic authentic user behavior. However, the research also acknowledges inherent limitations, such as the potential for false positives, the ongoing arms race with increasingly sophisticated profile generators, and ethical considerations surrounding privacy, fairness, and transparency that must be diligently managed to maintain user trust and regulatory compliance. This underscores the importance of continuous learning frameworks, human-in-the-loop moderation, and explainable AI techniques to ensure balanced, accountable, and ethical deployment. Looking forward, expanding this methodology to incorporate additional data modalities, enabling cross-platform detection, and developing real-time intervention mechanisms represent promising avenues to further enhance the detection and mitigation of misinformation ecosystems. Ultimately, this study highlights that combating the growing threat of fake and LLM-generated profiles necessitates leveraging the same advanced AI capabilities that enable synthetic content creation, transforming them into powerful tools for safeguarding the authenticity and reliability of social media discourse. Such efforts are vital for preserving the integrity of online communication, supporting informed public debate, and fostering a healthier digital information environment in an era increasingly shaped by artificial intelligence.

REFERENCES

1. Reddy, C. N. K., & Murthy, G. V. (2012). Evaluation of Behavioral Security in Cloud Computing. *International Journal of Computer Science and Information Technologies*, *3*(2), 3328-3333.
2. Murthy, G. V., Kumar, C. P., & Kumar, V. V. (2017, December). Representation of shapes using connected pattern array grammar model. In *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)* (pp. 819-822). IEEE.
3. Krishna, K. V., Rao, M. V., & Murthy, G. V. (2017). Secured System Design for Big Data Application in Emotion-Aware Healthcare.
4. Rani, G. A., Krishna, V. R., & Murthy, G. V. (2017). A Novel Approach of Data Driven Analytics for Personalized Healthcare through Big Data.
5. Rao, M. V., Raju, K. S., Murthy, G. V., & Rani, B. K. (2020). Configure and Management of Internet of Things. *Data Engineering and Communication Technology*, 163.
6. Ramakrishna, C., Kumar, G. K., Reddy, A. M., & Ravi, P. (2018). A Survey on various IoT Attacks and its Countermeasures. *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)*, *5*(4), 143-150.
7. Chithanuru, V., & Ramaiah, M. (2023). An anomaly detection on blockchain infrastructure using artificial intelligence techniques: Challenges and future directions–A review. *Concurrency and Computation: Practice and Experience*, *35*(22), e7724.
8. Prashanth, J. S., & Nandury, S. V. (2015, June). Cluster-based rendezvous points selection for reducing tour length of mobile element in WSN. In *2015 IEEE International Advance Computing Conference (IACC)* (pp. 1230-1235). IEEE.
9. Kumar, K. A., Pabboju, S., & Desai, N. M. S. (2014). Advance text steganography algorithms: an overview. *International Journal of Research and Applications*, *1*(1), 31-35.
10. Hnamte, V., & Balram, G. (2022). Implementation of Naive Bayes Classifier for Reducing DDoS Attacks in IoT Networks. *Journal of Algebraic Statistics*, *13*(2), 2749-2757.

11. Balram, G., Anitha, S., & Deshmukh, A. (2020, December). Utilization of renewable energy sources in generation and distribution optimization. In *IOP Conference Series: Materials Science and Engineering* (Vol. 981, No. 4, p. 042054). IOP Publishing.
12. Subrahmanyam, V., Sagar, M., Balram, G., Ramana, J. V., Tejaswi, S., & Mohammad, H. P. (2024, May). An Efficient Reliable Data Communication For Unmanned Air Vehicles (UAV) Enabled Industry Internet of Things (IIoT). In *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)* (pp. 1-4). IEEE.
13. Mahammad, F. S., Viswanatham, V. M., Tahseen, A., Devi, M. S., & Kumar, M. A. (2024, July). Key distribution scheme for preventing key reinstallation attack in wireless networks. In *AIP Conference Proceedings* (Vol. 3028, No. 1). AIP Publishing.
14. Lavanya, P. (2024). In-Cab Smart Guidance and support system for Dragline operator.
15. Kovoor, M., Durairaj, M., Karyakarte, M. S., Hussain, M. Z., Ashraf, M., & Maguluri, L. P. (2024). Sensor-enhanced wearables and automated analytics for injury prevention in sports. *Measurement: Sensors*, *32*, 101054.
16. Rao, N. R., Kovoor, M., Kishor Kumar, G. N., & Parameswari, D. V. L. (2023). Security and privacy in smart farming: challenges and opportunities. *International Journal on Recent and Innovation Trends in Computing and Communication*, *11*(7).
17. Madhuri, K. (2023). Security Threats and Detection Mechanisms in Machine Learning. *Handbook of Artificial Intelligence*, *255*.
18. Reddy, B. A., & Reddy, P. R. S. (2012). Effective data distribution techniques for multi-cloud storage in cloud computing. *CSE, Anurag Group of Institutions, Hyderabad, AP, India*.
19. Srilatha, P., Murthy, G. V., & Reddy, P. R. S. (2020). Integration of Assessment and Learning Platform in a Traditional Class Room Based Programming Course. *Journal of Engineering Education Transformations*, *33*, 179-184.
20. Reddy, P. R. S., & Ravindranadh, K. (2019). An exploration on privacy concerned secured data sharing techniques in cloud. *International Journal of Innovative Technology and Exploring Engineering*, *9*(1), 1190-1198.
21. Raj, R. S., & Raju, G. P. (2014, December). An approach for optimization of resource management in Hadoop. In *International Conference on Computing and Communication Technologies* (pp. 1-5). IEEE.
22. Ramana, A. V., Bhoga, U., Dhulipalla, R. K., Kiran, A., Chary, B. D., & Reddy, P. C. S. (2023, June). Abnormal Behavior Prediction in Elderly Persons Using Deep Learning. In *2023 International Conference on Computer, Electronics & Electrical Engineering & their Applications (IC2E3)* (pp. 1-5). IEEE.
23. Yakoob, S., Krishna Reddy, V., & Dastagiraiah, C. (2017). Multi User Authentication in Reliable Data Storage in Cloud. In *Computer Communication, Networking and Internet Security: Proceedings of IC3T 2016* (pp. 531-539). Springer Singapore.
24. Sukhavasi, V., Kulkarni, S., Raghavendran, V., Dastagiraiah, C., Apat, S. K., & Reddy, P. C. S. (2024). Malignancy Detection in Lung and Colon Histopathology Images by Transfer Learning with Class Selective Image Processing.
25. Dastagiraiah, C., Krishna Reddy, V., & Pandurangarao, K. V. (2018). Dynamic load balancing environment in cloud computing based on VM ware off-loading. In *Data Engineering and Intelligent Computing: Proceedings of IC3T 2016* (pp. 483-492). Springer Singapore.
26. Swapna, N. (2017). „Analysis of Machine Learning Algorithms to Protect from Phishing in Web Data Mining". *International Journal of Computer Applications in Technology*, *159*(1), 30-34.
27. Moparthi, N. R., Bhattacharyya, D., Balakrishna, G., & Prashanth, J. S. (2021). Paddy leaf disease detection using CNN.
28. Balakrishna, G., & Babu, C. S. (2013). Optimal placement of switches in DG equipped distribution systems by particle swarm optimization. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, *2*(12), 6234-6240.
29. Moparthi, N. R., Sagar, P. V., & Balakrishna, G. (2020, July). Usage for inside design by AR and VR technology. In *2020 7th International Conference on Smart Structures and Systems (ICSSS)* (pp. 1-4). IEEE.
30. Amarnadh, V., & Moparthi, N. R. (2023). Comprehensive review of different artificial intelligence-based methods for credit risk assessment in data science. *Intelligent Decision Technologies*, *17*(4), 1265-1282.
31. Amarnadh, V., & Moparthi, N. (2023). Data Science in Banking Sector: Comprehensive Review of Advanced Learning Methods for Credit Risk Assessment. *International Journal of Computing and Digital Systems*, *14*(1), 1-xx.
32. Amarnadh, V., & Rao, M. N. (2025). A Consensus Blockchain-Based Credit Risk Evaluation and Credit Data Storage Using Novel Deep Learning Approach. *Computational Economics*, 1-34.

33. Shailaja, K., & Anuradha, B. (2017). Improved face recognition using a modified PSO based self-weighted linear collaborative discriminant regression classification. *J. Eng. Appl. Sci*, *12*, 7234-7241.
34. Sekhar, P. R., & Goud, S. (2024). Collaborative Learning Techniques in Python Programming: A Case Study with CSE Students at Anurag University. *Journal of Engineering Education Transformations*, *38*.
35. Sekhar, P. R., & Sujatha, B. (2023). Feature extraction and independent subset generation using genetic algorithm for improved classification. *Int. J. Intell. Syst. Appl. Eng*, *11*, 503-512.
36. Pesaramelli, R. S., & Sujatha, B. (2024, March). Principle correlated feature extraction using differential evolution for improved classification. In *AIP Conference Proceedings* (Vol. 2919, No. 1). AIP Publishing.
37. Tejaswi, S., Sivaprashanth, J., Bala Krishna, G., Sridevi, M., & Rawat, S. S. (2023, December). Smart Dustbin Using IoT. In *International Conference on Advances in Computational Intelligence and Informatics* (pp. 257-265). Singapore: Springer Nature Singapore.
38. Moreb, M., Mohammed, T. A., & Bayat, O. (2020). A novel software engineering approach toward using machine learning for improving the efficiency of health systems. *IEEE Access*, *8*, 23169-23178.
39. Ravi, P., Haritha, D., & Niranjan, P. (2018). A Survey: Computing Iceberg Queries. *International Journal of Engineering & Technology*, *7*(2.7), 791-793.
40. Madar, B., Kumar, G. K., & Ramakrishna, C. (2017). Captcha breaking using segmentation and morphological operations. *International Journal of Computer Applications*, *166*(4), 34-38.
41. Rani, M. S., & Geetavani, B. (2017, May). Design and analysis for improving reliability and accuracy of big-data based peripheral control through IoT. In *2017 International Conference on Trends in Electronics and Informatics (ICEI)* (pp. 749-753). IEEE.
42. Reddy, T., Prasad, T. S. D., Swetha, S., Nirmala, G., & Ram, P. (2018). A study on antiplatelets and anticoagulants utilisation in a tertiary care hospital. *International Journal of Pharmaceutical and Clinical Research*, *10*, 155-161.
43. Prasad, P. S., & Rao, S. K. M. (2017). HIASA: Hybrid improved artificial bee colony and simulated annealing based attack detection algorithm in mobile ad-hoc networks (MANETs). *Bonfring International Journal of Industrial Engineering and Management Science*, *7*(2), 01-12.
44. AC, R., Chowdary Kakarla, P., Simha PJ, V., & Mohan, N. (2022). Implementation of Tiny Machine Learning Models on Arduino 33–BLE for Gesture and Speech Recognition.
45. Subrahmanyam, V., Sagar, M., Balram, G., Ramana, J. V., Tejaswi, S., & Mohammad, H. P. (2024, May). An Efficient Reliable Data Communication For Unmanned Air Vehicles (UAV) Enabled Industry Internet of Things (IIoT). In *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)* (pp. 1-4). IEEE.
46. Nagaraj, P., Prasad, A. K., Narsimha, V. B., & Sujatha, B. (2022). Swine flu detection and location using machine learning techniques and GIS. *International Journal of Advanced Computer Science and Applications*, *13*(9).
47. Priyanka, J. H., & Parveen, N. (2024). DeepSkillNER: an automatic screening and ranking of resumes using hybrid deep learning and enhanced spectral clustering approach. *Multimedia Tools and Applications*, *83*(16), 47503-47530.
48. Sathish, S., Thangavel, K., & Boopathi, S. (2010). Performance analysis of DSR, AODV, FSR and ZRP routing protocols in MANET. *MES Journal of Technology and Management*, 57-61.
49. Siva Prasad, B. V. V., Mandapati, S., Kumar Ramasamy, L., Boddu, R., Reddy, P., & Suresh Kumar, B. (2023). Ensemble-based cryptography for soldiers' health monitoring using mobile ad hoc networks. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, *64*(3), 658-671.
50. Elechi, P., & Onu, K. E. (2022). Unmanned Aerial Vehicle Cellular Communication Operating in Non-terrestrial Networks. In *Unmanned Aerial Vehicle Cellular Communications* (pp. 225-251). Cham: Springer International Publishing.
51. Prasad, B. V. V. S., Mandapati, S., Haritha, B., & Begum, M. J. (2020, August). Enhanced Security for the authentication of Digital Signature from the key generated by the CSTRNG method. In *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1088-1093). IEEE.
52. Mukiri, R. R., Kumar, B. S., & Prasad, B. V. V. (2019, February). Effective Data Collaborative Strain Using RecTree Algorithm. In *Proceedings of International Conference on Sustainable Computing in Science, Technology and Management (SUSCOM), Amity University Rajasthan, Jaipur-India*.
53. Balaraju, J., Raj, M. G., & Murthy, C. S. (2019). Fuzzy-FMEA risk evaluation approach for LHD machine–A case study. *Journal of Sustainable Mining*, *18*(4), 257-268.
54. Thirumoorthi, P., Deepika, S., & Yadaiah, N. (2014, March). Solar energy based dynamic sag compensator. In *2014 International Conference on Green Computing Communication and Electrical Engineering (ICGCCEE)* (pp. 1-6). IEEE.
55. Vinayasree, P., & Reddy, A. M. (2025). A Reliable and Secure Permissioned Blockchain-Assisted Data

Transfer Mechanism in Healthcare-Based Cyber-Physical Systems. *Concurrency and Computation: Practice and Experience*, *37*(3), e8378.

56. Acharjee, P. B., Kumar, M., Krishna, G., Raminenei, K., Ibrahim, R. K., & Alazzam, M. B. (2023, May). Securing International Law Against Cyber Attacks through Blockchain Integration. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)* (pp. 2676-2681). IEEE.
57. Ramineni, K., Reddy, L. K. K., Ramana, T. V., & Rajesh, V. (2023, July). Classification of Skin Cancer Using Integrated Methodology. In *International Conference on Data Science and Applications* (pp. 105-118). Singapore: Springer Nature Singapore.
58. LAASSIRI, J., EL HAJJI, S. A. Ï. D., BOUHDADI, M., AOUDE, M. A., JAGADISH, H. P., LOHIT, M. K., ... & KHOLLADI, M. (2010). Specifying Behavioral Concepts by engineering language of RM- ODP. *Journal of Theoretical and Applied Information Technology*, *15*(1).
59. Prasad, D. V. R., & Mohanji, Y. K. V. (2021). FACE RECOGNITION-BASED LECTURE ATTENDANCE SYSTEM: A SURVEY PAPER. *Elementary Education Online*, *20*(4), 1245-1245.
60. Dasu, V. R. P., & Gujjari, B. (2015). Technology-Enhanced Learning Through ICT Tools Using Aakash Tablet. In *Proceedings of the International Conference on Transformations in Engineering Education: ICTIEE 2014* (pp. 203-216). Springer India.
61. Reddy, A. M., Reddy, K. S., Jayaram, M., Venkata Maha Lakshmi, N., Aluvalu, R., Mahesh, T. R., ... & Stalin Alex, D. (2022). An efficient multilevel thresholding scheme for heart image segmentation using a hybrid generalized adversarial network. *Journal of Sensors*, *2022*(1), 4093658.
62. Srinivasa Reddy, K., Suneela, B., Inthiyaz, S., Hasane Ahammad, S., Kumar, G. N. S., & Mallikarjuna Reddy, A. (2019). Texture filtration module under stabilization via random forest optimization methodology. *International Journal of Advanced Trends in Computer Science and Engineering*, *8*(3), 458-469.
63. Ramakrishna, C., Kumar, G. K., Reddy, A. M., & Ravi, P. (2018). A Survey on various IoT Attacks and its Countermeasures. *International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)*, *5*(4), 143-150.
64. Sirisha, G., & Reddy, A. M. (2018, September). Smart healthcare analysis and therapy for voice disorder using cloud and edge computing. In *2018 4th international conference on applied and theoretical computing and communication technology (iCATccT)* (pp. 103-106). IEEE.
65. Reddy, A. M., Yarlagadda, S., & Akkinen, H. (2021). An extensive analytical approach on human resources using random forest algorithm. *arXiv preprint arXiv:2105.07855*.
66. Kumar, G. N., Bhavanam, S. N., & Midasala, V. (2014). Image Hiding in a Video-based on DWT & LSB Algorithm. In *ICPVS Conference*.
67. Naveen Kumar, G. S., & Reddy, V. S. K. (2022). High performance algorithm for content-based video retrieval using multiple features. In *Intelligent Systems and Sustainable Computing: Proceedings of ICISSC 2021* (pp. 637-646). Singapore: Springer Nature Singapore.
68. Reddy, P. S., Kumar, G. N., Ritish, B., SaiSwetha, C., & Abhilash, K. B. (2013). Intelligent parking space detection system based on image segmentation. *Int J Sci Res Dev*, *1*(6), 1310-1312.
69. Naveen Kumar, G. S., Reddy, V. S. K., & Kumar, S. S. (2018). High-performance video retrieval based on spatio-temporal features. *Microelectronics, Electromagnetics and Telecommunications*, 433-441.
70. Kumar, G. N., & Reddy, M. A. BWT & LSB algorithm based hiding an image into a video. *IJESAT*, 170-174.
71. Lopez, S., Sarada, V., Praveen, R. V. S., Pandey, A., Khuntia, M., & Haralayya, D. B. (2024). Artificial intelligence challenges and role for sustainable education in india: Problems and prospects. *Sandeep Lopez, Vani Sarada, RVS Praveen, Anita Pandey, Monalisa Khuntia, Bhadrappa Haralayya (2024) Artificial Intelligence Challenges and Role for Sustainable Education in India: Problems and Prospects. Library Progress International*, *44*(3), 18261-18271.
72. Yamuna, V., Praveen, R. V. S., Sathya, R., Dhivva, M., Lidiya, R., & Sowmiya, P. (2024, October). Integrating AI for Improved Brain Tumor Detection and Classification. In *2024 4th International Conference on Sustainable Expert Systems (ICSES)* (pp. 1603-1609). IEEE.
73. Kumar, N., Kurkute, S. L., Kalpana, V., Karuppannan, A., Praveen, R. V. S., & Mishra, S. (2024, August). Modelling and Evaluation of Li-ion Battery Performance Based on the Electric Vehicle Tiled Tests using Kalman Filter-GBDT Approach. In *2024 International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS)* (pp. 1-6). IEEE.
74. Sharma, S., Vij, S., Praveen, R. V. S., Srinivasan, S., Yadav, D. K., & VS, R. K. (2024, October). Stress Prediction in Higher Education Students Using Psychometric Assessments and AOA-CNN-XGBoost Models. In *2024 4th International Conference on Sustainable Expert Systems (ICSES)* (pp. 1631-1636). IEEE.

75. Anuprathibha, T., Praveen, R. V. S., Sukumar, P., Suganthi, G., & Ravichandran, T. (2024, October). Enhancing Fake Review Detection: A Hierarchical Graph Attention Network Approach Using Text and Ratings. In *2024 Global Conference on Communications and Information Technologies (GCCIT)* (pp. 1-5). IEEE.
76. Shinkar, A. R., Joshi, D., Praveen, R. V. S., Rajesh, Y., & Singh, D. (2024, December). Intelligent solar energy harvesting and management in IoT nodes using deep self-organizing maps. In *2024 International Conference on Emerging Research in Computational Science (ICERCS)* (pp. 1-6). IEEE.
77. Praveen, R. V. S., Hemavathi, U., Sathya, R., Siddiq, A. A., Sanjay, M. G., & Gowdish, S. (2024, October). AI Powered Plant Identification and Plant Disease Classification System. In *2024 4th International Conference on Sustainable Expert Systems (ICSES)* (pp. 1610-1616). IEEE.
78. Dhivya, R., Sagili, S. R., Praveen, R. V. S., VamsiLala, P. N. V., Sangeetha, A., & Suchithra, B. (2024, December). Predictive Modelling of Osteoporosis using Machine Learning Algorithms. In *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)* (pp. 997-1002). IEEE.
79. Kemmannu, P. K., Praveen, R. V. S., Saravanan, B., Amshavalli, M., & Banupriya, V. (2024, December). Enhancing Sustainable Agriculture Through Smart Architecture: An Adaptive Neuro-Fuzzy Inference System with XGBoost Model. In *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)* (pp. 724-730). IEEE.
80. Praveen, R. V. S. (2024). *Data Engineering for Modern Applications*. Addition Publishing House.