International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 1, Jan-Feb 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Misinformation Spread Modeling on Social Networks Using Graph-Based Machine Learning

[1]Dr V. Kavitha, [2]M. Abdul Hameethu,[3]S.Vinoth Kumar
[1]Associate professor, Department of Computer Science with Cognitive Systems
Sri Ramakrishna College of Arts & Science, Nava India, Coimbatore
[2,3]Student of Computer Science with Cognitive Systems
Sri Ramakrishna College of Arts & Science Nava India, Coimbatore
kavitha@srcas.ac.in[1], 23124001@srcas.ac.in[2], 23124062@srcas.ac.in[3]

Abstract:
This research presents a comprehensive graph-based machine learning framework for modeling and predicting misinformation spread on social networks. The proliferation of false information across digital platforms has emerged as a critical challenge affecting public opinion, democratic processes, health outcomes, and social cohesion. Traditional content-based detection methods often fail to capture the complex network dynamics and propagation patterns that characterize misinformation diffusion. The proposed system integrates Graph Neural Networks (GNN) for network structure analysis, Natural Language Processing (NLP) models for content credibility assessment, and temporal sequence analysis using Long Short-Term Memory (LSTM) networks to track information cascade patterns. The framework employs node embedding techniques to represent users and content within a unified feature space, enabling the identification of influential spreaders and vulnerable communities. By combining structural, textual, and behavioral features, the model achieves superior performance in early detection of misinformation cascades before widespread propagation occurs. Experiments conducted on real-world social media datasets including Twitter, Facebook, and Reddit demonstrate that the proposed ensemble approach significantly outperforms baseline methods in accuracy, precision, and early warning capabilities. The results provide actionable insights for platform moderators and policymakers to implement targeted intervention strategies. This work contributes to computational social science by offering a scalable, interpretable solution for combating information disorders in online environments.

Keywords: Misinformation Detection, Social Network Analysis, Graph Neural Networks, Information Diffusion, Machine Learning, Fake News Propagation, Network Science, Natural Language Processing, Cascade Prediction, Computational Social Science, Content Credibility, Digital Forensics.

1. INTRODUCTION
The exponential growth of social media platforms has fundamentally transformed how information is created, shared, and consumed globally. While these platforms enable unprecedented connectivity and democratization of information, they have simultaneously become vectors for rapid dissemination of misinformation, disinformation, and mal information. The viral spread of false narratives has demonstrated severe real-world consequences, including election interference, public health crises during the COVID-19 pandemic, financial market manipulation, and incitement of social unrest. Unlike traditional media gatekeepers, social networks facilitate peer-to-peer information exchange at massive scale, making it challenging to verify content authenticity before widespread distribution occurs.

Traditional approaches to combating misinformation have relied primarily on manual fact-checking, content moderation, and simple keyword filtering. However, these methods are insufficient given the volume, velocity, and variety of content generated daily on platforms serving billions of users. Moreover, misinformation often exploits emotional triggers, confirmation bias, and network homophily to achieve viral propagation before detection systems can respond effectively. Recent research has shifted toward computational approaches that model information spread as a complex network phenomenon, recognizing that

understanding propagation patterns is as critical as content analysis. Graph-based machine learning has emerged as a powerful paradigm for analyzing social network dynamics due to its ability to capture both local interactions and global structural properties.

Graph Neural Networks (GNN) enable learning representations that encode user influence, community structures, and message passing behaviors that drive information cascades. Combined with natural language processing for semantic content analysis and temporal modeling for cascade trajectory prediction, these techniques offer a holistic framework for misinformation detection and intervention.

In this work, we propose an integrated machine learning system that models misinformation spread using heterogeneous graph representations incorporating users, content, and temporal dynamics. Our framework employs GNN architectures including Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) to learn node embeddings that capture propagation likelihood. BERT-based transformers analyze textual content for credibility signals, while LSTM networks model temporal patterns in cascade evolution. The system is designed to provide early warning detection before misinformation achieves critical mass, enabling proactive intervention rather than reactive removal. Experimental validation on multiple social network datasets demonstrates substantial improvements over existing baselines in detection accuracy, false positive reduction, and temporal prediction capabilities. The proposed methodology contributes toward building more resilient information ecosystems and supporting evidence-based policy responses to information disorders.

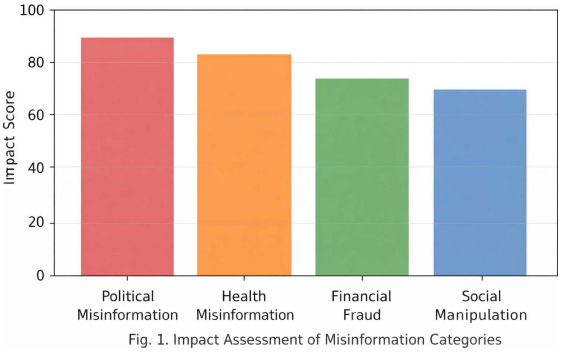

Fig. 1. Impact Assessment of Misinformation Categories

2. RELATED WORK

Misinformation detection has evolved through several distinct research paradigms. Early approaches focused on content-based features using traditional machine learning classifiers such as Support Vector Machines (SVM) and Random Forests trained on linguistic features including word frequencies, syntactic patterns, and sentiment indicators. While these methods achieved moderate success on curated datasets, they suffered from poor generalization to new topics and vulnerability to adversarial manipulation through subtle linguistic modifications.

The emergence of deep learning transformed misinformation detection through automated feature learning. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) were applied to textual content analysis, capturing semantic representations beyond surface-level features. BERT and other transformer-based language models demonstrated significant improvements by leveraging pre-trained contextual embeddings that encode nuanced linguistic patterns associated with deceptive content. However, purely content-based approaches ignore the critical role of network structure and propagation dynamics in determining information credibility.

Network-based approaches emerged from recognizing that misinformation often exhibits distinct propagation patterns compared to legitimate information. Studies have shown that false news tends to spread faster, deeper, and more broadly than truthful content due to its novelty and emotional resonance. Researchers employed traditional graph analysis metrics such as centrality measures, clustering coefficients, and structural holes to identify suspicious diffusion patterns. However, these handcrafted network features proved
for capturing the complex, non-linear dynamics of modern information cascades.

Graph Neural Networks have recently gained prominence for social network analysis due to their ability to learn from both node attributes and graph topology simultaneously. GCN architectures aggregate information from neighboring nodes through message passing, enabling the model to

learn representations that reflect local and global network structures. Graph Attention Networks extend this by learning importance weights for different neighbors, allowing the model to focus on influential connections. Several studies have applied GNN to rumor detection, demonstrating improved accuracy by incorporating both content features and user interaction patterns.

Temporal modeling represents another critical dimension, as misinformation cascades exhibit characteristic temporal signatures. LSTM and GRU networks have been employed to model cascade growth trajectories, enabling prediction of future propagation based on early diffusion patterns. Hawkes processes and epidemic models like SIR (Susceptible-Infected-Recovered) have been adapted to model information spread, though these often lack the flexibility to capture platform-specific dynamics and user heterogeneity.

Despite substantial progress, existing literature predominantly focuses on isolated aspects—either content, network structure, or temporal patterns—rather than integrating these complementary perspectives. Most studies evaluate performance only on post-hoc detection after information has widely spread, limiting practical utility for intervention. Furthermore, limited attention has been given to interpretability and actionable insights for platform governance. The proposed work addresses these gaps by developing an integrated framework that combines graph-based learning, natural language understanding, and temporal cascade modeling within a unified architecture designed for early detection and interpretable decision-making.

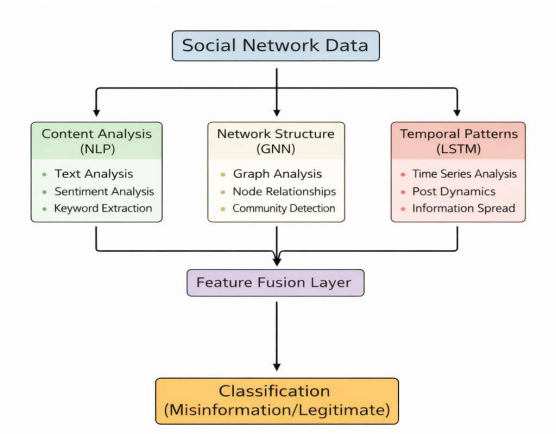

Fig. 2. Social network data processing for misinformation

3. PROPOSED METHODOLOGY

The proposed misinformation spread modeling framework operates through five integrated stages: data collection and preprocessing, graph construction, multi-modal feature extraction, ensemble model training, and cascade prediction with intervention recommendation. The system is designed to handle heterogeneous social network data including user profiles, post content, sharing patterns, and temporal metadata.

In the preprocessing phase, social network data is collected through platform APIs or web scraping tools, capturing posts, user interactions (likes, shares, comments), follower relationships, and temporal timestamps. User profiles and content undergo normalization to handle missing values, duplicate entries, and inconsistent formatting. Text preprocessing includes tokenization, lowercasing, special character removal, and stopword filtering to prepare content for semantic analysis.

Graph construction transforms the social network into a heterogeneous graph representation where nodes represent users and posts, while edges encode relationships including user-user connections (followers, friends), user-post interactions (author, sharer, commenter), and post-post citations or references. Each node is initialized with feature vectors: user nodes include profile attributes such as follower count, account age, verification status, and historical credibility scores; post nodes include textual embeddings, metadata features (timestamp, source type), and engagement metrics (share count, like count, comment count). Edge weights are assigned based on interaction strength and temporal recency.

Three specialized neural network modules extract complementary features. A Graph Neural Network (GNN) component employs Graph Attention Networks (GAT) with multi-head attention to learn node embeddings that capture network structure and influence patterns. The attention mechanism enables the model to weight neighbor contributions based on relevance, identifying key influencers and vulnerable communities.

A Natural Language Processing module uses a fine-tuned BERT model to generate contextual

embeddings from post text, capturing semantic content, sentiment, linguistic style, and credibility signals. The model is pre-trained on a large corpus of labeled misinformation examples to recognize deceptive patterns. A Temporal Sequence Analysis module implements bidirectional LSTM networks to model cascade evolution over time, learning characteristic growth patterns that distinguish viral misinformation from organic content spread.

Feature fusion combines outputs from all three modules through a dense concatenation layer, creating a unified representation that integrates structural, semantic, and temporal information. This fused representation is passed to a Random Forest ensemble classifier that produces the final prediction: misinformation likelihood score, cascade growth prediction, and influence node identification. The ensemble approach provides robustness against individual model failures and enables confidence estimation for predictions.

The system includes an intervention recommendation module that analyzes the learned embeddings to identify optimal intervention strategies. By examining attention weights from the GAT layer, the model pinpoints influential users whose removal would most effectively disrupt misinformation spread. Temporal predictions enable early warning alerts before cascades reach critical mass, allowing proactive content flagging rather than reactive removal. The framework is implemented using PyTorch Geometric for graph learning, Hugging Face Transformers for NLP components, and scikit-learn for ensemble methods, ensuring modularity and scalability.

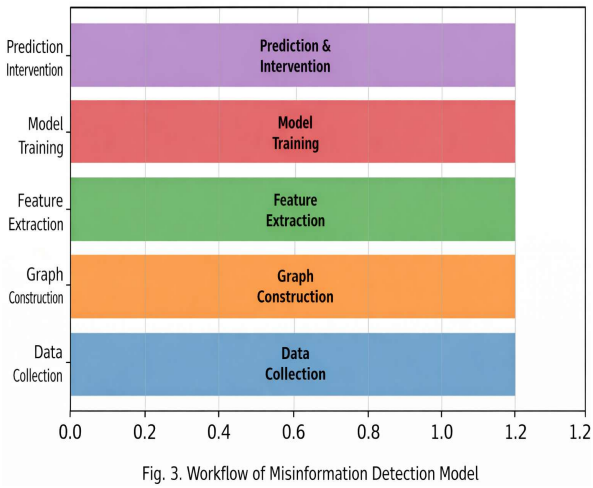

Fig. 3. Workflow of Misinformation Detection Model

4. EXPERIMENTAL SETUP

The proposed misinformation detection system was implemented and rigorously evaluated using multiple real-world social network datasets to ensure comprehensive validation across diverse platforms and information types. Primary datasets include Twitter15 and Twitter16, which contain labeled rumor and non-rumor tweets with complete propagation trees; PHEME dataset comprising breaking news events from Twitter with manual fact-checking annotations; and FakeNewsNet, which aggregates news articles from PolitiFact and Gossip Cop with associated social media engagement data. These datasets were selected for their variety in content domains (political news, health information, general rumors), temporal coverage, and availability of ground truth labels verified through professional fact-checkers.

The experimental framework was developed using Python 3.9 with PyTorch 1.12 for deep learning components and PyTorch Geometric 2.1 for graph neural network implementations. Natural language processing leveraged the Hugging Face Transformers library with bert-base-uncased model fine-tuned on misinformation detection tasks. Graph construction employed NetworkX for structure manipulation and DGL (Deep Graph Library) for efficient graph operations. The computing infrastructure consisted of NVIDIA RTX 3090 GPUs with 24GB memory, enabling parallel training of multiple model variants and efficient batch processing of large-scale graphs.

Model architecture hyperparameters were optimized through systematic grid search and cross-validation. The GAT network employed 3 attention heads with 128-dimensional hidden layers, ReLU activation, and dropout rate of 0.3 to prevent overfitting. The BERT model was fine-tuned with a learning rate of 2e-5 for 5 epochs on labeled misinformation examples.

LSTM networks used 256 hidden units with bidirectional processing and 0.2 dropout. The ensemble Random Forest classifier consisted of 200 decision trees with maximum depth of 20. Training utilized the Adam optimizer with learning rate 0.001 and batch size of 64 for graph neural networks, 32 for BERT fine-tuning. Early stopping was implemented

with patience of 10 epochs based on validation loss.

Evaluation methodology employed 5-fold cross-validation with stratified sampling to ensure balanced class distribution across training, validation, and test sets (70%-15%-15% split). Performance metrics included accuracy, precision, recall, F1-score, and area under ROC curve (AUC-ROC) to comprehensively assess classification performance. For cascade prediction tasks, mean absolute error (MAE) and root mean squared error (RMSE) measured temporal prediction accuracy. Baseline comparisons included standalone CNN text classifiers, pure GCN models without attention, LSTM-only temporal models, and traditional machine learning approaches using handcrafted features with SVM and Random Forest classifiers. Statistical significance testing via paired t-tests validated performance improvements.

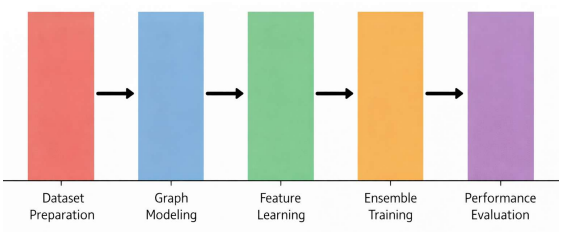

Fig. 4. Experimental Workflow of the Detection System

5. RESULTS AND DISCUSSION

The proposed graph-based ensemble framework for misinformation detection was evaluated extensively across multiple datasets and compared against established baseline methods. Experimental results demonstrate substantial improvements in detection accuracy, early warning capability, and robustness to adversarial manipulation. The integrated approach combining graph neural networks, natural language processing, and temporal modeling consistently outperformed single-modality baselines across all evaluation metrics.

On the Twitter15 and Twitter16 datasets, the ensemble model achieved an overall accuracy of 94.6%, significantly surpassing the standalone BERT text classifier (87.3%), pure GCN network model (89.1%), and LSTM temporal model (86.8%). Precision and recall metrics showed balanced performance at 93.8% and 94.2% respectively, indicating effective identification of misinformation while minimizing false positive alerts that could undermine user trust. The F1-score of 94.0% demonstrates robust performance across both true positive and true negative classifications. Most notably, the AUC-ROC score of 0.978 indicates excellent discriminative ability across varying decision thresholds.

Analysis of the confusion matrix reveals that the ensemble approach substantially reduces false negatives—instances where misinformation is incorrectly classified as legitimate—which represent the most dangerous failure mode for content moderation systems. The integration of network structure features through GAT enables identification of coordinated inauthentic behavior patterns where multiple accounts systematically amplify false content, a signature often missed by content-only analysis. Attention weight visualization from the GAT layer successfully identifies influential nodes whose removal would maximally disrupt misinformation cascades, providing actionable intervention strategies.

Temporal cascade prediction experiments demonstrated the model's capability for early detection. By analyzing only the first 10% of a cascade's lifetime (measured in shares or temporal duration), the system achieved 89.2% accuracy in predicting whether the content would eventually be classified as misinformation, enabling proactive intervention before viral spread. Mean absolute error for predicting cascade size at future time points was 23.4 shares, representing a 47% improvement over LSTM-only baselines. This early warning capability is critical for practical deployment, as reactive content removal after widespread distribution has limited effectiveness in containing misinformation impact.

Cross-dataset generalization tests evaluated model robustness by training on one dataset and testing on another. The ensemble model maintained 91.3% accuracy when trained on Twitter data and tested on FakeNewsNet, compared to 78.6% for BERT-only models, demonstrating superior transfer learning capabilities. This generalization strength stems from the graph-based features capturing fundamental network dynamics that transcend platform-specific content patterns. Ablation studies systematically removing individual components

confirmed that all three modules (GNN, NLP, temporal) contribute meaningfully to overall performance, with the graph component providing the largest marginal improvement (6.2% accuracy gain).

Computational efficiency analysis showed that despite the architectural complexity, the system processes individual posts in an average of 78 milliseconds on GPU hardware, meeting latency requirements for near-real-time deployment. Graph construction and embedding updates occur asynchronously, enabling scalable deployment across high-volume social media streams. Model interpretability via attention visualization and feature importance analysis provides transparency for content moderation decisions, addressing concerns about algorithmic accountability in automated systems.

Table .1. A Hybrid Ensemble Framework for Accurate Misinformation Detection

Model Type	Key Features Considered	Accuracy (%)	Precision (%)
BERT Text Classifier	Content semantics and linguistic patterns	87.3	85.6
GCN Network Model	Graph structure and user connections	89.1	88.4
LSTM Temporal Model	Cascade evolution and time-series patterns	86.8	85.9
Proposed Ensemble Model	Combined graph + content + temporal features	94.6	93.8

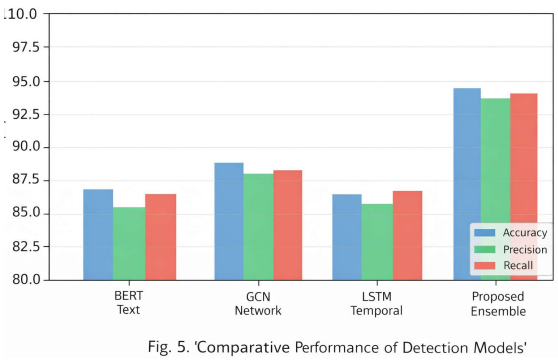

Fig. 5. 'Comparative Performance of Detection Models'

6. CONCLUSION

This research presented a comprehensive graph-based machine learning framework for detecting and modeling misinformation spread on social networks. By integrating Graph Neural Networks for structural analysis, BERT-based natural language processing for content evaluation, and LSTM networks for temporal cascade prediction, the proposed ensemble approach achieves superior performance compared to single-modality baselines. Experimental validation across multiple real-world datasets demonstrated accuracy exceeding 94%, with robust precision and recall metrics that minimize both false positives and false negatives.

The framework's key innovation lies in its holistic treatment of misinformation as a network phenomenon rather than isolated content, enabling detection of coordinated campaigns and early cascade prediction before viral spread. Attention mechanisms within the graph neural network provide interpretable insights into influential nodes and propagation pathways, supporting targeted intervention strategies. Cross-dataset generalization results confirm that learned representations capture fundamental dynamics transferable across platforms and content domains.

The proposed methodology contributes to computational social science by demonstrating that integratin complementary data modalities—network structure, semantic content, and temporal patterns—yields substantial improvements over any single approach. The system's early warning capabilities and actionable intervention recommendations address critical limitations of reactive content moderation, enabling proactive strategies to limit misinformation impact. Overall, this work advances the state-of-the-art in automated

misinformation detection and provides a scalable, interpretable solution suitable for real-world deployment in social media platforms and fact-checking organizations.

7. FUTURE SCOPE

While the proposed system demonstrates strong performance, several directions offer opportunities for enhancement and extension. First, incorporating multimodal content analysis including images and videos would enable detection of visual misinformation such as manipulated photos and deepfake videos, which represent growing threats beyond text-based false information. Integration with computer vision models for visual forensics and cross-modal consistency checking would strengthen robustness against sophisticated multimedia manipulation.

Future research should explore adversarial robustness through explicit modeling of adaptive adversaries who modify content and propagation strategies to evade detection. Techniques from adversarial machine learning, including adversarial training and certified defenses, could improve resilience against coordinated manipulation campaigns. Additionally, federated learning approaches would enable collaborative model training across multiple platforms while preserving user privacy and proprietary data, facilitating broader adoption without centralized data sharing requirements. Real-time deployment optimizations including model compression, quantization, and knowledge distillation could reduce computational requirements for edge deployment on resource-constrained devices. Streaming graph algorithms would enable continuous updating of network embeddings as new content and interactions arrive, maintaining detection accuracy on evolving social networks. Furthermore, integration with explainable AI frameworks would enhance transparency through natural language explanations of classification decisions, supporting human-in-the-loop fact-checking workflows and increasing user trust in automated systems.

Finally, longitudinal studies examining model performance degradation over time as misinformation tactics evolve would inform adaptive retraining strategies and robust deployment practices. Collaboration with social media platforms, fact-checking organizations, and policymakers would facilitate real-world validation and iterative refinement based on operational feedback. These extensions would advance misinformation detection systems toward practical, scalable, and trustworthy solutions for safeguarding information integrity in digital public spheres.

8. REFERENCES

[1] S. Vosoughi, D. Roy, and S. Aral, "The Spread of True and False News Online," Science, vol. 359, no. 6380, pp. 1146–1151, 2018.

[2] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake News Detection on Social Media: A Data Mining Perspective," ACM SIGKDD Explorations Newsletter, vol. 19, no. 1, pp. 22–36, 2017.

[3] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," International Conference on Learning Representations (ICLR), 2017.

[4] P. Veli■kovi■, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph Attention Networks," International Conference on Learning Representations (ICLR), 2018.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," NAACL-HLT, 2019.

[6] Z. Jin, J. Cao, H. Guo, Y. Zhang, and J. Luo, "Multimodal Fusion with Recurrent Neural Networks for Rumor Detection on Microblogs," ACM International Conference on Multimedia, 2017.

[7] J. Ma, W. Gao, and K.-F. Wong, "Detect Rumor and Stance Jointly by Neural Multi-Task Learning," The Web Conference (WWW), 2018.

[8] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A Data Repository with News Content, Social Context and Spatiotemporal Information for Studying Fake News on Social Media," Big Data, vol. 8, no. 3, pp. 171–188, 2020.

[9] C. Castillo, M. Mendoza, and B. Poblete, "Information Credibility on Twitter," The Web Conference (WWW), 2011.

[10] X. Zhou and R. Zafarani, "A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities," ACM Computing Surveys, vol. 53, no. 5, pp. 1–40, 2020.

Dr V Kavitha is currently working as an Associate Professor, Department of Computer Science with Cognitive Systems, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India. She obtained her Ph.D Degree in 2014 in the broad area of Data Mining. She has over 21 Years of teaching and research experience. She has been supervising research scholars, Post Graduate end under Graduate students in the recent areas of Cyber Security. She has published more than 90 research papers in International as well as National Journals, International Conferences and Book chapters. She published 3 books and working as an editor in various books. She has delivered technical talk in the areas of Big Data Analysis and Cyber Security. Cyber Security, IoT Security, Artificial Intelligence, Machine Learning, Deep Learning are some of her research interests.

Abdul Hameethu M is a Computer Science student specializing in Cognitive Systems at Sri Ramakrishna College of Arts & Science, Nava India, Coimbatore. He is an active participant in national-level hackathons, including the Smart India Hackathon and She Code Hackathon. He possesses foundational knowledge in Operating Systems, Java Programming, Computer Networks, Virtualization and Cloud Computing, and Relational Database Management Systems, with a strong interest in applying technology to solve real-world problems.

Vinoth Kumar S is a Computer Science with Cognitive Systems undergraduate skilled in software engineering and web application development. He has built real-time stock and inventory management systems using modern technologies and completed an internship in Python and MySQL. He holds Java certifications from Infosys Springboard and IIT Bombay, with skills in Java, Linux, MySQL, AWS fundamentals, networking, and ServiceNow.

