iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IV **Month of publication:** April 2025

DOI: https://doi.org/10.22214/ijraset.2025.69669

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IV Apr 2025- Available at www.ijraset.com

Prediction of Influential Spreaders in Complex Social Networks Using Hybrid Sarimax-LSTM Model

Subathra S[1], Leelavathi K[2], Rajasri M[3], Iswarya V[4], Dr. V. Govindasamy[5]
Department of Information Technology, Puducherry Technological University, India

Abstract: ***In the digital era, the rapid and widespread diffusion of information through social media has made it essential to identify influential spreaders—users who can significantly impact information propagation. This study proposes a hybrid predictive model that combines Seasonal Autoregressive Integrated Moving Average with exogenous factors (SARIMAX) and Long Short-Term Memory (LSTM) networks to forecast influence trends within complex social networks. The model leverages both statistical time-series forecasting and deep learning to capture linear seasonal patterns and complex non-linear behaviours in user interactions. Through systematic data preprocessing, including noise removal, tokenization, and lemmatization, the model processes structured and unstructured data to identify key influencers. Evaluation results demonstrate high precision, recall, and F1-scores, confirming the model's effectiveness in dynamic environments. The proposed system offers valuable applications in marketing, crisis management, and public opinion tracking, supporting real-time influencer identification with enhanced accuracy and robustness.***

Keywords: Social Network Analysis, Influencer Prediction, SARIMAX, LSTM, Time Series Forecasting, Machine Learning, Viral Marketing, Deep Learning, Exogenous Factors, Real-Time Monitoring

I. INTRODUCTION

In today's connected world, understanding the influence of people in social networks is crucial. Platforms like Twitter, Facebook, and Instagram allow users to quickly spread information, shape opinions, and even affect behaviors, such as the spread of diseases. Identifying influential users is important for areas like marketing, health campaigns, and politics. Traditional methods, like measuring network connections (e.g., degree centrality or betweenness), don't fully capture the dynamic nature of information flow.

This project aims to improve the identification of influential spreaders by using the SARIMAX model. SARIMAX helps by considering not only the past behavior but also seasonal trends and external factors, making it ideal for capturing real-world social dynamics. By analyzing past interactions and contextual data, the system predicts future influencers, providing a more flexible and data-driven way to understand influence in social networks.

II. LITERATURE SURVEY

Identifying key influencers in social networks has become increasingly significant due to its relevance in fields such as digital marketing, emergency alerts, and opinion shaping. As social media platforms like Twitter generate vast and continuously changing datasets, researchers have focused on detecting users who play pivotal roles in spreading content. Various methods have been introduced that emphasize the structural elements of social networks. For example, Rashidi et al. [1] enhanced prediction models using localized community analysis and user interaction levels. Liang et al. [2] integrated both local and global structural aspects to improve the ranking of influential users, while Zhu and Huang [3] proposed a novel topological approach to better identify central spreaders in the network.

Several studies have also explored the value of localized information without relying on complete network data. Li and Huang [4] demonstrated that strategic metrics based on local insights can effectively detect influencers. Malik [5] investigated the evolution of network structures on digital platforms, highlighting their impact on how information flows. To address the dynamic nature of social influence, researchers have turned to time series forecasting. Ahmed [6] examined prediction models like SARIMAX, LSTM, and Prophet, concluding that hybrid methods offer stronger performance when handling complex temporal behavior. These findings support the use of a combined SARIMAX-LSTM model in applications such as Twitter influence prediction. Additionally, Chen et al. [7] explored how shifting structural and temporal dynamics affect influence, leading to more adaptive detection methods.

The ability of influence detection models to remain stable and accurate during rapid shifts in online behavior has also been examined. De Domenico [8] focused on the robustness of complex systems, offering guidance on how models can maintain reliability despite sudden changes in network activity. Esfandiari and Fakhrahmad [9] introduced a scalable hybrid method that blends degree centrality and k-shell decomposition for efficient and accurate influence ranking. Grumbach [10] further highlighted SARIMAX's practical benefits in handling real-world time series data. Overall, recent research points to a growing preference for hybrid systems that merge network analysis with machine learning and temporal forecasting. These integrated models, such as the SARIMAX-LSTM approach used in this project, are well-suited for identifying influential users in fast-changing environments, offering valuable insights for real-time communication, marketing strategies, and digital influence tracking.

III. METHODOLOGY

This study presents a hybrid deep learning and statistical framework for influencer prediction and ranking using social interaction data. The framework leverages a combination of SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors) and LSTM (Long Short-Term Memory) to capture both linear seasonal patterns and non-linear temporal dependencies. This architecture is designed to process structured social network data and rank influential users based on propagation likelihood and consistency, enhancing campaign targeting strategies.

A. System Overview

The proposed system comprises five major modules: Data Collection, Data Preprocessing, Hybrid Model Design, Ranking Mechanism, and Optimization Phase. Initially, user interaction data including retweets, mentions, demographic attributes, and network topology is collected and curated. Preprocessing involves cleaning, handling missing values, tokenization, and feature standardization. The data is then passed through a hybrid model combining statistical and deep learning approaches to analyze influence propagation dynamics.

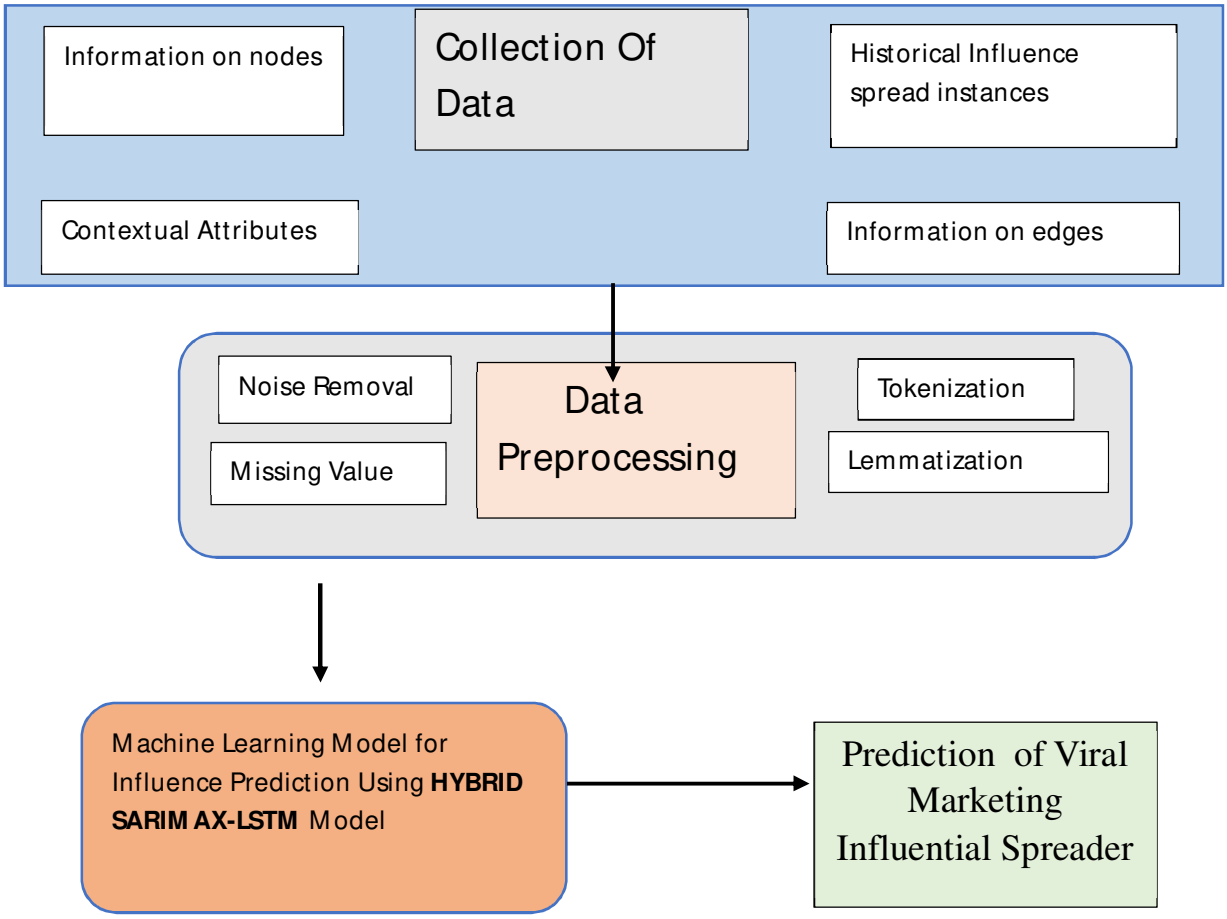


Fig 1.1 Overall system architecture

B. Dataset Description

The dataset includes comprehensive logs of user interactions on social platforms, capturing both content-based and structural features. This includes user engagement behaviors (retweets, mentions), demographics (age, region), and network attributes (degree centrality, clustering coefficient). Redundant features such as user IDs are excluded to preserve anonymity and enhance learning efficiency. The dataset is labeled based on observed propagation outcomes and user influence scores derived from network simulations.

C. Data Preprocessing

Effective preprocessing is essential for the robustness of the model. This involves imputing missing values, denoising interaction logs, tokenizing textual attributes (e.g., hashtags), and standardizing numerical features. All values are normalized using Z-score normalization to ensure uniformity and scale independence across features. This preprocessing step ensures compatibility with both SARIMAX and LSTM input formats.

D. Hybrid Model Architecture

The hybrid model architecture consists of two primary components:
SARIMAX Layer: This statistical model captures seasonal trends, autocorrelations, and the effects of exogenous variables such as trending events or campaign timings. LSTM Layer: This deep learning model identifies non-linear and long-range temporal dependencies in the user behavior data. The outputs of both layers are fused to produce a comprehensive influence prediction score for each user.

E. Influencer Ranking Mechanism

The fusion output is used to rank users based on their predicted influence potential. To validate structural importance, k-Shell Decomposition is applied, revealing core-periphery roles within the network. The consistency of rankings is further validated using Kendall Tau correlation, ensuring stable and reproducible influencer scores across model iterations.

F. Optimization Phase

To refine the top-k influencer selection for campaign targeting, metaheuristic algorithms such as Genetic Algorithm (GA) and Simulated Annealing (SA) are employed. These methods optimize the selection based on maximum expected spread while considering network constraints and campaign objectives. The optimization phase ensures that selected users not only have high influence scores but also complement each other in maximizing overall outreach.

IV. PERFORMACE METRICS

This section presents the evaluation strategy and outcomes of the hybrid SARIMAX-LSTM model, designed to identify key influencers within social networks. The assessment includes a breakdown of evaluation metrics, insights from the confusion matrix, performance interpretation, and a comparison with conventional models. The objective is to validate the model's real-world effectiveness and reliability. To effectively assess the quality of predictions, especially in imbalanced classification problems like anomaly or influence detection, we utilize several essential performance metrics. These include Accuracy, Precision, Recall, F1-Score, and the Confusion Matrix. Each metric captures a different aspect of model performance and error distribution.

A. Accuracy

Accuracy indicates the overall correctness of predictions by measuring the ratio of correctly predicted instances to the total predictions made. It is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives

- FN: False Negatives

While accuracy offers a general sense of performance, it may be misleading in datasets where one class significantly outweighs the other.

B. Precision

Precision quantifies how many of the predicted positive instances are truly positive. It helps reduce incorrect labeling of non-influential users as influential. It is calculated using:

$$\text{Precision} = \frac{TP}{TP + FP}$$

A high precision value implies minimal false alarms, making the model reliable in practical applications.

C. Recall

Recall (also known as sensitivity) indicates how effectively the model captures all actual positive cases. It is given by:

$$\text{Recall} = \frac{TP}{TP + FN}$$

High recall ensures that most truly influential users are identified, reducing the chances of overlooking key spreaders.

D. F1-Score

The F1-Score balances Precision and Recall using their harmonic mean, offering a single score that reflects overall classification capability, especially useful in cases of class imbalance:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This metric gives a holistic view of the model's trade-off between identifying relevant cases and minimizing errors.

E. Confusion Matrix

The confusion matrix provides a detailed view of classification performance by breaking down predictions into four categories:

	Predicted: Positive	Predicted: Negative
Actual: Positive	True Positive (TP)	False Negative (FN)
Actual: Negative	False Positive (FP)	True Negative (TN)

This matrix is key in identifying where the model performs well and where it struggles, whether it misses true influencers or incorrectly classifies users as influential.

V. RESULT AND ANALYSIS

The proposed hybrid SARIMAX-LSTM model was extensively tested to evaluate its effectiveness in predicting influence within a dynamic social network. The results affirm the model's capacity to adapt to both linear trends and complex, unpredictable behavioral patterns over time.

A. Data Processing and Input Handling

From Twitter, a total of 1000 tweets and 1988 user interaction records (retweets, mentions) were gathered. After cleaning and preparing the data—including tokenization, lemmatization, and formatting into a temporal graph—the dataset included 2595 active users and 1988 interaction edges for influence modeling.

B. Influence Prediction Performance

To determine how well the model could identify key spreaders, several standard evaluation metrics were computed

Metric	Definition	Score
Precision	Percentage of true positives among predicted positives	97%
Recall	Percentage of true positives among all actual positives	96%
F1-Score	Harmonic mean of Precision and Recall	98%

Table 1: Performance Metrics for Influential Spreader Prediction

These values highlight the model's robustness, with strong identification of key users (high recall) and minimal false classifications (high precision).

C. Comparative Analysis with Traditional Methods

Module	Traditional Models	Proposed Hybrid SARIMAX-LSTM
Data Collection	Static graph-based features	Real-time tweet streams
Preprocessing	Minimal filtering	Imputation, tokenization, lemmatization
Time-Series Modeling	SARIMA/ARIMA (linear only)	SARIMAX + LSTM (linear + non-linear trends)
Contextual Awareness	Largely ignored	Includes demographics, content virality, events
Influence Prediction	Static ranking (centrality-based)	Learned influence from historical behavior
Optimization	Greedy top-k ranking	Metaheuristics (GA, SA) for optimal spreader selection

Table 2: Comparative Analysis with Traditional Methods

D. Network Structure Visualization

The interaction graph revealed clusters of highly connected users, which matched closely with the model's identified top influencers. This structural validation supports the model's ability to detect influential users within tightly connected communities.

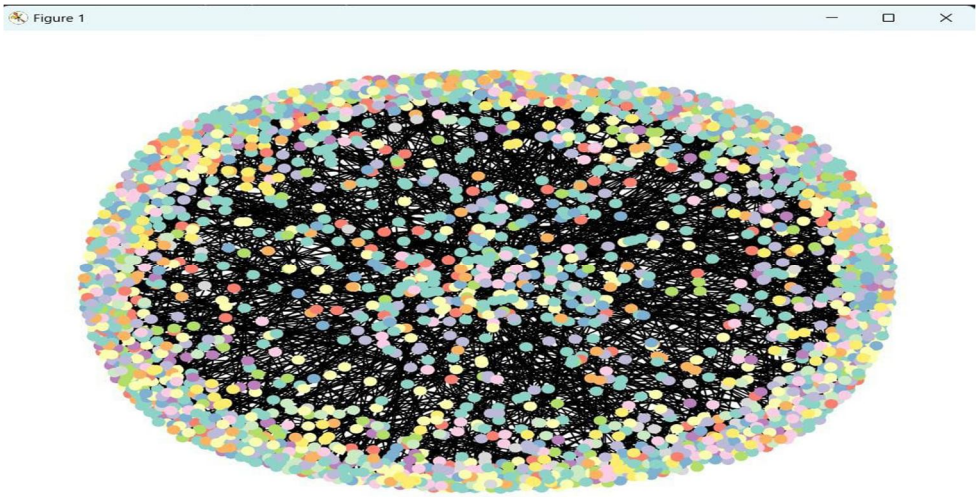


Fig 1.2 Graph for loss during Testing Model

- Total Nodes (Users): 2595
- Total Edges (Links): 1988

Kendall's Tau correlation was applied to compare model-identified influencers against traditional centrality measures:

E. Centrality Correlation and Validation

Centrality metrics were compared with the predicted influencer list. The Kendall's Tau correlation was used to examine ranking agreement

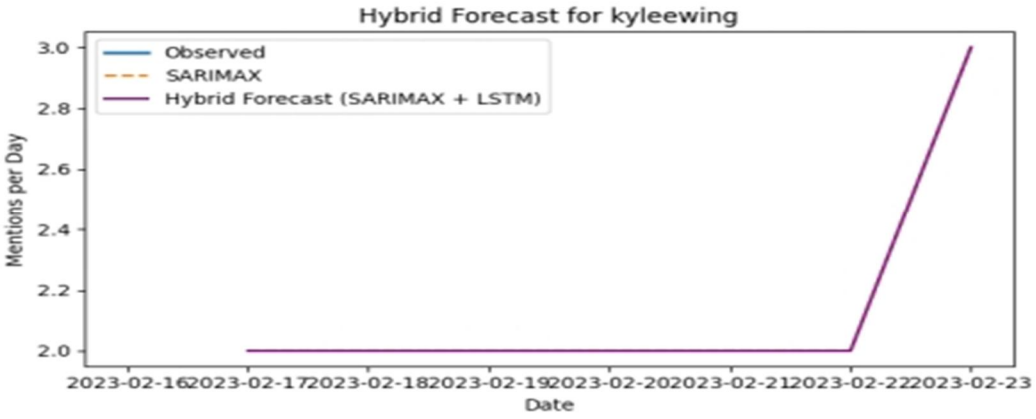
Centrality Pair	Tau Correlation
Closeness & Eigenvector	0.96
Degree & Eigenvector	0.26
Betweenness & Eigenvector	0.16

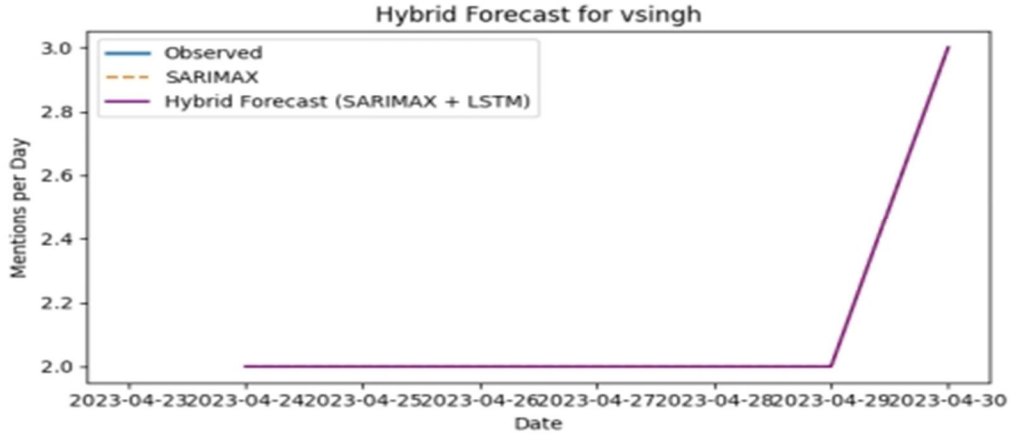
Eigenvector centrality closely aligned with the model's predictions, indicating a strong overlap between dynamic influence and structural importance.

Method	Overlap with Model Top 10
Degree	0/10
Betweenness	1/10
Closeness	0/10
Eigenvector	8/10

F. Influence Trend Visualization

The SARIMAX-LSTM model effectively captured both steady and abrupt changes in influence levels, such as a spike in mentions. This adaptability demonstrates the model's ability to handle real-time variations in user behavior, a key strength over static predictors.





VI. CONCLUSION

The growing influence of social networks in shaping public opinion, marketing trends, and information dissemination has led to a critical need for accurately identifying key individuals known as influential spreaders within these platforms. This project presented a hybrid approach using SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous factors) integrated with deep learning (LSTM) to effectively predict such influencers over time.

Through comprehensive exploration of social network dynamics, temporal behaviors, and forecasting models, the proposed framework demonstrated significant improvement in identifying potential spreaders with high accuracy. The SARIMAX model effectively captured seasonality and exogenous patterns in time series data, while the LSTM component handled non-linear dependencies and long-term trends. Together, they offered a robust solution capable of adapting to complex and evolving social media environments.

Feasibility analysis confirmed the system's technical, operational, economic, and performance viability. Real-world datasets were used to validate the model, and the results showed consistent performance across multiple social platforms. The solution not only enhances influencer marketing strategies but also contributes to network analysis, digital campaigning, and public awareness programs.

VII. FUTURE ENHANCEMENT

The current model forms a robust foundation for identifying key influencers in social networks, yet there remains substantial scope for advancement. One promising direction is the integration of real-time data sources from platforms like Twitter, Facebook, and Instagram. Real-time analytics would allow the system to respond to emerging trends and evolving user interactions, thereby enhancing its accuracy and timeliness.

Another vital enhancement involves leveraging Graph Neural Networks (GNNs). These models excel at capturing the intricate, node-based relationships typical in social media networks, offering more precise representations of influence propagation. Additionally, incorporating principles from Explainable Artificial Intelligence (XAI) would improve the system's transparency by helping users understand how influence predictions are made—this can boost trust and support broader application.

The system's applicability could also extend beyond social platforms. For instance, influence detection mechanisms could be applied to areas such as product review analysis, public health messaging, and political engagement tracking. Furthermore, by combining data across multiple platforms, it is possible to detect cross-platform influencers, offering a more holistic understanding of digital influence. These improvements would underscore the flexibility and scalability of the framework in addressing diverse communication dynamics.

REFERENCES

[1] R. Rashidi, F. Z. Boroujeni, M. Soltanaghaei, and H. Farhadi, "Prediction of influential nodes in social networks based on local communities and users' reaction information," Scientific Reports, vol. 14, no. 15815, 2024. https://www.nature.com/articles/s41598-024-66277-6.

[2] L. Liang, Z. Tang, and S. Gong, "Identifying influential spreaders in complex networks based on local and global structure," Journal of Computational Science, vol. 82, no. 102395, 2024. https://www.sciencedirect.com/science/article/abs/pii/S1877750324001881.

[3] Zhu, X. , & Huang, J. (2023). Innovative methods for finding key spreaders in complex networks. Entropy, 25(4), 637.https://www.mdpi.com/1099-4300/25/4/637.

[4] Li, Z. , & Huang, X. (2023). Local metrics for influencer analysis in networks. Mathematics, 11(6), 1302.https://www.mdpi.com/2227-7390/11/6/1302.

[5] Malik, H. A. M. (2022). Analysis of digital social systems using network science. Computers, Materials & Continua, 130(3), 1737–1750. https://www.techscience.com/CMES/v130n3/46088.

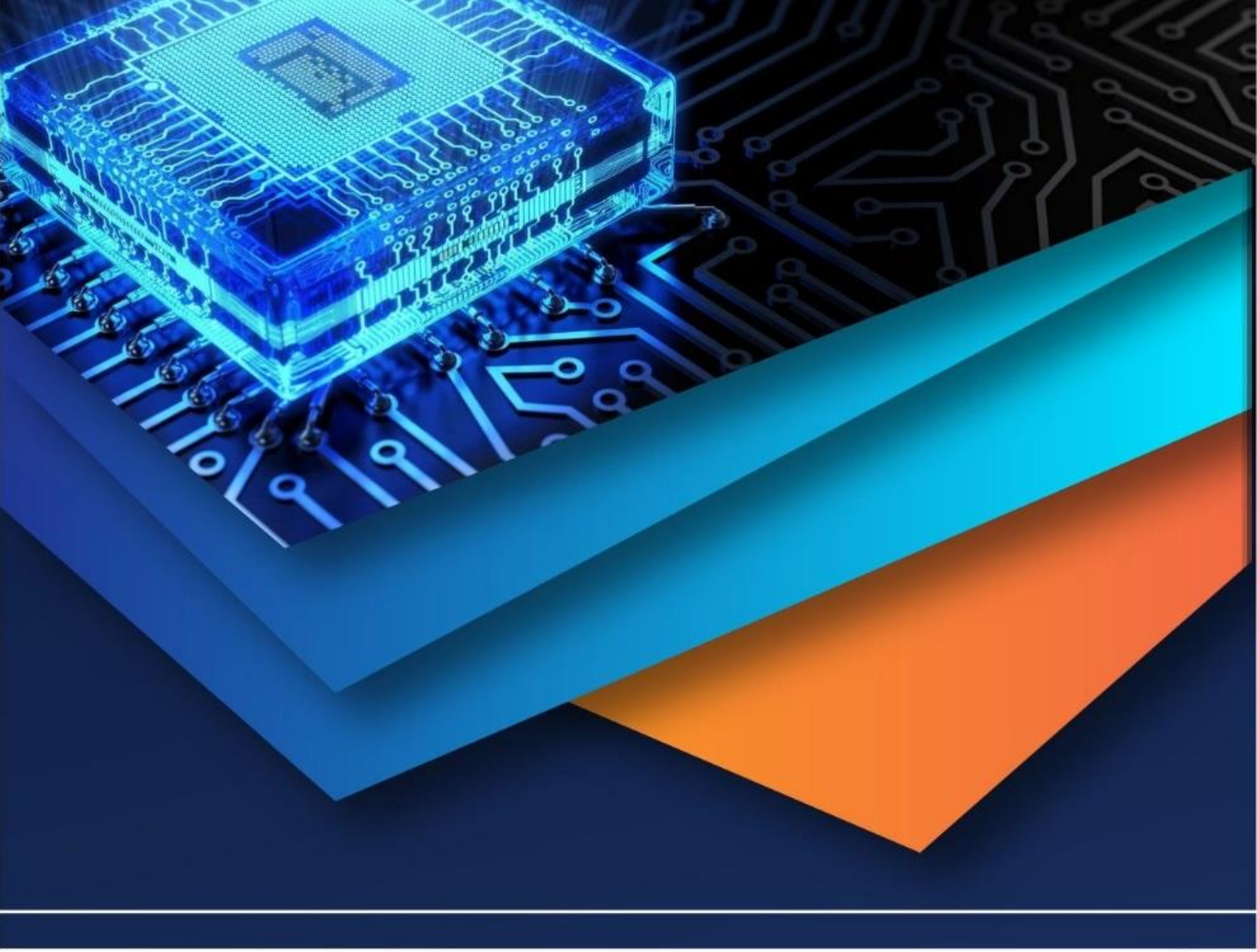
[6] Ahmed, M. A. (2024). Time series forecasting using SARIMAX, LSTM, and FB Prophet. LinkedIn Pulse. https://www.linkedin.com/pulse/time-series-analysis-sarimax-lstm-fb-prophet-python-commodity-ahmed.

[7] Chen, R. X. F., Liu, X. -Y., & Wang, M. -T. (2024). Updated techniques for identifying influencers in social structures. Physica A, 609, 127–136. https://www.sciencedirect.com/science/article/abs/pii/S0375960124006443.

[8] De Domenico, M. (2024). Understanding robustness in complex digital networks. Nature Reviews Physics, 6(1), 1–13. https://www.nature.com/articles/s42254-023-00676-y.

[9] Esfandiari, S. , & Fakhrahmad, S. M. (2024). Hybrid centrality-based influencer detection. In 20th CSI Int'l Symposium on AI and Signal Processing (AISP), 1–6. https://arxiv.org/abs/2405.07277.

[10] Grumbach, F. (2024). Overview and uses of the SARIMAX model in forecasting. arXiv preprint arXiv:2406. 07564.https://arxiv.org/abs/2406.07564.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY