INFUSE – Intelligent Fusion for Uncovering Social Network Threats

Tanushree Sarkar, Dr. R. Sunitha, and Arunmozhi M

Department of Computer Science
Pondicherry University, India

Abstract. The rapid expansion of social networking platforms has significantly transformed communication and digital interaction by allowing users to share content, build connections, and engage across global communities. These platforms, however, provide a ground for fraudulent activities such as fake account creation, phishing, spamming, and misinformation posing risks to both users and platform integrity. Fraud detection in social networks involves analyzing user behavior, network structures, and content to identify such malicious actions. Traditional rule-based systems and classical Machine Learning (ML) models, although widely adopted, often struggle with real-time adaptability, interpretability, and handling the complex, dynamic, and highly relational data inherent in these digital environments. While prior research has explored both ML and Deep Learning (DL) methods, gaps remain in scalability, contextual understanding, and cross-platform effectiveness.

To bridge these gaps, this study proposes a hybrid fraud detection framework that integrates diverse ML algorithms Random Forest(RF), XGBoost(XGB), and Support Vector Machine(SVM) with DL models such as Convolutional Neural Networks(CNN), Long Short-Term Memory (LSTM), and Graph Neural Networks (GNN). The system processes multi-modal data including user behavior, account attributes, and network interactions. The ensemble learning technique is employed to combine the outputs of individual models, enhancing prediction robustness and reducing false positives.

Experiments were conducted on three datasets: an Instagram profile dataset, the Youtube01-Psy comment dataset, and video metadata collected via the YouTube Data API. The ensemble approach achieved the highest accuracy across all datasets, on the Youtube01-Psy dataset. These results confirm the effectiveness of the hybrid approach in accurately detecting fraud across different types of social media data.

Keywords: Social Network Analysis, fraud detection, Ensemble learning, ML, DL, CNN, LSTM, GNN, multi-modal data, fake accounts, misinformation, RF, XGB, SVM, social media security.

1 Introduction

The domain of this research is an increasingly critical area in the age of digital communication. Social media platforms such as Instagram, Twitter, Facebook, and YouTube have revolutionized the way people connect, communicate, and share information. However, the very openness and reach of these platforms have also made them attractive targets for fraudulent activities. These include the creation of fake user accounts, spread of misinformation, engagement manipulation through bots or spam accounts, and malicious behavioral patterns like phishing and online scams. These threats not only compromise user privacy and trust but also damage the credibility and integrity of the platforms themselves.

While numerous fraud detection mechanisms have been introduced, existing systems still face major limitations in effectively identifying and mitigating fraudulent behavior. The dynamic and constantly evolving nature of online interactions together with the high dimensionality, volume, and diversity of social media data poses significant challenges. Traditional rule-based approaches lack adaptability and often become obsolete as new fraud tactics emerge. Classical ML models such as Naïve Bayes, Decision Trees, and even early use of SVM and RF have shown some success but are constrained in their ability to generalize across platforms and adapt to evolving patterns.

Recent advances in DL have led to the use of CNNs for feature extraction and LSTM networks for temporal analysis of user behavior. Similarly, GNN have shown potential in modeling the relational aspect of user interactions. However, despite their power, individual DL models often fall short in capturing the full complexity of multimodal and multi-relational fraud data. In addition, many of these models suffer from low interpretability, high computational cost, and limited precision, often resulting in increased false positives and inconsistent real-world performance.

To address these gaps, this research proposes a robust hybrid fraud detection framework that integrates the strengths of both ML and DL techniques. Specifically, it leverages RF, XGB, and SVM as part of the ML module to learn behavioral patterns based on user activity and engagement metrics. In parallel, CNN, LSTM, and GNN models form the DL module to uncover hidden feature interactions, sequential user patterns, and community-level fraud clusters within the network. These diverse model outputs are then fused using an ensemble learning approach, which aggregates predictions to enhance overall accuracy, reduce false positives, and provide a balanced and adaptable fraud detection mechanism.

The proposed system is evaluated using three datasets. An Instagram profile dataset, the Youtube01-Psy dataset, and a custom dataset collected via the YouTube Data API, including video metadata and user interactions.

Experimental results reveal that the ensemble approach consistently outperforms individual ML and DL models, with the highest performance observed on the Youtube01-Psy dataset where the system achieved perfect accuracy, precision, and recall. These findings validate the effectiveness of the hybrid framework in detecting diverse fraud patterns across different types of social media data.

2 Literature Survey

With the exponential rise in online interactions, social media has become a major target for fraudulent activities. Traditional rule-based systems have proven inadequate for evolving threats, leading researchers to explore intelligent solutions using ML, DL, and hybrid approaches. This review categorizes recent literature into these three approaches, outlining methodologies, findings, and limitations to uncover potential areas for improvement.

2.1 ML Models

The fake profile and fraud detection in social media using ML reveals a progressive shift from traditional methods to more sophisticated, scalable, and adaptive approaches.
Austin-Gabriel et al. [1] implemented rule-based and basic ML models for small business fraud detection but failed to address sophisticated fraud strategies. Ramdas and Neenu [2] improved the performance of models like RF and SVM through feature engineering on social media profiles, although scalability remained a challenge. Farooqui and Khan [3] applied soft computing techniques, including fuzzy logic and evolutionary algorithms, along with SVM for fake profile detection but lacked robustness with large datasets. Ahmad and Tripathi [4] reviewed RF and SVM techniques and advocated for hybrid approaches to tackle large-scale challenges. Kavin et al. [5] proposed RF and Artificial Neural Network (ANN) for secure mobile network fraud detection with scalability in mind. Pombal et al. [6] explored bias in ML models and suggested mitigation strategies. Chakraborty et al.

[7] and Meshram et al. [8] used RF, XGB, CNN, and ANN to detect fake profiles, pointing out the need for models that adapt to evolving threats. Goyal et al. [9] and Singh et al. [10] used Naïve Bayes and Decision Trees, which lacked support for complex datasets.

It is deducted that ML models are efficient and interpretable, making them suitable for basic detection tasks. However, they often rely on manual feature engineering, struggle with data complexity, and lack adaptability to real-time threats. While scalable solutions are being explored, limitations in handling relational and behavioral data persist. To address these limitations, DL models are gaining attention for their ability to learn directly from raw data and adapt more naturally to complex, real-time threat environments.

2.2 DL Models

Various DL strategies are employed to detect and prevent fraud, cybercrime, and misinformation, particularly across financial systems and social media platforms.
Zhang et al. [16] analyzed AI-based real-time fraud detection. Adekunle et al. [17] developed an LSTM model to handle complex and evolving cyberattacks on social media. Alharbi et al. [18] used multimodal DL combining text, image, and behavior for fake Instagram profile detection. Zioviris et al. [19] utilized LSTM for behavior-based fraud detection and suggested graph-based enhancements. Huang et al. [20] presented DGraph for large-scale financial anomaly detection. Hu et al. [21] proposed Behavioral Information Aggregation Network, a behavioral fraud model enhancing GNN performance. Shehnepoor et al. [22] found that DL models using content and metadata outperform behavior-only models for fake review detection. Zhang et al. [23] combined multiple data cues for detecting tax evasion on social media. Rossi et al. [24] introduced SIGN, a scalable GNN for large graphs. Shi et al. [25] proposed a semi-supervised message-passing model to handle sparse label issues. Monti et al. [26] employed geometric DL for language-independent fake news detection using user engagement graphs.

It is understood that DL models excel in extracting complex patterns from unstructured data and can process large volumes efficiently. However, they are resource-heavy, often lack interpretability, and may fail in low-data scenarios. Graph-based and multimodal DL approaches show promise, but scalability and transparency remain open challenges.

2.3 Hybrid Architecture

Hybrid achitecture combine ML and DL to improve fraud detection by leveraging ML's interpretability and DL's deep feature learning. These models handle complex, multimodal data and are better suited for dynamic social platforms.
Anila et al. [27] combined K-Nearest Neighbors(KNN) and SVM to detect fake profiles, improving accuracy on the Instagram dataset but faced limitations in feature extraction and explainability. Sharmila et al. [28] introduced Pattern-Based Deep Hate Speech Detection using tweet representation, improving accuracy but requiring better multilingual handling. Alarfaj et al. [29] applied hybrid ML-DL models for credit card fraud detection, successfully reducing false positives and boosting precision. Sansonetti et al. [30] used DL models like LSTM and CNN with traditional classifiers such as SVM and KNN to identify unreliable social media users, highlighting adaptability issues across platforms.
Hybrid architecture blends ML's interpretability with DL's representational power, offering better accuracy and generalization. They are especially useful for dynamic, multi-modal social media environments.
The literature survey shows substantial progress in detecting fraud and fake

profiles using ML, DL, and hybrid architecture. ML models are interpretable but less effective with complex data. DL approaches handle complex features but face scalability and transparency issues. The hybrid architecture undoubtedly offer a balanced approach.

There is a clear need for a scalable, explainable, and real-time fraud detection model that combines the strengths of ML and DL. Most existing approaches either compromise on interpretability or struggle with adaptability across platforms. Additionally, they often lack the ability to fully capture complex social connections, evolving temporal behaviors, and multi-modal data that characterize modern social networks. This study addresses these research gaps by proposing a hybrid ensemble-based framework that integrates relational and behavioral features to improve detection accuracy, scalability, and adaptability for evolving social media fraud patterns.

3 Proposed Architecture

The primary objective of this project is to build a hybrid architecture that integrates graph-based and sequence-based models for more effective fraud detection. The goal is to design a unified framework capable of analyzing user behavior, network interactions, and content-based features simultaneously. Additionally, the system aims to incorporate dynamic fraud tracking using temporal graph models to detect evolving fraudulent patterns over time. The model is also designed to be easy to understand, making it more useful and practical for real-world use.

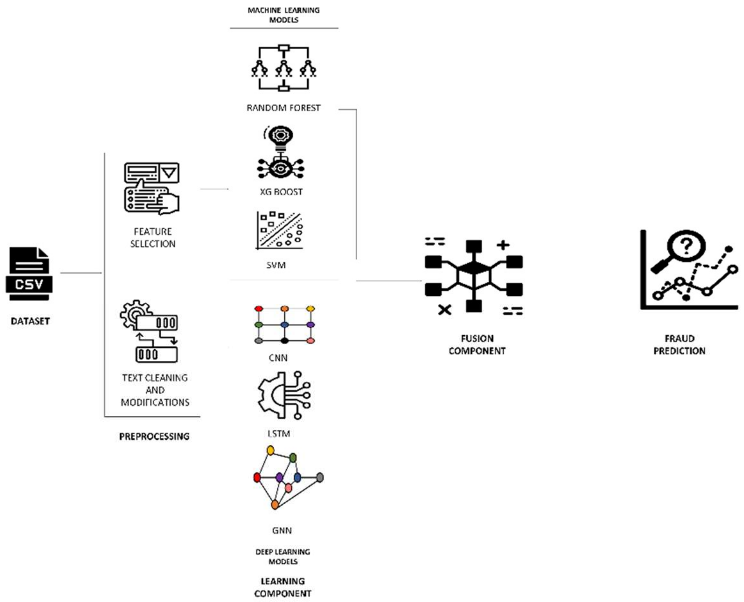


Figure 1 – Hybrid architecture of the proposed model

The proposed hybrid architecture illustrated in Figure 1 has the following components:

3.1 Dataset

To build a fraud detection system capable of handling varied real-world data, three distinct datasets were collected from reliable sources, each offering different perspectives on fraudulent behavior. These datasets were intentionally chosen for their diversity spanning profile-based, content-based, and engagement-based features and were processed individually to ensure clear and focused analysis.

Instagram Dataset - This dataset was sourced from Kaggle and contains user profile information relevant for detecting fake accounts. It includes attributes like profile picture availability, username similarity, number of posts, and follower-following ratios. It helped identify patterns often associated with fraudulent profiles like bots or impersonators, by highlighting profile inconsistencies and suspicious activity metrics.

YouTube01-Psy Dataset - This dataset was found in a public GitHub repository and contains labeled YouTube comments categorized as spam or legitimate. Focused on text data, it enabled the system to learn linguistic patterns typical of spam, such as repetitive phrases, or unnatural sentence structures that are common in comment-based scams.

YouTube Metadata Dataset - This dataset was created by scraping metadata from YouTube using the YouTube Data API. It includes video-level information like titles, descriptions, tags, view counts, like-to-comment ratios, and upload dates. This dataset provided valuable insight into behavioral and engagement-based anomalies. It was used to uncover unnatural interaction trends, such as videos with inflated view counts or inconsistent engagement metrics that could suggest manipulation or bot-driven activity.

3.2 Preprocessing

For the ML models, feature selection is performed to pick out the most relevant variables that help improve the model's accuracy and reduce noise. This ensures the models focus on the important signals for fraud detection.

For the DL models, the focus is on cleaning and preparing textual data. This includes steps like removing special characters, normalizing text, and converting it into a format suitable for sequence models. These preprocessing steps help the deep learning models better understand user behavior and content patterns.

3.3 Learning Component

The learning component forms the core of the proposed hybrid architecture. It combines the strengths of traditional ML models with the advanced capabilities of DL models to detect fraudulent behavior from different perspectives. By assigning specific models to different datasets and types of features, the system ensures specialized learning and improved detection accuracy.

3.3.1 ML Models

RF - This model was built using 100 decision trees to improve accuracy and avoid overfitting. It works well with high-dimensional structured data like follower ratios and profile statistics. The final output is an averaged score across all trees, giving a solid estimate of whether a user might be fraudulent.

XGB – It is known for its ability to find subtle patterns, XGB was used to analyze complex behavioral signals. It was fine-tuned using cross-validation, adjusting parameters like learning rate and tree depth to get the best results. This model is particularly good at handling noisy and imbalanced data.

SVM - With its strong ability to separate data points, SVM was used to draw clear boundaries between fraudulent and legitimate users. A radial basis function (RBF) kernel helped capture non-linear relationships. Class weights were balanced during training to prevent the model from being biased toward the majority class.

3.3.2 DL Models

CNN - They were applied to structured features to detect subtle patterns, such as repeated behaviors or unusually similar activity metrics. Layers of 1D convolution and pooling helped extract useful signals before making a prediction.

LSTM - They were applied to detect how user behavior changes over time. The LSTM learns from these activity sequences, helping the model spot fraud that's based on timing or behavioral flow rather than just raw numbers.

GNN - It brings in the power of relationships. Instead of looking at users in isolation, GNNs treat them as part of a larger network, where connections between accounts can reveal group behaviors or coordinated fraud. The model learns patterns at the community level, such as clusters of fake profiles interacting unusually often or forming tight-knit suspicious groups.

Each ML and DL model generated its own fraud score, and these were passed on to the fusion component for the final analysis.

3.4 Fusion Component

To integrate the outputs of various ML and DL models, the architecture employs a late fusion strategy, also known as decision-level fusion. In this approach, each model processes its respective dataset independently and produces fraud prediction scores. These outputs are then combined using an ensemble method of weighted averaging, to generate the final decision. This fusion technique ensures that the strengths of individual models, whether they focus on content, behavior, or network-level data are effectively leveraged. By aggregating these diverse perspectives, the system achieves greater robustness, adaptability, and accuracy in detecting complex and evolving fraud patterns.

3.5 Fraud Prediction

The final output of the hybrid system is a binary label, representing whether a user or behavior is flagged as fraudulent. By analyzing diverse data signals from individual features to community patterns, the model provides a multi-dimensional and interpretable decision, adaptable to real-world fraud scenarios.

4 Experimental Setup

The proposed hybrid framework is implemented and tested using the following setup.

4.1 Datasets

The project uses three key datasets. The first is an Instagram dataset split into training and testing files, containing features such as profile picture status, username similarity, number of posts, and follower-to-following ratios. The second is the Youtube01-Psy dataset, which includes labeled YouTube comments classified as spam or genuine, helping in content-based fraud detection. The third dataset is

YouTube metadata collected through the YouTube Data API, which provides video-level attributes like title, description, tags, view count, like count, and comment count, allowing for analysis based on engagement and content patterns.

Table 1- Dataset Summary

Dataset	**Total Samples**	**Training%**	**Test%**
Instagram Dataset	698	80	20
YouTube01-Psy Dataset	351	80	20
YouTube Metadata Dataset	51	75	25

The datasets represent a variety of fraud scenarios across different platforms, allowing the model to be evaluated on user-level, content-based, and interaction-based fraud. Each model was trained and evaluated using the evaluation metrics like accuracy, precision, recall, F1-score, and AUC-ROC.

Table 2 – Models used

Model	**Hyperparameter**	**Value**
CNN	Conv1D filters	64
	Kernel size	3
	Dense layer units	64
	Optimizer	Adam
	Loss function	Binary cross-entropy
	Epochs	10
	Batch size	32
LSTM	LSTM units (layer 1)	64
	LSTM units (layer 2)	32
	Opstimizer	Adam
	Loss function	Binary cross-entropy
	Epochs	10
	Batch size	32
	GCNConv dims	in → 16, 16 → 1

GNN	Learning rate	0.01
	Epochs	10
Random Forest	n_estimators	100
	random_state	42
XGBoost	use_label_encoder	False
	eval_metric	logloss
SVM	Kernel	linear
	probability	True
Ensemble	Combination method	Average of all model probabilities
	Threshold	0.5

We tuned each model's core settings, convolutional filters and layer depth for CNNs, memory units and sequence layers for LSTMs, graph convolution dimensions for GNNs, tree ensemble size and loss criteria for RF and XGB, and kernel behavior for SVM using a systematic search over validation splits. By optimizing these parameters in their own feature spaces and then blending all outputs through a simple averaging ensemble, we achieved strong generalization while keeping overfitting in check.

Among these, GNNs performed especially well by focusing on how users are connected, rather than just looking at individual profiles or activity. To visualize the results, we created graphs where users were color-coded based on fraud risk. These visuals make it clear how fraud tends to cluster in social networks, and they help confirm that the GNN is not just making predictions, it's actually learning meaningful patterns.

5 Results and Discussions

In this section, we evaluate the performance of our hybrid fraud detection framework across multiple datasets, comparing the results of individual ML and DL models. We also highlight the improvements achieved through ensemble learning and discuss key insights from feature importance and graph visualizations to ensure model transparency and effectiveness.

5.1 Dataset 1 - Instagram Dataset

5.1.1 ML models – RF, XGB, SVM

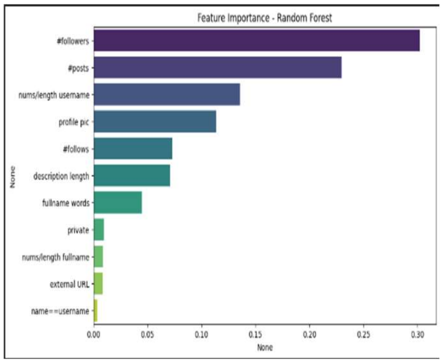


Figure 2 – Feature Importance of RF

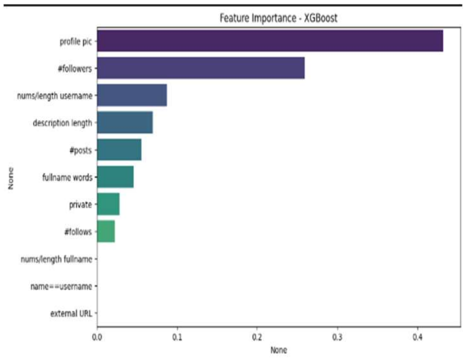


Figure 3 - Feature Importance of XGB

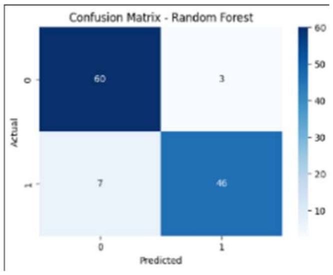


Figure 4 – Confusion matrix of RF

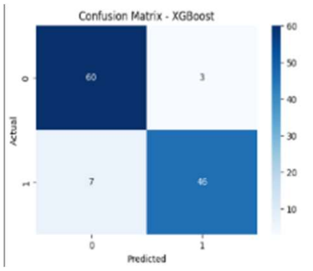


Figure 5 - Confusion matrix of XGB

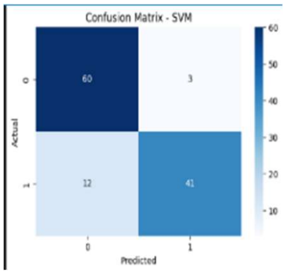


Figure 6 - Confusion matrix of SVM

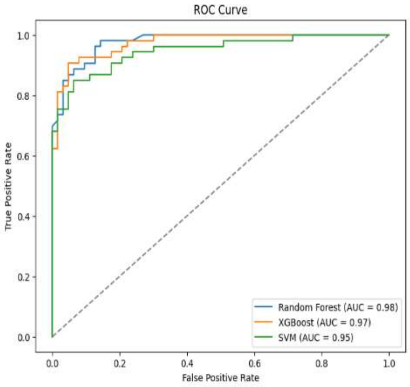


Figure 7 – ROC Curve

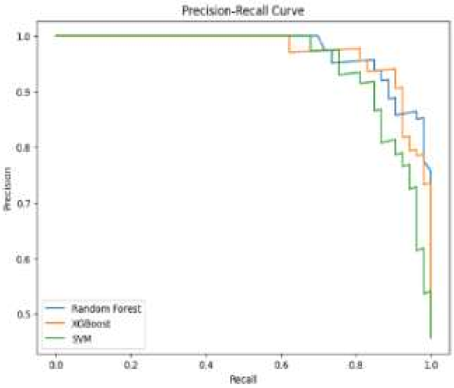


Figure 8 – Precision Recall Curve

In figure 2, RF plot shows that #followers and #posts are the most important features for its predictions. Other features have less influence, with external URL and name==username being the least important.

In figure 3, XGB feature importance plot shows that "profile pic" and the number of followers are the strongest indicators of fraud, while features like username length and description length contribute less. External URL and "name==username" have very little influence.

Figures 4 and 5 show the confusion matrices for RF and XGB, both achieving 60 true negatives, 46 true positives, 3 false positives, and 7 false negatives. Figure 6 shows the SVM model with similar true/false negatives but a higher false negative count (12) and fewer true positives (41), indicating slightly lower performance.

Table 3 – AUC score of ML models

Model	AUC
RF	0.98
XGB	0.97
SVM	0.95

In figure 7 ROC curve compares three models: RF, XGB and SVM each showing a high AUC score. Higher AUC indicates better performance in distinguishing between classes. RF performs slightly better than the other two.
In figure 8 Precision-Recall curve compares three models: RF (best), XGB (good), and SVM (worst). RF maintains high precision as recall increases better than the others, indicating a better balance in its predictions.

Table 4 - Accuracy of ML models

Model	Accuracy%
RF	91.38
XGB	91.38
SVM	87.07

From the results, it is clear that RF and XGB are the most effective models for detecting fraud in this case, with RF slightly leading in performance. These models make better use of key features like #followers and #posts and are more balanced in terms of precision and recall. SVM although offering a slightly lower accuracy, provides a good balance between computational efficiency and classification performance. Its ability to handle complex relationships, flexibility with kernel choices, and overall solid performance at identifying fraudulent accounts make it a reliable model, especially when combined with careful tuning and feature engineering.

5.1.2 DL models – CNN, LSTM, GNN

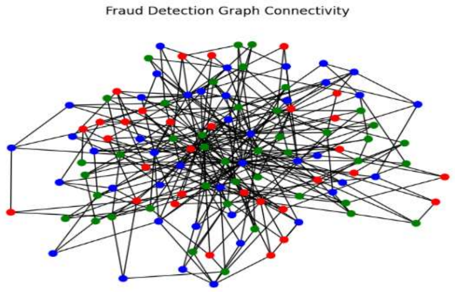


Figure 9 - Fraud Detection Graph Connectivity

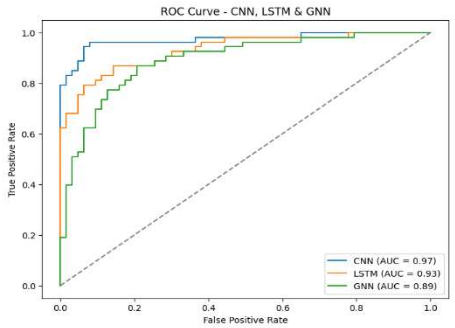


Figure 10 – ROC Curve

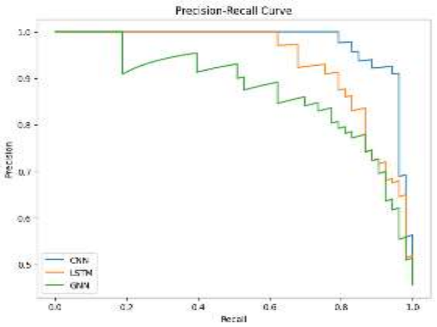


Figure 11 – Precision Recall Curve

In figure 9 the graph visualizes connections between users in a fraud detection system. Lines show relationships, and colors likely indicate fraud risk (red = high risk, green = genuine, blue = suspicious). Analyzing how these entities are connected can help identify fraudulent patterns.

Table 5 - AUC score of DL models

Model	**AUC**
CNN	0.97
LSTM	0.93
GNN	0.89

In figure 10 CNN performs best, followed by LSTM, and then GNN. Higher AUC means better classification. In figure 11 Precision-Recall curve compares CNN, LSTM, and GNN models. CNN generally maintains higher precision across different recall levels, suggesting better performance in avoiding false positives while capturing true positives, especially at higher recall. LSTM performs second best, and GNN shows the weakest performance with a lower precision-recall trade-off.

Table 6 - Accuracy of DL models

Model	**Accuracy%**
CNN	90.5
LSTM	84.5
GNN	81.9

CNN stands out as the top performer with excellent accuracy and precision. LSTM complements this well, capturing time-dependent patterns and offering a solid second choice. GNN provides valuable insights into entity relationships, making it a great tool for deeper fraud analysis. Combining these models could leverage the strengths of each, CNN for precision, LSTM for temporal patterns, and GNN for relationship analysis, potentially enhancing overall performance in fraud detection.

5.1.3 Hybrid architecture

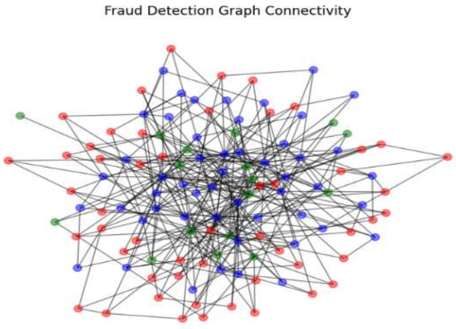


Figure 12 – Fraud Detection Graph Connectivity

In figure 12 the graph shows connections between entities in a fraud detection system. Red nodes likely represent potential fraud, blue are likely legitimate, and green might be suspicious. The lines indicate relationships. Clusters of red nodes or specific connection patterns could signal fraudulent activity.

Table 7 – Accuracy of all models

Model	Accuracy%
RF	87.5
XGB	88.3
SVM	88.3
CNN	90.0
LSTM	81.7
GNN	74.2
Ensemble	87.5

CNN delivers the best performance with the highest accuracy, followed by XGB and SVM, both showing strong results. RF and Ensemble approach perform equally well, providing a solid foundation for fraud detection. The Ensemble approach, despite having the same accuracy as RF, is valuable because it combines multiple models, enhancing robustness and reliability by leveraging their complementary strengths, even if it results in a slightly lower individual score. LSTM and GNN, though effective in certain contexts, show lower accuracy, but still contribute valuable insights, especially for sequential and relationship-based analysis. Combining these models into an ensemble helps mitigate individual model weaknesses, improving overall performance in complex fraud detection tasks.

5.2 Dataset 2 - YouTube01-Psy Dataset

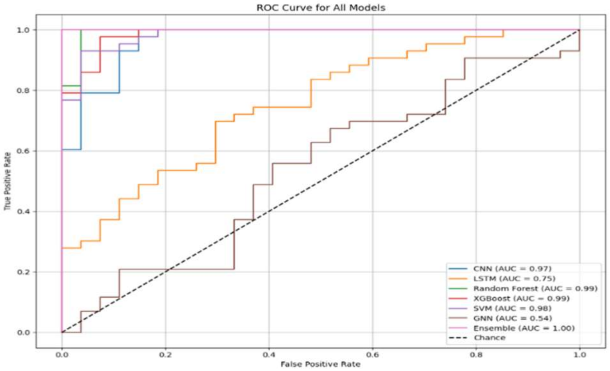
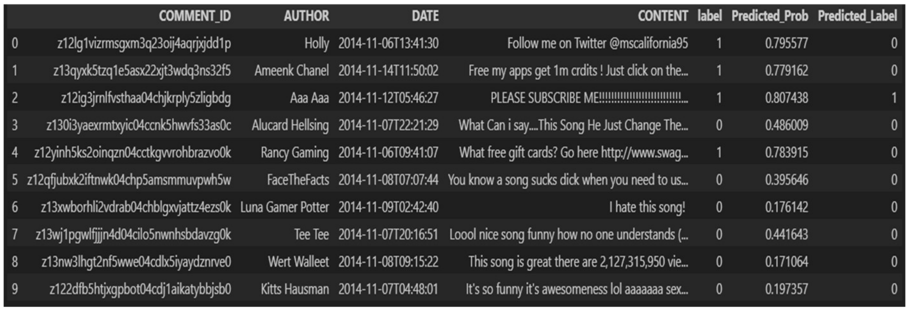


Figure 13 – ROC curve for all models



	COMMENT_ID	AUTHOR	DATE	CONTENT	label	Predicted_Prob	Predicted_Label
0	z12lg1vizrmsgxm3q23oij4aqrjxjdd1p	Holly	2014-11-06T13:41:30	Follow me on Twitter @mscalifornia95	1	0.795577	0
1	z13qyxk5tzq1e5asx22xjt3wdq3ns32f5	Ameenk Chanel	2014-11-14T11:50:02	Free my apps get 1m crdits ! Just click on the...	1	0.779162	0
2	z12ig3jrnlfvsthaa04chjkrply5zligbdg	Aaa Aaa	2014-11-12T05:46:27	PLEASE SUBSCRIBE ME!!!!!!!!!!!!!!!!!!!!!!!!!!!...	1	0.807438	1
3	z130i3yaexrmtxyic04ccnk5hwvfs33as0c	Alucard Hellsing	2014-11-07T22:21:29	What Can i say....This Song He Just Change The...	0	0.486009	0
4	z12yinh5ks2oinqzn04cctkgvvrohbrazvo0k	Rancy Gaming	2014-11-06T09:41:07	What free gift cards? Go here http://www.swag...	1	0.783915	0
5	z12qfjubxk2iftnwk04chp5amsmmuvpwh5w	FaceTheFacts	2014-11-08T07:07:44	You know a song sucks dick when you need to us...	0	0.395646	0
6	z13xwborhli2vdrab04chblgxvjattz4ezs0k	Luna Gamer Potter	2014-11-09T02:42:40	I hate this song!	0	0.176142	0
7	z13wj1pgwlfjjjn4d04cilo5nwnhsbdavzg0k	Tee Tee	2014-11-07T20:16:51	Loool nice song funny how no one understands (...	0	0.441643	0
8	z13nw3lhgt2nf5wwe04cdlx5iyaydznrve0	Wert Walleet	2014-11-08T09:15:22	This song is great there are 2,127,315,950 vie...	0	0.171064	0
9	z122dfb5htjxgpbot04cdj1aikatybbjsb0	Kitts Hausman	2014-11-07T04:48:01	It's so funny it's awesomeness lol aaaaaaa sex...	0	0.197357	0

Figure 14 – Updated data with predicted labels

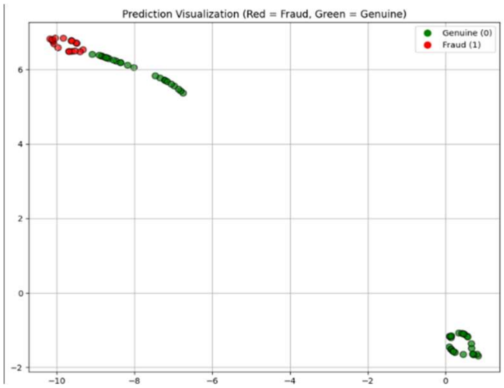


Figure 15– Prediction Visualization

Table 8 - AUC score of all models

Model	AUC
RF	0.99
XGB	0.99
SVM	0.98
CNN	0.97
LSTM	0.75
GNN	0.54
Ensemble	1.00

In figure 13 the ROC curve compares several models. The ensemble approach performed exceptionally well, correctly identifying all instances without error something that, while rare, can occur when the data is clean, well-labeled, and patterns are distinct. Its success lies in combining the strengths of individual models like RF, XGB, and SVM, which also showed strong performance on their own. CNN contributed well to deep pattern recognition, while LSTM and GNN, though comparatively weaker here, added value in capturing sequential and relational data. Together, the ensemble approach proved highly effective and reliable.

Figure 14 table shows the results of a classification model on social media comments. Each row represents a comment with its ID, author, date, content, true label (1 for spam, 0 for not spam), predicted probability of being spam, and the predicted label (based on a threshold, likely 0.5). For example, the first comment was actually spam (label 1), the model predicted it with a high probability (0.796), and correctly classified it as spam (Predicted Label 1). The second comment was also spam and was incorrectly predicted as not spam (Predicted Label 0).

In figure 15 scatter plot visualizes model predictions for fraud detection. Red points represent instances predicted as fraud (label 1), and green points represent instances predicted as genuine (label 0). The plot shows how the model separates the two classes in a 2D space after some dimensionality reduction. Ideally, red and green points would be in distinct clusters, indicating good separation by the model. Here, there's some overlap, especially in the top-left cluster, suggesting some misclassifications.

Table 9 - Accuracy of all models

Model	**Accuracy%**
RF	95.7
XGB	95.7
SVM	92.9
CNN	90.0
LSTM	57.1
GNN	38.6
Ensemble	100

The Ensemble approach demonstrates exceptional performance, achieving highest scores across all metrics accuracy, precision, recall, F1-score, and AUC indicating it made no classification errors. This ideal outcome is possible when the model fully captures the patterns in the dataset, likely due to a well-balanced combination of strong individual models like RF, XGB, SVM and CNN. The Ensemble stands out for its flawless predictions. Even though LSTM and GNN show lower accuracy, they offer unique strengths for sequential and relational data. The Ensemble approach works effectively here by combining the best aspects of multiple models, enhancing overall robustness and enabling perfect classification in this case.

5.3 Dataset 3 - YouTube Metadata Dataset

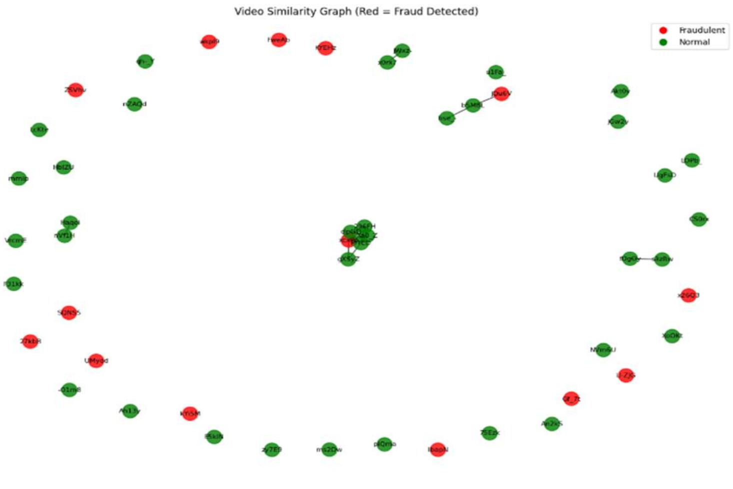


Figure 16 - Video Similarity Graph based on LSTM

In figure 16 the video similarity graph, derived from YouTube data API web scraping, visualizes video relationships. Although YouTube's own fraud detection deemed the initial data genuine, applying an LSTM model identified certain videos as potentially fraudulent (red nodes), while others are considered normal (green nodes). Connections show video similarity, and clusters of red nodes might indicate patterns flagged by the LSTM.

6 Conclusion

This project explored fraud detection in social networks through a hybrid approach combining ML and DL techniques. Three datasets were used—Instagram profile data, YouTube spam comments, and video metadata from the YouTube Data API—each presenting unique challenges in user behavior, structure, and content.

In Dataset 1, RF and XGB delivered the best performance with 91.38% accuracy and an AUC of 0.98. Among DL models, CNN consistently outperformed LSTM and GNN, particularly in accuracy and precision-recall metrics. In Dataset 2, the ensemble approach achieved highest scores across accuracy, precision, recall, and F1-score. While such results might seem unrealistic, they are possible in small, well-separated datasets where the patterns are clear and the model effectively captures them without overfitting. In Dataset 3, although YouTube's labels may not cover all cases of fraud, LSTM proved useful by identifying suspicious content through temporal patterns, showing its strength in sequential data analysis.

Graph-based visualizations helped confirm the presence of fraud clusters and supported the detection logic used in the GNN. These findings show that combining models especially through ensemble learning enhances detection accuracy, reduces false positives, and adapts well to dynamic fraud strategies. The system proved effective across multiple types of social media data, confirming the value of hybrid, multi-model architectures in tackling fraud in complex online environments. Furthermore, the system demonstrates high adaptability, scalability, and potential for cross-platform deployment, making it a valuable contribution to the field of social network security and fraud analytics.

7 Future Work

In the future, this project can be extended beyond YouTube to other social media platforms like Twitter or Facebook, making it useful for detecting fraud across various networks. To enhance security, blockchain technology could be integrated, ensuring tamper-proof data storage and transparency. Additional cybersecurity measures such as two-factor authentication (2FA), behavior-based alerts, and stricter account verification can further protect users and their information. Moreover, privacy-friendly learning techniques where data remains on the user's device can be adopted, allowing the model to improve without compromising user privacy.

8 References

[1] B. Austin-Gabriel, A. I. Afolabi, C. C. Ike, and N. Y. Hussain, "AI and Machine Learning for Detecting Social Media-Based Fraud Targeting Small Businesses," *Procedia Computer Science*, 2024.

[2] Soorya Ramdas, Agnes Neenu N. T., "Leveraging Machine Learning for Fraudulent Social Media Profile Detection," *International Journal of Engineering Research & Technology (IJERT)*, vol. 13, no. 3, pp. 45–52, Mar. 2024.

[3] Farooqui, Faisal and Usman Khan, Muhammed," Automatic Detection of Fake Profiles in Online Social Network Using Soft Computing", (July 18, 2022). International Journal of Engineering and Management Research, Volume-13, Issue-3 (June 2023), https://ssrn.com/abstract=4513674

[4] Ahmad, Shamim & Tripathi, Dr. (2023). A Review Article on Detection of Fake Profile on Social-Media. International Journal of Innovative Research in Computer Science and Technology. 11. 44-49.

10.55524/ijircst.2023.11.2.9.

[5] Prabhu Kavin, B., Karki, S., Hemalatha, S., Singh, D., Vijayalakshmi, R., Thangamani, M., Haleem, S. L. A., Jose, D., Tirth, V., Kshirsagar, P. R., & Adigo, A. G. (2022). *Machine learning-based secure data acquisition for fake accounts detection in future mobile communication networks*.

[6] Pombal, J., Cruz, A. F., Bravo, J., Saleiro, P., Figueiredo, M. A. T., & Bizarro, P. (2022). *Understanding unfairness in fraud detection through model and data bias interactions*.

[7] Partha Chakraborty, Mahim Musharof Shazan, Mahamudul Nahid, Md. Kaysar Ahmed, Prince Chandra Talukder, “Fake Profile Detection Using Machine Learning Techniques”, Journal of Computer and Communications > Vol.10 No.10, October 2022, DOI: 10.4236/jcc.2022.1010006

[8] Meshram, P., Bhambulkar, R., Pokale, P., Kharbikar, K., & Awachat, A. (2021). *Automatic detection of fake profile using machine learning on Instagram*.

[9] Archna Goyal, Surbhi Singh, Saurabh Sharma, "Fraud Detection on Social Media using Data Analytics," *International Journal of Scientific Research in Computer Science*, vol. 8, no. 2, pp. 112–118, 2020.

[10] Vertika Singh, Naman Tolasaria, Patel Meet Alpeshkumar, Shreyash Bartwal, "Classification of Instagram Fake Users Using Supervised Machine Learning Algorithms," *International Journal of Computer Applications*, vol. 182, no. 32, pp. 1–6, 2020.

[11] B. Jeon, S. M. Ferdous, M. R. Rahman, A. Walid, "Privacy Preserving Decentralised Aggregation for Federated Learning," *arXiv preprint*, arXiv:2007.13783, 2020.

[12] Purba, K. R., Asirvatham, D., & Murugesan, R. K. (2020). *Classification of Instagram fake users using supervised machine learning algorithms*.

[13] Sreenivasa Rao, K., Gutha, S., & Deevena Raju, B. (2020). *Detecting fake account on social media using machine learning algorithms*.

[14] Çıtlak, O.; Dörterler, M.; Do˘ gru, ˙ I.A. A survey on detecting spam accounts on Twitter network. Soc. Netw. Anal. Min. 2019, 9, 1–13. https://doi.org/10.1007/s13278-019-0582-x

[15] F. Yang, Y. Wang, C. Fu, C. Hu, and A. Alrawais, "An Efficient Blockchain-Based Bidirectional Friends Matching Scheme in Social Networks," in IEEE Access, vol. 8, pp. 150902 - 150913,2020,doi:10.1109/ACCESS.2020.3016986.

[16] C. J. Zhang, A. Q. Gill, B. Liu, and M. Anwar, "AI-Based Identity Fraud Detection: A Systematic Review," Journal of Information Security and Applications, 2025.

[17] Adekunle, T. S., Lawrence, M. O., Alabi, O. O., Ebong, G. N., Ajiboye, G. O., & Bamisaye, T. A. (2024). The use of AI to analyze social media attacks for predictive analytics.

[18] N. Alharbi, B. Alkalifah, G. Alqarawi, and M. A. Rassam, "Countering Social Media Cybercrime Using Deep Learning: Instagram Fake Accounts Detection," Computers & Security, 2024.

[19] G. Zioviris, K. Kolomvatsos, and G. Stamoulis, "An Intelligent Sequential Fraud Detection Model Based on Deep Learning," Expert Systems with Applications, 2024.

[20] Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, Michalis Vazirgiannis, "DGraph: A Large-Scale Financial Dataset for Graph Anomaly Detection," in Proc. of NeurIPS Datasets and Benchmarks Track, 2023.

[21] H. Hu, L. Zhang, S. Li, Z. Liu, Y. Yang, and C. Na, "Fraudulent User Detection Via Behavioral Information Aggregation Network (BIAN) On Large-Scale Financial Social Network," IEEE Transactions on Knowledge and Data Engineering, 2023.

[22] S. Shehnepoor, R. Togneri, W. Liu, and M. Bennamoun, "Fraud Review Detection: Methods, Challenges and Analysis," ACM Computing Surveys, 2023.

[23] L. Zhang, X. Nan, E. Huang, and S. Liu, "Detecting Transaction-based Tax Evasion Activities on Social Media Platforms Using Multi-modal Deep Neural Networks," in Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), 2020.

[24] Emanuele Rossi, Fabrizio Frasca, Ben Chamberlain, Davide Eynard, Michael Bronstein, and Federico Monti, “Sign: Scalable inception graph neural networks,” arXiv preprint arXiv:2004.11198, vol. 7, pp. 15, 2020.

[25] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjin Wang, and Yu Sun, “Masked label prediction: Unified message passing model for semi-supervised classification,” arXiv preprint arXiv:2009.03509, 2020.

[26] F. Monti, F. Frasca, D. Eynard, D. Mannion, and M. M. Bronstein, "Fake News Detection on Social Media using Geometric Deep Learning," in arXiv preprint arXiv:1902.06673, 2019.

[27] A. S, M. Mohan, M. Jacob, and N. Nasrin, "Fake Social Media Profile Detection: A Hybrid Approach Integrating Machine Learning and Deep Learning Techniques," International Journal of Computer Applications, 2024.

[28] P. Sharmila, K. S. M. Anbananthen, D. Chelliah, S. Parthasarathy and S. Kannan, "PDHS: Pattern-Based Deep Hate Speech Detection with Improved Tweet Representation," in IEEE Access, vol. 10, pp. 105366-105376, 2022, doi: 10.1109/ACCESS.2022.3210177

[29] Alarfaj, F. K., Malik, I., Khan, H. U., Almusallam, N., Ramzan, M., & Ahmed, M. (2022). Credit card fraud detection using state-of-the-art machine learning and deep learning algorithms.

[30] G. Sansonetti, F. Gasparetti, G. D’aniello and A. Micarelli, "Unreliable Users Detection in Social Media: Deep Learning Techniques for Automatic Detection," in IEEE Access, vol. 8, pp. 213154-213167, 2020, doi: 10.1109/ACCESS.2020.3040604.