RESEARCH ARTICLE OPEN ACCESS

Quantum-Inspired Machine Learning Framework for Social Media Account Authentication

Namana S
Dept of ISE
Don Bosco Institute of Technology
Bengaluru, India
namana608kumar@gmail.com

Monisha B S
Dept of ISE
Don Bosco Institute of Technology
Bengaluru, India
monishabs83@gmail.com

Rachana T M
Dept of ISE
Don Bosco Institute of Technology
Bengaluru, India
rachanatm17@gmail.com

Vanishree B S
Dept of ISE
Don Bosco Institute of Technology
Bengaluru, India
vanishreebs105@gmail.com

Abstract:

Social media platforms have seen a sharp rise in fake, automated, and impersonated accounts, making it harder to preserve trust, control misinformation, and maintain platform safety. This paper proposes a quantum-inspired machine learning framework for evaluating whether a social media account is authentic by learning useful patterns from commonly available profile, content, and interaction features. These include signals such as account age, posting frequency, timing regularity, follower-following relationships, engagement behaviour, writing-style indicators, and network connectivity. To study performance, multiple classifiers including logistic regression, decision tree, random forest, and gradient boosting are trained and tested using stratified train-test splits. The models are compared using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, with particular focus on minimizing false positives for genuine users and false negatives for abusive accounts. The best model is then deployed through a user-friendly web application that provides probability-based authenticity predictions along with confidence scores and interpretable explanations supported by SHAP visualizations. These explanations help moderators and analysts understand which features influenced each prediction, improving transparency and supporting better decision-making. Overall, the work shows that combining quantum-inspired learning with interpretable machine learning can strengthen fake-account detection, improve platform reliability, and support scalable authenticity assessment in modern social media systems.

I. INTRODUCTION

Social media platforms now play a major role in news sharing, public conversation, marketing, and community interaction. At the same time, their openness and large scale make them easy targets for bots, fake profiles, spam networks, and impersonation attempts. These inauthentic accounts can influence trending topics, spread misleading information, support scams, and gradually weaken user confidence in the platform. Because large platforms generate enormous volumes of activity, manual moderation by itself is no longer enough, which makes automated authenticity assessment an important requirement.

Machine learning offers a practical solution because it can identify authenticity-related patterns from observable signals. These signals may come from profile details such as account age or biography completeness, behavioural traces such as posting rate and timing consistency, content-based clues such as repetition or link density, and network properties such as follower-following balance or community connectivity. After suitable preprocessing, including handling missing values, normalizing numeric data, encoding categorical variables, and addressing class imbalance, models such as logistic regression, decision trees, random forests, and gradient-boosting methods can learn both simple and complex relationships that help distinguish genuine users from automated or malicious accounts.

Although classical machine learning already delivers strong results, social media data is usually high-

dimensional, noisy, and highly correlated across behaviour, content, and network activity. This creates an opportunity to explore quantum-inspired machine learning, which borrows ideas such as high-dimensional feature mapping, angle or amplitude style encodings, and quantum-inspired optimization while still running on classical hardware. In this work, the proposed framework combines feature engineering from profile, behavioural, content, and network signals with quantum-inspired transformations to improve class separability, and it produces interpretable outputs using probability scores and explanation methods such as SHAP. The overall design is intended for scalable moderation support, helping reviewers focus on suspicious accounts, lower false alarms on legitimate users, and respond more effectively to emerging abuse patterns.
That brings us to the solution in this research GREEN LOOP. It's a digital, cloud-based platform built to power smart waste management and promote serious recycling. Everything's connected: barcode scanning, AI-powered image verification, cloud data management, and reward systems. When people want to recycle, they just scan a barcode, upload a photo to prove they recycled something, and rack up points once their effort gets verified. Those points

Typical indicators of inauthentic accounts include:

- Very high posting frequency or highly regular, automated timing patterns
- Unusual follower-following ratios, low reciprocity, or suspiciously dense follow bursts
- Repetitive content, high URL/hashtag density, or near-duplicate posts across many accounts
- Recently created accounts with incomplete profiles or inconsistent identity signals
- Coordinated engagement patterns, such as synchronized likes or retweets, within tight clusters

II. RELATED WORK

Varol et al. [1] presented a well-known machine-learning approach for detecting social bots by using profile metadata, content, network behaviour, sentiment, and temporal activity. Their work showed that supervised classifiers can separate bot accounts from human users, but also noted that changing bot behaviour makes cross-dataset generalization difficult.

Cresci et al. [2] demonstrated that modern social spambots can imitate human-like profiles and interactions, making simple rule-based detection less effective. Their findings highlighted the need to study coordinated behaviour such as synchronized posting, dense retweet activity, and group-level regularities.

Feng et al. [3] introduced TwiBot-20, a large-scale benchmark that combines profile attributes, tweet semantics, and neighborhood relations. Their study showed that older bot-detection models often perform poorly on realistic datasets, emphasizing the importance of strong feature engineering and non-linear learning models.

Ferrara et al. [4] reviewed social-bot detection techniques and grouped authenticity signals into profile-based, content-based, behavioural, temporal, and network features. Their survey also identified major challenges such as label noise, dataset shift, adversarial adaptation, and scalable feature extraction.

Wu, Rim, and Lee [5] proposed a multimodal authenticity detection approach that combines user metadata, post content, and network interaction cues. Such combined representations improve robustness against impersonation and coordinated campaigns, while explanation methods such as SHAP help moderators understand why an account is flagged.

Davis et al. [6] developed BotOrNot, later known as Botometer, which assigns bot-likeness scores using a supervised-learning pipeline built from metadata, content, and temporal features. This work helped establish practical scoring-based evaluation for real-world moderation systems.

Graph-based learning methods have also influenced fake-account detection. ChebNet [7] and GraphSAGE [8] showed how graph convolution and neighborhood aggregation can model irregular network structures, which is useful for detecting suspicious follower-following patterns and coordinated clusters in social media graphs.

From the quantum-learning perspective, Havlicek et al. [9] demonstrated that quantum feature maps can create richer feature spaces for supervised classification. This idea motivates quantum-inspired feature transformations on classical hardware, where

high-dimensional mappings and kernel-style similarity measures may improve separation between genuine and inauthentic accounts.

Based on these studies, the proposed work combines classical supervised models with quantum-inspired feature transformation and probability-based authenticity scoring. The aim is to improve detection performance on noisy and high-dimensional social-media data while maintaining interpretability for moderation- oriented decision support.

Yang et al. [10] extended this direction through TwiBot-22, a larger benchmark designed to represent more realistic Twitter ecosystems with richer user relationships and graph structures. Their work shows that reliable account authentication should not depend only on isolated profile attributes, but should also analyze how accounts connect, interact, and evolve within the wider social network.

Lee, Eoff, and Caverlee [11] studied social honeypots and content polluters, showing that malicious accounts often reveal repeated behavioural patterns over time. Their findings support the use of temporal features such as posting bursts, abnormal following activity, and repeated promotional content for identifying suspicious or automated users.

Beskow and Carley [12] proposed an ensemble-based bot detection approach that combines several classifiers and feature groups to improve robustness. This idea is useful for social media authentication because fake accounts may expose different signals through their profile, content, behaviour, or network activity, and a combined model can capture these variations more effectively.

Overall, these additional studies confirm that fake-account detection is most effective when multiple evidence sources are combined. Therefore, the proposed framework uses profile, behavioural, content, and network-based features together with quantum-inspired transformations to improve classification while keeping the final prediction interpretable for moderation use.

Recent graph neural network studies further improve bot detection by learning from both account metadata and relationship structures. Neural architecture search based models [13] automatically identify suitable graph model designs, reducing manual model selection and improving performance on large bot-detection benchmarks.

Community-aware graph contrastive learning methods [14] also show that social media accounts should be evaluated in relation to their surrounding communities. By comparing similar and dissimilar users within a heterogeneous graph, such methods can capture hidden coordination patterns that may not bevisible from profile features alone.

Federated graph-based bot detection [15] addresses privacy and data-sharing limitations by training models across distributed datasets without directly exposing user data. This direction is relevant for real-world deployment because platforms may need collaborative detection while preserving user privacy and organizational data boundaries.

Recent explainable fake-profile detection studies [16] emphasize that high accuracy alone is not sufficient for account authentication. Moderation systems must also explain the contribution of important features such as account age, posting frequency, follower ratio, content similarity, and engagement behaviour so that decisions can be reviewed with greater confidence.

III. PROPOSED SYSTEM

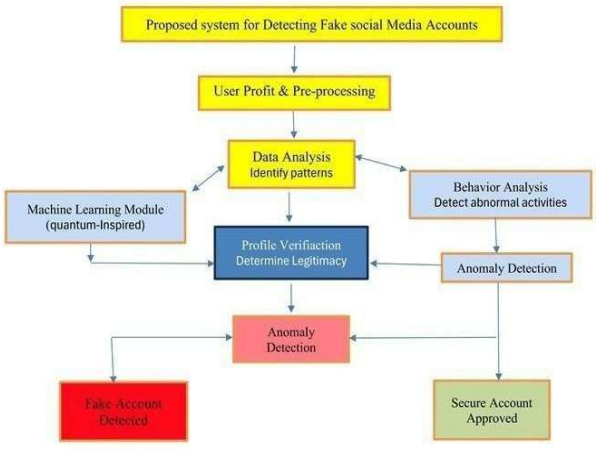

Fig. 3.1 Proposed Quantum-Inspired Machine Learning Framework for Social Media Account Authentication

Figure 3.1 outlines the proposed quantum-inspired machine learning framework for assessing the authenticity of social media accounts. The process begins with data sources, where account-level data

is gathered from social media APIs or trusted public datasets. This data may include profile metadata such as account age and bio completeness, content samples like posts, hashtags, and URLs, temporal behaviour such as posting rate and inter-post intervals, and network or interaction signals including followers-following structure, mentions, replies, and reshares. The next stage, data preparation, organizes these records, removes duplicates, and
assigns labels such as genuine or inauthentic using verified indicators or expert annotation.

During the pre-processing stage, missing values are, numerical attributes are normalized or scaled, categorical fields are encoded, and textual content is transformed into measurable representations. Techniques such as class weighting or SMOTE are also used to reduce bias when genuine and inauthentic accounts are not evenly distributed. After this, the feature extraction module builds a unified feature vector that captures profile, content, temporal, and graph-based patterns. To better represent non-linear relationships among these different signals, the framework applies a quantum-inspired feature mapping on classical hardware. The transformed representation is then passed to the ML model training stage, where classifiers such as logistic regression, decision tree, random forest, and gradient boosting are trained and compared.

Performance evaluation uses metrics such as accuracy, precision, recall, F1-score, and ROC-AUC, with emphasis on reducing false positives (flagging genuine users) and false negatives (missing malicious accounts). Finally, the selected model is integrated with an explainability layer (feature importance/SHAP) and deployed through a web UI/API to output an authenticity label and probability score, while a monitoring component tracks drift and feedback to trigger periodic retraining as adversarial behaviours evolve, ensuring reliable and scalable authenticity screening.

IV. FLOW DIAGRAM

- Figure 4.1 outlines the end-to-end stages of the proposed account authenticity assessment system. **Users** of the system include platform moderators, analysts, and investigators. They start an authenticity check by providing an account identifier, such as a username or profile URL, or by selecting accounts from a flagged queue. This stage defines the operational context, where the system must return fast, probability-based decisions while keeping false positives low so legitimate users are not affected unnecessarily.

  Web UI: The web user interface provides a simple workflow to submit an account for analysis and view results. It displays the predicted label (genuine/inauthentic), an authenticity probability score, and an explanation panel (top contributing features) so that moderators can quickly understand why an account was flagged and decide on the next action.

- **API:** The Application Programming Interface (API) links the UI and downstream analytics. It receives the account identifier and requested scope (profile-only, content + network, etc.), fetches the required signals through approved connectors (APIs/datasets), invokes preprocessing and model inference, and returns a structured response containing the authenticity score, label, and explanation details.

- **Pre-processing:** This stage cleans and prepares raw social-media signals for machine learning by handling missing values, removing duplicates, normalizing/scaling numerical features (e.g., posting rates), encoding categorical fields, and converting text content into measurable representations. Class-imbalance handling is applied to reduce bias because inauthentic accounts are often rarer than genuine ones.

- **Database:** The database stores extracted features (where permitted), prediction outputs (scores/labels), explanation summaries, and review outcomes from moderators. This history supports auditing, performance analysis, and supervised feedback loops for improving future model versions.

- **External Data:** External data can include

curated bot/fake-account benchmark datasets,threat-intelligence feeds, blocklists of malicious URLs/domains, and cross-platform signals (when ethically and legally allowed). These sources enrich feature extraction, improve coverage of emerging attack patterns, and help validate generalization across datasets.

- **Security:** Security ensures that account data and moderation decisions are protected from unauthorized access. It includes authentication/authorization, secure storage of sensitive fields, audit logging and request validation are also used to prevent abuse of the scoring API.

- **Monitoring:** Monitoring continuously tracks data drift and model performance (precision/recall, false-positive rate, latency) as attacker behaviour evolves. It flags abnormal shifts in feature distributions, supports periodic retraining using newly reviewed cases, and ensures the deployed model remains reliable for large-scale authenticity screening.

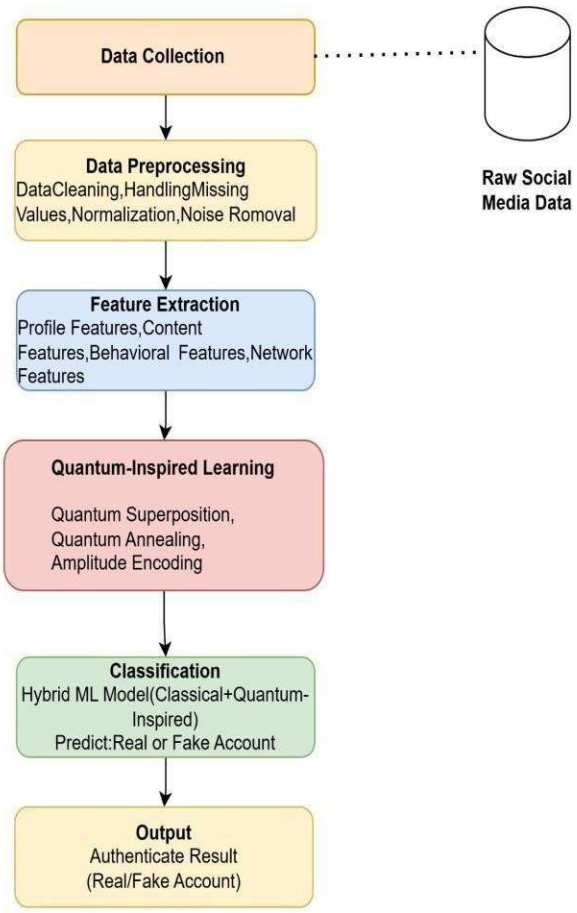

Fig. 4.1 Proposed Social Media Account Authentication System Flow

V. Comparative Study of Social Media Account Authenticity Assessment Using ML Classifier

Existing System

Different machine learning approaches have been explored to identify whether a social media account is genuine or fake. Traditional models such as Support Vector Machines (SVM) and Random Forests (RF) are often considered strong starting points because they are reliable, easier to implement, and work well with structured data. These models can produce good results when the selected features clearly capture account behaviour. However, their performance can drop when the data is noisy, highly imbalanced, or constantly changing, which is common in real-world social media platforms.

To address these challenges, more advanced methods such as ensemble learning and deep learning have also been applied. These approaches are generally better at identifying complex and hidden patterns in the data, which can improve detection accuracy. At the same time, they often need larger datasets, longer training time, and

greater computational resources, which may not always be practical in every setting.

In this work, quantum-inspired machine learning is treated as an enhancement to classical methods rather than a replacement for them. It helps represent input features in a richer way and can improve the separation between genuine and suspicious accounts, especially when the data is high-

dimensional and contains subtle patterns. Overall, the comparison suggests that each classifier has its own strengths, but a hybrid framework that combines classical machine learning with quantum-inspired techniques offers a more balanced solution by improving accuracy, scalability, and robustness for social media account authenticity assessment.

Proposed System

The proposed system uses a quantum-inspired machine learning framework to improve social media account authenticity assessment. Instead of depending only on basic profile or content features, it combines profile details, behavioural patterns, content characteristics, and network-based signals to form a stronger feature set for classification.

Uses quantum-inspired feature transformation to represent complex, high-dimensional social media data more effectively on classical hardware.

- Compares multiple classifiers such as logistic regression, decision tree, random forest, and gradient boosting to identify the best-performing model.
- Generates probability-based authenticity scores instead of only giving a fixed genuine or fake label.
- Includes explainability support using feature importance and SHAP-style explanations so moderators can understand why an account is flagged.

Overall, the proposed system aims to provide a more accurate, scalable, and interpretable solution for detecting fake, automated, or suspicious social media accounts while reducing false alarms on genuine users

TABLE I

Paper Title	Year	Techniques Used	Platform /Dataset	Accuracy /Performance	Advantages	Limitations
Quantum-Inspired Hybrid Machine Learning Framework for Scalable Computational Next-Generation Technologies	2026	Quantum-Inspired Hybrid ML, Quantum-Classical Models	Big Data & AI Systems	High scalability and computational efficiency	Faster processing and improved optimization	Not specifically for fake account detection
PegasosQSVM: A Quantum Machine Learning Approach for Accurate Fake News Detection	2025	Pegasos Quantum SVM, Quantum Kernels	BUZZFEED Dataset	95.63% Accuracy	Better than classical ML methods	Quantum hardware noise issues
Quantum Machine Learning Algorithms for Anomaly Detection: A Review	2025	Quantum Neural Networks, Quantum SVM	Cybersecurity & Anomaly Detection Data	Improved anomaly detection capabilities	Comprehensive review of QML methods	Mostly theoretical implementation
Detecting Fake Accounts on Instagram Using Machine Learning and Hybrid Optimization Algorithms	2024	BGWO-PSO, ANN, SVM, KNN	Instagram Dataset	Improved fake account detection accuracy	Better feature optimization	Complex preprocessing overhead
Detection of Fake Instagram Accounts via Machine Learning Techniques	2024	Random Forest, Logistic Regression, SVM, KNN	Public Instagram Data	High accuracy with fewer features	Effective small dataset handling	Limited public data availability
Social Media Fake Account Identification Using Machine Learning	2023	SVM, KNN, Random Forest, ANN	Facebook, Twitter, Instagram	High fake account identification accuracy	Multi-step fake profile analysis	Depends heavily on feature engineering

VI. METHODOLOGY

This task is treated as a probabilistic binary classification problem, where the aim is to estimate how likely a social media account is to be inauthentic, such as a bot, fake profile, spam account, or impersonator, using observable account signals. Instead of depending on a fixed decision threshold, the model produces calibrated probabilities so moderation teams can choose thresholds that fit their policies and risk tolerance. The methodology focuses on signals available before enforcement action is taken, including profile metadata, posting behaviour, content-related indicators, and network or interaction patterns, helping the system adapt better to new campaigns and changing attacker strategies.

A strong methodology begins with data governance and data quality. Account information should be collected only through approved APIs or licensed datasets and managed using privacy-aware practices such as data minimization, access control, and secure storage. Labels used to identify accounts as genuine or inauthentic should come from reliable sources such as enforcement outcomes, trusted annotations, or benchmark datasets, while recognizing that borderline cases may still introduce some label noise. To avoid information leakage, every preprocessing step, including imputation, scaling, encoding, feature selection, and any quantum-inspired mapping parameters, is learned only from the training split and then applied to the validation and test data.

Feature representation is designed to capture multiple sources of evidence without losing interpretability. Numerical features such as posts per day, inter-post variance, follower–following ratio, and engagement rate are normalized for algorithms that are sensitive to scale, including LR, SVM, and KNN, while tree-based models such as DT, RF, and Gradient Boosting are generally more robust. Low-cardinality categorical fields, such as verified status, account type, and language, are one-hot encoded, and missing categories may be retained as an "Unknown" value because missing information can also be meaningful. Content is represented using lightweight text statistics such as URL density, hashtag density, repetition rate, and average length, while network and interaction signals are

summarized using graph statistics such as reciprocity, clustering tendency, and neighbour activity similarity. If quantum-inspired feature

mapping is applied, it is computed after standardization to support stable downstream learning.

The modeling stage compares several classification strategies. Logistic regression provides a simple and interpretable baseline, while SVM and KNN offer alternative decision boundaries for high-dimensional features. Tree-based ensembles such as Random Forest and Gradient Boosting (XGBoost/LightGBM) are included because they often perform well on heterogeneous tabular signals and can learn non-linear interactions between features. The proposed framework also evaluates a quantum-inspired feature mapping stage, which applies a high-dimensional mapping or kernel-style similarity transformation before training to improve the separability of subtle authenticity cues.

Hyperparameters are optimized using stratified k-fold cross-validation, and performance is evaluated mainly with precision, recall, F1-score, and ROC-AUC rather than accuracy alone. When class imbalance is severe, PR-AUC is also considered. The imbalance between genuine and inauthentic accounts is handled through class weights or resampling techniques such as SMOTE, and decision thresholds are adjusted based on moderation costs and operational priorities.

The quantum-inspired transformation stage is applied after preprocessing and before classifier training. In this stage, normalized input features are projected into a richer representation space using angle-style or kernel-style mappings. The purpose is to make subtle differences between genuine and suspicious accounts more visible to the classifier. Since the transformation runs on classical hardware, it remains practical for implementation while still benefiting from ideas inspired by quantum feature spaces.

REFERENCES

1. "Fake Account Detection Using Machine and Deep Techniques in Social Media," IEEE Xplore, 2024, Art. no. 10612549.
2. "Fake Social Media Detection using Machine Learning," IEEE Xplore, 2024, Art. no. 10725280.
3. "Fake Profile Detection Using Machine Learning," IEEE Xplore, 2024, Art. no. 10459570.
4. "Fake Account Detection Using ANN Based Model in Machine Learning," IEEE Xplore, 2024, Art. no. 10561061.
5. "Comparative Analysis of Fake Account Detection Using Machine Learning," IEEE Xplore, 2024, Art. no. 10870733.
6. "Dispelling the Fake: Social Bot Detection Based on Edge Confidence," IEEE Xplore, 2024, Art. no. 10530431.
7. "BotScan: An Unsupervised Bot Detection Based on Adversarial Learning," IEEE Xplore, 2024, Art. no. 10665678.
8. "A Hybrid Deep Learning Architecture for Social Media Bots Detection," IEEE Xplore, 2024, Art. no. 10602502.
9. "Evolution of Malicious Social Bot Detection: From Individual Profiling to Collective Behavior," IEEE Xplore, 2024, Art. no. 11184733.
10. "Fine -Tuned Understanding: Enhancing Social Bot Detection With Large Language Models," IEEE Xplore, 2024, Art. no. 10630818.
11. "Twitter Bot Detection with Multi-Head Attention," IEEE Xplore, 2024, Art. no. 10873626.
12. "Effective Bot Detection in Twitter using Deep Boltzmann Machine," IEEE Xplore, 2024, Art. no. 10533382.

13. "Quantum -Inspired Machine Learning Framework using a Quantum Ising Solver," IEEE Xplore, 2024, Art. no. 10639986.
14. "Quantum -Inspired Optimization Algorithms for Scalable Machine Learning," IEEE Xplore, 2024, Art. no. 10840586.
15. "Projected Quantum Kernel for IoT Data Analysis," IEEE Xplore, 2024, Art. no. 10795417.
16. "The Impact of Feature Embedding Placement in the Ansatz for Quantum Kernels," IEEE Xplore, 2024, Art. no. 10821392.
17. " LLM-BotGuard: A Novel Framework for Detecting LLM-Driven Bots," IEEE Xplore, 2025, Art. no. 10924316.
18. "Fake Social Media Profile Detection Using Machine Learning and Deep Learning," IEEE Xplore, 2025, Art. no. 11069891.
19. "Fake Social Media Accounts Detection using Machine Learning," IEEE Xplore, 2025, Art. no. 11035486.
20. "Advanced Feature Engineering for Twitter Bot Detection," IEEE Xplore, 2025, Art. no. 11364472.
21. "A Comprehensive Study of Bot Detection in Twitter," IEEE Xplore, 2025, Art. no. 11101394.
22. "LGB: Language Model and Graph Neural Network-Driven Social Bot Detection," IEEE Xplore, 2025, Art. no. 11015729.
23. "Graph Neural Networks for Detecting Coordinated Cyber Attacks," IEEE Xplore, 2026, Art. no. 11495613.
24. "Multimodal Learning-Based Relational Graph Neural Network for Social Bot Detection," IEEE Xplore, 2026, Art. no. 11406181.

