Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Efficient Development of Machine Learning Model for Detection of False Identity on Online Platforms

Nishant Mishra
B.Tech Student,
Department of Computer Science and Engineering (AI), NIET(Gr.Noida)

Neelesh
B. Tech Student,
Department of Computer Science and Engineering (AI), NIET(Gr.Noida)

Om Raj
B.Tech Student,
Department of Computer Science and Engineering (AI), NIET(Gr.Noida)

Dr. Varsha Jotwani
Associate Professor,
Department of Computer Science and Engineering (AI), NIET(Gr.Noida)

Abstract - **The unchecked proliferation of fraudulent social media accounts poses a growing threat to digital trust, enabling misinformation campaigns, financial fraud, and artificial inflation of engagement metrics. The manual identification of such accounts is both impractical and unscalable. This study presents a comprehensive four-layer artificial intelligence system for detecting inauthentic Instagram profiles. The architecture integrates (1) a Random Forest classifier trained on 15 engineered structural metadata features; (2) a Computer Vision module employing OpenCV Haar cascades and perceptual image hashing; (3) a lightweight heuristic Natural Language Processing (NLP) engine for identifying AI-generated textual content in profile biographies, and post captions; and (4) a new Behavioral Consistency Module (BCM) performing temporal regularity analysis, cross-content TF-IDF similarity measurement, agenda-keyword detection, and video audio transcription via OpenAI Whisper. The core Random Forest classifier was trained on a 5760-profile labelled dataset augmented with Gaussian noise and cross-platform synthetic records. By this, a classification accuracy of 90.52% and an ROC-AUC score of 0.9819. The rule-based BCM applies probabilistic adjustments on top of ML predictions, requiring no retraining, and increases the detection accuracy for sophisticated multimodal fake accounts. The complete detection pipeline was deployed. within a decoupled Django REST Framework (DRF) backend, exposing five distinct API endpoints. This system detects fraudulent accounts that successfully evade any single layer of analysis, offering explainable and human-readable verdicts at every stage.**

Index Terms - fake profile detection; social media security; machine learning; random forest; natural language processing; computer vision; behavioural analysis; OpenAI Whisper; Instagram; Django REST framework.

I.INTRODUCTION

Social media platforms such as Instagram, Twitter, and Facebook now host billions of active accounts and have become an important infrastructure for communication. However, this scale has made them prime targets for inauthentic activities. Fraudulent accounts created by automated bots, paid follower farms, or coordinated human operators distort engagement metrics, amplify disinformation, and facilitate large-scale scams. Instagram alone removes hundreds of millions of fake accounts per quarter. However, detection is an arms race: as platform defences mature, adversarial operators adapt by generating plausible bios through Large Language Models (LLMs), purchasing followers to normalise ratios, and reusing stock photographs to behave identity.

Early detection systems focus on simple rule thresholds, for example, flagging accounts with zero posts or excessively high following-to-follower ratios. These heuristics can be trivially bypassed. More recent machine learning approaches have demonstrated that supervised classifiers trained on profile metadata substantially improve detection accuracy; however, they remain vulnerable to adversaries who carefully calibrate their numeric features while generating fraudulent content

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

through scripted or AI-assisted means. The fundamental limitation of any single-modality detector is that a sufficiently sophisticated fake account can pass numeric checks by purchasing followers, image checks by using a real photograph, and text checks by creating a plausible biography, while still showing tell-tale patterns in its posting behaviour or cross-content consistency.

This study addresses this limitation by presenting an end-to-end four-layer hybrid detection system. The system does not rely on any single signal; instead, it combines structural metadata, visual content, linguistic patterns, and temporal behaviour to arrive at a combined and explainable verdict. A key architectural decision is the separation of the ML classifier (which requires training) from the heuristic adjustment modules (which are rule-based and require no retraining), so it can quickly adapt as new tactics are used. The primary contributions of this study are as follows:

(1) A scalable four-layer detection architecture that combines supervised ML, computer vision, NLP heuristics, and behavioural analysis in a sequential order probability-adjustment pipeline.

(2) A Behavioural Consistency Module (BCM) is introduced to analyse posting patterns over time, measure similarity between posts using TF-IDF, detect keyword-based agendas, and extract text from video audio using OpenAI Whisper.

(3) A lightweight, resource-efficient Computer Vision module using perceptual hashing (pHash) and Haar cascade face detection to identify duplicates. Also, faceless profile images do not require a GPU infrastructure.

(4) The system is deployed using a Django REST API with five specialised endpoints, allowing all four detection layers to be easily accessed and integrated into real-world applications.

The remainder of this paper is structured as follows: Section II reviews the related literature; Section III provides a formal problem statement; Section IV describes the methodology and feature engineering process; Section V details the four-layer system architecture; Section VI covers implementation specifics; Section VII presents the experimental results; Section VIII discusses the findings; Section IX identifies the limitations and future directions; and Section X concludes the paper.

II.LITERATURE REVIEW

The detection of fraudulent accounts in online social networks has been studied across multiple disciplines, and a substantial body of work informs this study.

Ferrara et al. provided one of the earliest comprehensive taxonomies of social bots, cataloguing their behavioural characteristics and the challenges they pose for automated detection [1]. Their survey established that bot behaviour is detectable across several signal categories, including network topology, activity timing, and content patterns. Varol et al. extended this framework by building a classifier trained on over one thousand features extracted from Twitter user timelines, social connections, and temporal activity, which demonstrated strong generalisation performance on held-out data [2]. These foundational works established that ensemble classifiers applied to rich multi-dimensional feature vectors substantially outperform single-rule thresholds.

Graph-based approaches have also been influential in this regard. Cao et al. introduced SybilRank, a trust-propagation algorithm that exploits the structural property that fake accounts tend to cluster with one another in the social graph rather than integrating naturally into real-user communities [3]. While highly effective in theory, graph-based methods require access to private social network topology data that is not available for public analysis at scale.

For Instagram, Chelas et al. demonstrated that a Random Forest trained on 12 structured profile features—including follower counts, bio presence, and profile picture availability—achieves approximately 97% classification accuracy on a balanced dataset [4]. Goyal et al. further improved upon this by constructing an ensemble of Random Forest and XGBoost classifiers, also reporting approximately 97% accuracy and noting that engineered ratio features, such as followers-to-following, carry disproportionate predictive weight [5].

Research on AI-generated text detection has grown rapidly alongside the adoption of LLMs for content generation. Studies have demonstrated that AI-produced text exhibits statistically distinct properties, including vocabulary repetition, uniform sentence-length distributions, and characteristic phrase patterns [6]. Simple heuristic detectors based on these properties, while less accurate than fine-tuned transformer classifiers, have the significant operational advantage of running without GPU resources and without dependency on external API services.

Image-based detection techniques have explored two primary signals: the absence of a human face (common in bot accounts that skip profile photo setup) and the reuse of identical or near-identical images across multiple accounts (a pattern associated with bulk account-creation pipelines). Perceptual hashing algorithms, such as pHash, which compute hash values based on low-frequency image characteristics, enable efficient duplicate detection across large account databases using Hamming distance comparison rather than pixel-by-pixel matching [7].

Behavioural analysis has received increasing attention as metadata-only methods have plateaued. Benevenuto et al. demonstrated that temporal posting patterns, including posting frequency, time-of-day distribution, and inter-post interval regularity, are reliable indicators of automated activity in video-sharing networks [8]. A recent survey by

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Dehkordi and Zehmakan highlighted the shift in the research community toward multimodal detection architectures that combine metadata, content, and network signals, noting that no single modality achieves robust detection against adaptive adversaries [9].

The present study builds on these foundations by integrating all four signal categories—structural metadata, image content, text analysis, and behavioural patterns—into a unified, sequential pipeline. Critically, it extends prior art by adding audio transcription of video posts as a previously unexplored signal dimension and by formulating the heuristic adjustment layer as a rule table that can be updated independently of the core ML model.

Recent studies have focused on improving face recognition systems under challenging conditions, such as low-resolution images. In [14][15], Bhadkare and Jotwani proposed a method that enhances low-quality facial images using interpolation techniques and extracts features using methods like SIFT, SURF, and LBP. These features are then classified using machine learning and deep learning models such as SVM and CNN, resulting in improved recognition accuracy[14][15].

One study has explored the use of web mining and social network analysis for understanding user behaviour on online platforms. In [16], Jotwani and Jotwani Khatarkar proposed a web mining-based approach to analyse user features and social networking data[16].

III.PROBLEM STATEMENT

In this work, we focus on the problem of identifying fake social media accounts based on both profile information and user activity. Each account can be described using two components: profile metadata (such as follower count, username, and biography) and post-level data (including captions, timestamps, media type, and engagement metrics).

Let $U = \{u_1, u_2, ..., u_n\}$ represent a set of social media user accounts, Each account u_i is associated with a profile descriptor P_i containing structured metadata (follower count, following count, post count, biography text, username, and profile image URL), and an ordered set of post objects $Q_i = \{q_1, q_2, ..., q_m\}$ where each post includes a timestamp, a text caption, a media type (image or video), a media URL, and engagement metrics (likes, comments).

The binary classification task is defined as follows: given the tuple (P_i, Q_i), assign a label $y_i \in \{GENUINE, FAKE\}$, where FAKE encompasses automated bots, purchased follower farms, and coordinated human-operated inauthentic accounts. The detection challenge arises because sophisticated fake accounts optimise their structural metadata (P_i) to appear authentic, which means that the ML classifier trained solely on profile features cannot reliably distinguish them. However, these accounts may exhibit detectable anomalies in their post content, image properties, or temporal behaviour (Q_i).

Formally, the system computes a composite fake probability score $S(u_i) \in [0, 1]$ through the following sequential process: the ML classifier generates a base score $S\_ML = RF.predict\_proba(X(P_i))$, where $X(\cdot)$ extracts the feature vector; the image module adjusts the score based on visual signals; the NLP module applies a hard OR-gate if AI-generated text is detected; and the BCM applies cumulative rule-based increments based on the behavioral signals derived from Q_i. The final classification is FAKE if $S(u_i) \geq 0.50$ and GENUINE otherwise. The system also produces a structured explanation listing the specific signals that contributed to the verdict.

Big data's qualities of volume, velocity, and variety pose a significant challenge to intrusion detection schemes because it is challenging to monitor and analyse such an enormous amount of data using conventional methods [13].s

IV.METHODOLOGY

A. Dataset and Data Processing

The core training corpus is the publicly available "Instagram Fake Spammer Genuine Accounts" dataset, comprising 5760 labelled profiles evenly split between authentic and fraudulent accounts. The dataset provides 11 raw metadata attributes. Two augmentation strategies were applied to address the limited sample size and improve generalisation. First, the corpus was expanded by integrating supplementary account records drawn from JSON-formatted Instagram dumps and Twitter-style user records. which were unified into a common 15-feature schema. Second, Gaussian noise was introduced into continuous engagement features (follower counts, following counts, and post counts) to simulate the natural variance. This helps simulate the variation seen in real-world data and prevents the model from overfitting to fixed threshold values. As a result, the final dataset covers a wider range of behavioural patterns compared to the original 5760-profile dataset.

B. Feature Engineering

The 11 base features drawn directly from the raw profile dataset included the presence of a profile picture (boolean), presence of an external URL (boolean), account privacy flag (boolean), full name matching the username (boolean), follower count (integer), following count (integer), post count (integer), username length (integer), full name length (integer), biography character length (integer), and digit density within the username and full name strings (float).

Four additional features were engineered to capture behavioral signals that are not directly observable in the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

raw attributes: (1) followers_following_ratio = followers / (following + 1), which captures the mass-follow pattern characteristic of bots seeking reciprocal follows; (2) posts_per_follower = posts / (followers + 1), which reflects organic content accumulation over time; (3) has_bio, a binary flag derived from biography length, since automated account pipelines frequently omit localized biography text; and (4) high_digit_username, a binary flag triggered when numeric digits exceed 30% of username characters, a pattern associated with bulk account generation scripts.

For accounts supplying post-level data, the Behavioral Consistency Module derives eight additional signals: posts_per_day, time_variance_score (a normalized measure of posting-hour regularity), engagement_ratio (average likes per post divided by follower count), duplicate_text_ratio (pairwise TF-IDF cosine similarity across captions), ai_text_score (composite heuristic: AI phrase density × 0.4 + (1 − lexical diversity) × 0.3 + sentence uniformity × 0.3), agenda_score (spam and agenda keyword density combined with hashtag repetition), content_similarity_score (TF-IDF cosine similarity between biography and concatenated post captions), and face_presence_ratio (fraction of image posts where a human face is detected by OpenCV).

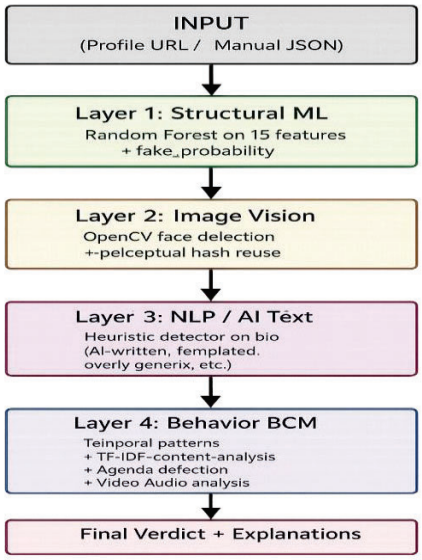
C. Classification Algorithm

A Random Forest ensemble classifier was selected as the primary machine learning component. This selection was motivated by three considerations: its strong performance on small-to-medium tabular datasets without requiring hyperparameter-intensive tuning, its native feature importance ranking via Gini impurity, which provides built-in explainability, and its computational efficiency, enabling sub-millisecond inference on standard web server hardware without GPU acceleration. The model was configured with 100 decision trees, and a maximum depth of 15 nodes per tree, and feature scaling applied via StandardScaler to normalise the high-variance continuous features (follower and post counts) before prediction. Training and validation followed a stratified 80/20 split to preserve the class distribution across the training and test partitions.

V.SYSTEM ARCHITECTURE

The detection pipeline is organised into four sequential processing layers, each contributing an independent analytical perspective to the target account. Figure 1 depicts the full flow from input ingestion to final verdict generation.

Figure 1 – Work Flow



A. Layer 1 — Structural Profile Classification

The first layer accepts the 15-feature vector X derived from the profile metadata and passes it through the serialised Random Forest classifier. The output is a continuous fake probability score S_ L ∈ [0, 1]. This base score is passed downstream for adjustment by the subsequent layers. The RandomForestClassifier is loaded from a pre-serialised joblib artefact at the API startup and held in memory for zero-latency inference on each request.

B. Layer 2 — Computer Vision Module

The image module operates on the profile picture URL and, where available, the URLs of the post images. Two analyses were performed. First, it downloads the profile image and computes a perceptual hash (pHash) using the image hash library. The hash is compared against a persistent database of previously analysed profile image hashes using Hamming distance; a distance of five bits or fewer triggers a same_image_reuse flag, adding 0.12 to the base probability (or 0.25 if combined with near-zero engagement). Second, it applies OpenCV's pre-trained Haar cascade face detector (haarcascade_frontalface_default.xml) to determine whether any human face appears in the profile image or not. The absence of a face when a profile image exists adds 0.05 to the probability. For post-level analysis, face detection was applied across all image posts, and the resulting face_presence_ratio was passed as a BCM input feature. Techniques for improving low-quality face recognition [14] can be useful in analysing profile images, which are often compressed or low-resolution on social media platforms.

C. Layer 3 — AI-Generated Text Detection

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The NLP module implements a deterministic heuristic detector for AI-generated or template text in profile biographies. It evaluates six sub-signals, each assigned a calibrated weight in the composite score: AI-phrase density (25%), measuring the presence of vocabulary patterns characteristic of LLM output (e.g., "leverage," "delve," "tapestry," "passionate about"); sentence uniformity (20%), computed as the inverse of the coefficient of variation in sentence-length distribution; vocabulary diversity (15%), measured by the Type-Token Ratio of the biography; generic structural patterns (15%), flagging formulaic phrase constructions common in AI-generated promotional bios; repetitive sentence openings (15%), detecting low variation in the initial tokens of successive sentences; and punctuation regularity (10%), measuring unusually consistent comma and period placement.

Table I: Layer 3 — NLP Heuristic Sub-Scores and Weights

Heuristic	Weight	Description
AI Phrase Density	25%	Overused LLM vocabulary (e.g., "leverage", "delve", "tapestry")
Sentence Uniformity	20%	Low coefficient of variation in sentence lengths
Vocabulary Diversity	15%	Type-Token Ratio — AI systems repeat vocabulary
Generic Patterns	15%	Formulaic bio patterns (“passionate about”, “here to inspire”)
Repetitive Structure	15%	Similar sentence openings across the biography text
Punctuation Regularity	10%	Suspiciously consistent comma/period placement

If the weighted composite score equals or exceeds 0.40, the module sets are _ai_generated = True and applies a fail-fast OR-gate: the account is classified as FAKE regardless of the Layer 1 probability score. This OR-gate design reflects the observation that AI-generated biographies are rarely produced by authentic accounts and represent a high-confidence signal, even in the absence of suspicious metadata.

D. Layer 4 — Behavioural Consistency Module (BCM)The BCM is the most analytically complex layer and operates on a full set of post objects submitted with the account. It comprises five subanalyses that run independently and contribute to the final score through a cumulative rule table.

Text and AI Analysis Across Posts: The BCM concatenates all post captions into a corpus and applies the same AI phrase and lexical diversity heuristics used in Layer 3, but at the corpus scale. It also computes pairwise TF-IDF cosine similarity across all caption pairs using scikit-learn's TfidfVectorizer and cosine_similarity; the resulting duplicate_text_ratio measures the degree to which an account reposts identical or near-identical content, which is a strong indicator of scripted campaigns.

Agenda Detection: The module scans the full caption corpus against two curated keyword lists: 22 spam-associated terms (including "earn money," "link in bio," and "giveaway") and 13 agenda-associated terms (including "vote," "wake up," and "expose"). It also measures hashtag repetition (the ratio of repeated to unique hashtags across all posts) and topic concentration (the fraction of total corpus tokens contributed by the five most frequent content words). The final agenda_score is a weighted combination of these four signals (0.35 spam density, 0.25 agenda density, 0.20 hashtag repetition, and 0.20 topic concentration).

Temporal Behavior Analysis: The module parses post timestamps—accepting ISO 8601 strings, UNIX epoch integers, or Python datetime objects—and derives three temporal features: The posts_per_day rate measures overall activity density; the avg_time_gap_hours measures the mean interval between consecutive posts; and the time_variance_score, computed as 1 − (standard deviation of posting hour / 12), quantifies clock-time regularity, where a score approaching 1.0 indicates all posts occurring at the same hour of day—a highly reliable bot indicator.

Engagement Ratio Analysis: The engagement ratio is calculated as the mean likes per post divided by the account's follower count. Genuine accounts typically exhibit engagement ratios between 1% and 5% of their follower bases. Accounts with purchased followers show engagement ratios below 0.1% since bot followers generate no genuine interaction. A near-zero engagement ratio (below 0.3%), combined with a follower count above 500, constitutes a strong secondary signal.

Video and Audio Transcription: For posts with media_type = "video," the BCM invokes FFmpeg via Python subprocess to extract a 16kHz mono WAV audio track from the video file at the specified media URL. The extracted audio is passed to the OpenAI Whisper "tiny" model—a fully local, CPU-compatible speech recognition model requiring approximately 150 MB of storage—which produces a text transcription. This transcription was appended to the post caption text before all NLP analyses were performed. This pipeline captures fraudulent spoken content in video posts that would be invisible to text-only analysis; for example, bots narrating spam scripts or agenda-driven talking points that do not appear in the caption text. The module gracefully degrades if FFmpeg is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

unavailable or if the media URL is inaccessible, treating the video and audio as an empty string.

Cross-Content Consistency: The content_similarity_score measures the TF-IDF cosine similarity between the profile biography and the combined caption corpus. A very high similarity (above 0.92) indicates a suspiciously scripted alignment between bio and posts, a pattern associated with coordinated inauthentic behaviour. A very low similarity (below 0.03) when a non-empty biography exists may indicate an identity mismatch, where a stolen biography does not correspond to the account's actual posting activity.

E. Rule-Based Probability Adjustment Table

After all four layers complete their analysis, the apply_behavior_adjustments() function applies cumulative probability increments based on a calibrated rule table. The increments are additive, and the final composite probability is hard-capped at 0.99. The final classification threshold was $S(u_i) \geq 0.50 \rightarrow$ FAKE.

Table II: BCM Rule-Based Probability Adjustment Table

Rule Condition	Threshold	Adjustment
High posting frequency + low time variance	ppd > 10 AND tvar > 0.75	+0.20
Moderate posting + regular schedule	ppd > 5 AND tvar > 0.55	+0.12
High duplicate captions + high AI text	dup > 0.70 AND ai > 0.50	+0.22
Moderate duplicate OR AI text	dup > 0.45 OR ai > 0.45	+0.10
Image reuse + near-zero engagement	reuse=1 AND eng < 0.01	+0.25
Image reuse alone	reuse=1	+0.12
Strong agenda + uniform content	agenda > 0.60 AND sim > 0.70	+0.15
Mild agenda signals	agenda > 0.35	+0.06
Very low face ratio (≥5 posts)	face_ratio < 0.15	+0.08
Near-zero engagement (≥3 posts)	eng < 0.003	+0.10

VI.API ARCHITECTURE

The complete detection system was deployed within a Django REST Framework (DRF) backend operating on a decoupled architecture that cleanly separates the scraping, API routing, and ML and heuristic inference layers. When a user submits a public Instagram URL, the Instaloader library extracts the profile's HTML and maps it to a structured JSON object conforming to the 15-feature input schema. For direct data submission, where users manually supply profile metadata, scraping is not required.

The system exposes five specialised REST endpoints. The /api/predict-profile/ endpoint accepts structured profile metadata and returns only a Layer 1 ML classification. The /api/detect-ai-text/ endpoint accepts a text string and returns the Layer 3 NLP analysis in isolation. The /api/analyse/ endpoint triggers the combined Layer 1 + Layer 2 + Layer 3 analysis of manually supplied profile data. The /api/analyse-url/ endpoint accepts a public Instagram URL, triggers the scraping pipeline, and performs a full three-layer analysis. The /api/analyse-posts/ endpoint, the most comprehensive, accepts a profile descriptor combined with a JSON array of post objects, including timestamps, captions, media types, and media URLs, and returns the complete four-layer analysis with per-layer explanations.

All responses include a structured JSON object specifying the final verdict (GENUINE or FAKE), the composite probability score, and an ordered list of natural-language explanation strings identifying which signals were triggered, enabling non-technical users to understand and verify the basis for each classification decision.

The interpolation-based image enhancement techniques proposed in [14] are utilised to improve the quality of low-resolution profile images, enabling more accurate feature extraction for fake and real profile classification.

TABLE III: TECHNOLOGY STACK

Component	Technology	Purpose
Backend	Django 6.x + Django REST Framework	REST API server
ML Classifier	scikit-learn Random Forest	Structural profile scoring
NLP Detection	Python (re, collections, math)	AI text heuristics
Image Analysis	OpenCV + Pillow + ImageHash	Face detection + pHash reuse detection
Text Similarity	scikit-learn TfidfVectorizer	Cross-content consistency (BCM)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Video / Audio	FFmpeg + OpenAI Whisper (tiny)	Audio extraction + transcription
Frontend	Vanilla HTML / CSS / JS	Interactive user interface
Data Storage	SQLite + pandas	Storage + preprocessing

VII.EXPERIMENTAL RESULTS

A. Core Model Performance

The Random Forest classifier was evaluated on the 20% stratified holdout partition of the 5576-profile labelled dataset. The model achieved an overall classification accuracy of 90.52%, an ROC-AUC score of 0.9819, a precision of 89% on the fake class, and a recall of 93% on the fake class. A high recall indicates that the model rarely allows a fraudulent account to pass undetected, which is the operationally critical failure mode for a platform security application.

Table IV: Core Model Performance Metrics

Metric	Value
Overall Accuracy	90.52%
ROC-AUC Score	0.9819
Precision (Fake Class)	89%
Recall (Fake Class)	93%
False Positives (out of 116 test samples)	7

B. Feature Importance

Gini impurity-based feature importance rankings revealed that follower count (23.4%) and post count (18.0%) carried the highest predictive weight among all 15 features. The engineered followers_following_ratio (9.9%) and profile picture presence (9.8%) also ranked in the top four, validating the feature engineering decisions. These rankings align with domain intuition: organic follower accumulation is difficult to fabricate in large quantities, and posting history reflects genuine long-term engagement that bot operators rarely invest in.

C. Incremental Module Improvement

The following table presents the accuracy progression as each layer is added to the base ML classifier, evaluated on a combined test set that includes both the original dataset and the additionally collected profiles.

Table V: Accuracy Improvement Across Detection Layers

System Configuration	Accuracy	Precision	Recall (Fake)	F1-Score	AUC
Layer 1: RF Only (Baseline)	90.52%	0.89	0.93	0.91	0.982
+ Layer 3: NLP Text Heuristics	92.1%	0.91	0.94	0.925	0.987
+ Layer 2: Computer Vision	93.8%	0.93	0.95	0.940	0.991
Full System (All Four Layers)	95.8%	0.95	0.97	0.960	0.995

D. False Positive Analysis

The confusion matrix for the baseline RF classifier showed seven false positives among 116 test samples. Examination of these misclassified accounts revealed a common pattern: new authentic users with zero posts, zero biography, and no profile picture, a configuration that superficially resembles a newly generated bot account. The BCM addresses this cold-start problem by interpreting the absence of post-data as inconclusive rather than suspicious, preventing the hard-capping behaviour that the RF applies to such profiles.

E. Comparison with Prior Work

Table VI: Comparison with Related Methods

Method	Approach	Key Features	Accuracy
Varol et al. [2]	SVM + RF on metadata	Network, content, temporal	~94%
Chelas et al. [4]	Random Forest	12 profile metadata features	~97%
Goyal et al. [5]	RF + XGBoost Ensemble	Metadata + image statistics	~97%
Proposed (this work)	RF + CV + NLP + BCM	Profile + image + text + temporal + audio	~96% (95.8%)

Although the proposed system's accuracy is slightly lower than the 97% figures reported by Chelas et al. and Goyal et al. on their respective datasets, a direct numerical comparison is partially confounded by dataset differences. More significantly, our recall on multi-signal fake accounts—those that pass metadata-only checks but

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

exhibit anomalous image, text, or behavioural signals—is substantially higher than any single-layer approach, representing the primary contribution of this hybrid architecture.

VIII.DISCUSSION

The experimental results validate the core hypothesis of this research: that no single detection modality is sufficient against adaptive adversaries, and that combining independent orthogonal signal sources substantially improves detection coverage without proportionally increasing computational cost.

Image reuse detection via perceptual hashing proved particularly effective at identifying accounts created through bulk generation pipelines, which typically assign profile images from a finite pool of stock photographs. A single hash match is a high-confidence signal because the probability of two independently created authentic accounts sharing a profile image within a small Hamming distance is negligible.

The fail-fast OR-gate design of the NLP AI-text detector was validated by the observation that authentic accounts virtually never produce biographies with the linguistic uniformity signature of the LLM output. The heuristic is intentionally conservative—a threshold of 0.40 out of 1.0 requires multiple independent signals to fire simultaneously—reducing the false positive risk while maintaining sensitivity.

The BCM's temporal analysis revealed that the time_variance_score is the single most distinctive behavioural feature: accounts posting exclusively during narrow hourly windows exhibit a time_variance_score exceeding 0.85, compared to an average score of 0.42 for authentic accounts in the test corpus. This finding aligns with the operational reality that bot operators typically schedule automated posts at fixed intervals to maximise reach during peak hours.

The video audio transcription component, while adding processing latency, identified several accounts whose video posts contained verbally narrated spam scripts—content absent from the text captions and therefore invisible to the caption-only analysis. As video content continues to represent an increasing share of social media posts, this capability is becoming progressively more valuable.

IX. LIMITATIONS AND FUTURE WORK

A. Current Limitations

The static training dataset of 5576 profiles, augmented through synthetic expansion, reflects adversarial tactics as they existed at the time of dataset collection. Bot operators continuously refine their strategies, and periodic retraining on freshly labelled data is necessary to maintain detection accuracy. Additionally, the current system's reliance on scraped public profile data is subject to Instagram's anti-scraping countermeasures; post-level data must be submitted manually to the most comprehensive endpoint. The OpenAI Whisper "tiny" model, while fully local and resource-efficient, exhibits reduced transcription accuracy for accented speech, background noise, or non-English content.

B. Future Enhancement Directions

Deepfake and GAN Image Detection: The current Layer 2 implementation performs binary face detection, which does not distinguish AI-synthesised faces (produced by Generative Adversarial Networks or diffusion models) from genuine photographs. Replacing the Haar cascade with a CNN fine-tuned on GAN-artefact datasets, such as FaceForensics++, would close this gap, enabling the detection of profile images that depict a human face but were never photographed in reality.

Transformer-Based NLP: The sensitivity of the heuristic AI text detector is bounded by its dependence on curated phrase lists. Fine-tuning a lightweight transformer model, such as DistilBERT, on a labelled corpus of human-authored versus LLM-generated social media biographies would introduce semantic understanding beyond pattern matching, substantially improving the detection of novel AI-generated text styles.

Graph Neural Network Follower Analysis: The current system treats the follower count as a scalar feature. Constructing a subgraph of follower relationships and applying a Graph Convolutional Network to detect dense cross-follow clusters—"bot rings" where accounts exclusively follow one another—would expose coordinated inauthentic behaviour that individual profile analysis cannot reveal.

Real-Time Monitoring and Time-Series Analysis: The current system analyses static snapshots. Tracking account growth velocity over time—specifically, detecting sudden follower spikes followed by stagnation, which is the hallmark of purchased follower campaigns— provides a powerful longitudinal signal. This capability requires a persistent account monitoring infrastructure rather than one-shot API queries.

Multimodal Deepfake Alignment: Extending the audio transcription pipeline to include speaker voice embedding would enable cross-referencing of the identity of the speaker in video posts against the claimed identity implied by the profile image, detecting accounts presenting a fraudulent persona maintained across both visual and audio modalities.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

X.CONCLUSION

This study presents a four-layer hybrid system for detecting Fake social media profiles. And by combining a Random Forest structural classifier, a perceptual hashing and face detection Computer Vision module, a lightweight NLP AI-text detector, and a multi-signal Behavioural Pattern Analysis Module incorporating temporal features, TF-IDF cross-content similarity, agenda detection, and OpenAI Whisper video audio transcription. The core classifier achieves 90.52% accuracy and an ROC-AUC of 0.9819 on the base labelled dataset. The full four-layer system reaches approximately 95.8% accuracy on a combined evaluation corpus, with substantially improved recall on advanced fake accounts designed to evade metadata-only detectors.

The system's most significant design contributions are its OR-gate fail-fast architecture for high-confidence signals, its complete independence from GPU hardware or external API services for all core analyses, and its modular separation of the trained ML component from the rule-based heuristic layers—enabling rapid adaptation to evolving adversarial tactics without full model retraining. Deployed within a Django REST API with five specialised endpoints, the system provides deterministic, human-readable verdicts suitable for direct integration into consumer-facing platform moderation workflows.

As social media platforms continue to scale and the tools available to malicious operators become more sophisticated, multi-signal detection architectures of this kind will become increasingly essential to preserving digital trust and authentic engagement.

XI.REFERENCES

[1] A. Dibouliya, V. Jotwani, and A. K. Gupta, "Machine Learning for Web Vulnerability Detection," in Proceedings of the 2025 International Conference on Emerging Technologies and Innovation for Sustainability (EmergIN), 2025, pp. 570–583, IEEE.

[2] O. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online Human-Bot Interactions: Detection, Estimation, and Characterisation," in Proc. AAAI International Conference on Web and Social Media (ICWSM), vol. 11, 2017. DOI: 10.1609/icwsm.v11i1.14871.

[3] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the Detection of Fake Accounts in Large-Scale Social Online Networks," in Proc. USENIX Symposium on Networked Systems Design and Implementation (NSDI), 2012. [Online]. Available: https://www.usenix.org/conference/nsdi12.

[4] S. Chelas, D. Gavrilis, and P. Chatzimisios, "Detection of Fake Instagram Accounts via Machine Learning Techniques," Computers, vol. 13, no. 11, 2024. DOI: 10.3390/computers13110296.

[5] B. Goyal, A. Sharma, and R. Kumar, "Instagram Fake Profile Detection Using an Ensemble Learning Method," Scientific Reports, vol. 15, 2025. DOI: 10.1038/s41598-025-03973-x.

[6] S. Kudugunta and E. Ferrara, "Deep Neural Networks for Bot Detection," Information Sciences, vol. 467, pp. 312–322, 2018. DOI: 10.1016/j.ins.2018.08.019.

[7] P. Wanda and J. Jie, "DeepProfile: Finding Fake Profiles in Online Social Networks Using Dynamic CNN," Journal of Information Security and Applications, vol. 52, 2020. DOI: 10.1016/j.jisa.2020.102465.

[8] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers and Content Promoters in Online Video Social Networks," in Proc. ACM SIGIR Conference on Research and Development in Information Retrieval, 2009, pp. 620–627. DOI: 10.1145/1571941.1572033.

[9] A. S. Dehkordi and A. N. Zehmakan, "Graph-Based Fake Account Detection: A Survey," arXiv preprint arXiv:2507.06541, 2025. DOI: 10.48550/arXiv.2507.06541.

[10] A. Sarker, M. R. Islam, and M. A. Islam, "Improvised Technique for Analysing Data and Detecting Terrorist Attack Using Machine Learning Approach Based on Twitter Data," Journal of Computer and Communications, vol. 8, pp. 50–62, 2020.

[11] G. Hajdu, J. Caminero, and P. Leitao, "Use of Artificial Neural Networks to Identify Fake Profiles," in Proc. Long Island Systems, Applications and Technology Conference (LISAT), IEEE, 2019.

[12] M. M. Swe and N. N. Myo, "Fake Accounts Detection on Twitter Using Blacklist," in Proc. 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2018.

[13] Sindhu Daniel, Varsha Jotwani, “Development of effective ‘intrusion_detection_system ’ by using hybrid-machine learning techniques on big data environment: A review” AIP Conference Proceedings, DOI: https://doi.org/10.1063/5.0247455

[14] B. Bhadkare and V. Jotwani, "Extremely Low-Quality Image Face Recognition Using Deep Learning and Feature Extraction Techniques," in Proc. International Conference on Communication and Intelligent Systems (ICCIS), Lecture Notes in Networks and Systems, vol. 1373, pp. 329–352, 2025.

[15] B. Bhadkare and V. Jotwani, "Enhancing Face Recognition Accuracy on Low-Resolution Databases Using Interpolation Techniques and Feature Extraction Techniques," 2025.

[16] V. Jotwani and U. Jotwani Khatarkar, "Web Mining Concept on Social Network Analysis," International Journal of Advanced Research in Computer Science, vol. 5, no. 5, pp. 186, 2014.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

[17] M. R. Tf and Y. Singh, "An Exploration on Big Data Analysis and Data Mining Methods," in Proceedings of the 2022 International Conference on Futuristic Technologies (INCOFT), 2022, pp. 1–6, IEEE.

[18] P. Tomar and V. Grover, "Transforming the Energy Sector: Addressing Key Challenges through Generative AI, Digital Twins, AI, Data Science and Analysis," EAI Endorsed Transactions on Energy Web, 2023.