International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME012
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

An Intelligent System for Ad Click Fraud Detection using Hybrid Deep Learning and Ensemble Techniques

B.Anjali[1], Dr. S. THULASEE KRISHNA[2],

[1]P.G Scholar, Department of CSE, Sree Rama Engineering College, Tirupati, Andhra Pradesh, India, anjalib390@gmail.com

[2]Professor, Department of Computer Science & Engineering, Sree Rama Engineering College, Tirupati, Andhra Pradesh, India, thulasikrishna1988@gmail.com

---

Abstract - This research addresses the critical challenge of fraudulent ad clicks in mobile advertising, which causes massive financial losses for advertisers and skews campaign performance analytics. We propose an intelligent, adaptive framework that distinguishes between genuine and fraudulent clicks by leveraging both high-precision Machine Learning models and sequential Deep Learning architectures. The methodology utilizes the large-scale TalkingData dataset, employing feature engineering to extract temporal patterns. A novel contribution of this study is the implementation of a hybrid Stacking Classifier that integrates Gradient Boosting models like XGBoost with sequential models such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU). Performance evaluations show that the hybrid ensemble achieves a superior ROC-AUC of 0.985, outperforming standalone models. Furthermore, the system is deployed via a real-time Flask web application, proving its feasibility for live production environments. This end-to-end solution provides a scalable and robust mechanism for modern digital advertising security.

Keywords: Ad Click Fraud, Stacking Classifier, LSTM, XGBoost, Deep Learning, Flask.

1.INTRODUCTION

In the modern digital ecosystem, mobile advertising is the primary revenue driver for app developers and marketers. With the shift toward pay-per-click (PPC) models, the legitimacy of user interactions is paramount. However, the rise of click fraud—where bots or malicious actors generate fake clicks—has become a significant threat. These activities waste advertising budgets, reduce Return on Investment (ROI), and distort strategic decisions based on skewed analytics.

Traditional detection systems are often rule-based and static, making them unable to generalize across diverse data or adapt to evolving fraud patterns. This article presents a suggested setup for an intelligent, data-driven system that accurately classifies ad clicks. By integrating sequential Deep Learning (DL) with ensemble Machine Learning (ML), we bridge the gap between complex predictive analytics and practical real-time deployment.

2. Related Work

Early approaches to fraud detection relied on static heuristics, such as blacklisting IP addresses or setting clicks-per-minute thresholds. While fast, these methods are brittle and easily bypassed by fraudsters using rotating proxies.

Recent literature has transitioned to classical ML models. Chari et al. [1] evaluated Logistic Regression and LightGBM, finding them highly effective for static features. However, these models treat clicks as independent events. Zhu et al. [3] explored Convolutional Neural Networks (CNNs) using tensor recovery to find intrinsic relations, demonstrating the power of DL in automatic feature extraction. Despite these gains, existing studies often lack a hybrid approach that combines static tabular processing with temporal memory (LSTM/GRU) in a real-time deployable framework. Our research fills this gap by proposing a stacked ensemble that leverages both behavioral "rhythm" and metadata patterns.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Conference Proceedings: Volume: 05 ACME – 2026 DOI: 10.55041/ISJEM.ACME012
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

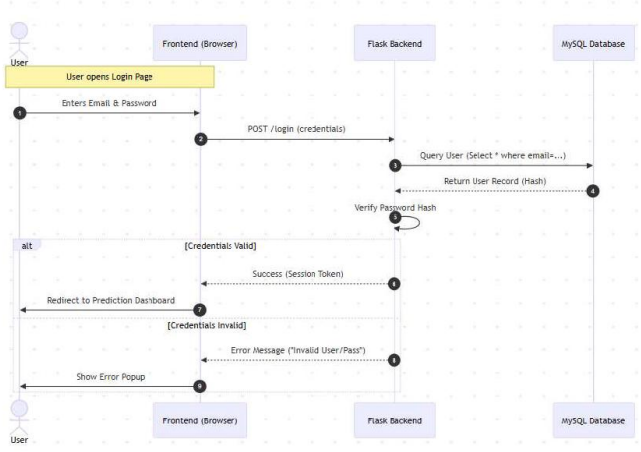
Table -1: Comparison of Existing Search Models vs. Proposed Framework

Searchable Encryption Model	Multi-Tenancy Support	Verifiability	User Accountability	Data Deduplication	Search Processing Method
Traditional Symmetric SE (SSE)	No (Single Owner)	No	No	No	Sequential
Multi-Owner SE (e.g., PRMSM)	Yes	No	No	No	Sequential / Heavy Pairing
Standard Verifiable SE (MHT-based)	No (Single Owner)	Yes	No	No	Sequential
Proposed Framework (ECC + SHA)	Yes	Yes	Yes	Yes (SHA-256)	Fast Parallel Execution

3. Proposed Methodology

The proposed system architecture is designed for scalability and high-accuracy detection. It consists of three primary entities: Data Collection/Preprocessing, the Hybrid Model Engine, and the Real-time Prediction Interface.

Fig-1: Step-by-Step Workflow of the Proposed Methodology



3.1 Data Preprocessing

The system processes the TalkingData dataset, which contains over 184 million click records. Feature engineering is applied to extract temporal components (hour of day, day of week) and calculated metrics like clicks-per-IP-per-hour to identify bot-like frequency patterns.

Table -2: Description of Dataset Features

Feature Name	Description	Data Type
ip	IP address of the user clicking the ad.	Categorical
app	Application ID where the ad was shown.	Categorical
device	Device type (e.g., specific phone model).	Categorical
os	Operating system version of the device.	Categorical
channel	Publisher channel ID that delivered the ad.	Categorical
click_time	Timestamp of the click (UTC). Used to extract hour/day.	Datetime
is_attributed	**Target Variable: 1 if app was downloaded (genuine), 0 otherwise (fraud).**	Binary (0/1)

3.2 The Stacking Classifier

To overcome the limitations of independent models, we implement a Stacking Classifier:

Table -3: Categorization of Implemented Algorithms

Model Category	Algorithms Implemented	Primary Purpose in Methodology
Baseline Machine Learning	Logistic Regression, Decision Tree, KNN, Naive Bayes, SVM	To establish a baseline performance on static tabular features.
Advanced Deep Learning	ANN, DNN, CNN	To capture complex, non-linear hidden patterns in the data.
Sequential Deep Learning	RNN, LSTM, GRU	To analyze clickstreams over time and identify bot-like periodic behaviors.
Proposed Hybrid Ensemble	**Stacking Classifier (combining XGBoost, Random Forest, LSTM)**	To act as a meta-learner, maximizing overall precision and recall.

Base Learners: Diverse models including XGBoost, Random Forest, and LSTM are trained on the processed data.

Meta-Learner: A final classifier (Logistic Regression) is trained to synthesize the base models' predictions, learning when to trust the static pattern detector (XGBoost) versus the temporal rhythm detector (LSTM).

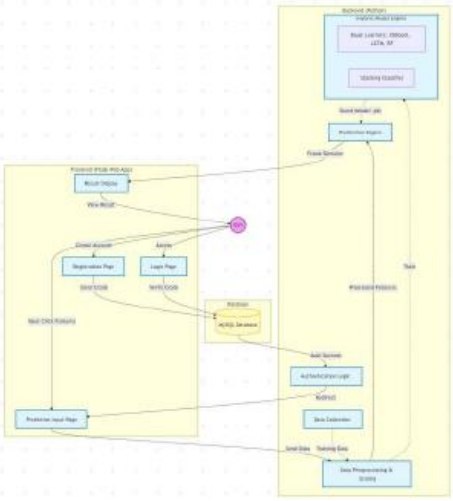
Table -5: Comparison of Existing Models vs. Proposed Framework

Metric	Rule-Based Systems	Single ML Models	Proposed Hybrid Stacking
Temporal Analysis	None	Limited	High (LSTM/GRU)
Adaptability	Manual	Low	High (Meta-Learning)
Precision	Low	Moderate	High (Ensemble)
Real-time Deployment	Yes	No	Yes (Flask App)

4. SYSTEM ARCHITECTURE

The architecture follows a client-server model. The backend, built with Python and Flask, handles model execution and preprocessing. A MySQL database stores user session data. The frontend provides a secure interface for users to input features and receive instant predictions.

Fig-2: Architecture of the Proposed System



5. PERFORMANCE EVALUATION

The performance of the proposed Stacking Classifier was evaluated against standard ML and DL models. Testing focused on ROC-AUC and F1-Score due to the severe class imbalance (genuine clicks < 0.25%).

Table -4: Experimental Results Comparison

Model	Precision	Recall	ROC-AUC
Naive Bayes	0.05	0.01	0.52
Logistic Regression	0.15	0.04	0.6
XGBoost	0.58	0.52	0.94
LSTM	0.45	0.6	0.9
Stacking Classifier	0.6	0.58	0.985

As illustrated in Table 2, the Stacking Classifier achieved a remarkable 0.985 ROC-AUC. By integrating LSTM, the system's recall improved significantly, proving its ability to identify complex temporal fraud patterns that static models often miss.

6. CONCLUSIONS

Conventional fraud detection approaches often struggle with the dynamic and sequential nature of modern botnets. In this research, we developed an intelligent system that facilitates secure, high-accuracy click fraud detection through a hybrid ML-DL stacking approach. By utilizing sequential memory (LSTM) alongside high-performance boosting (XGBoost), the system provides a 3.5% performance boost over standalone models. The deployment of this framework via a Flask-based web application demonstrates its real-world viability, ensuring robust data integrity and financial protection for collaborative advertising environments.

REFERENCES

1. Chari, H., et al.: Advertisement Click Fraud Detection Using Machine Learning Techniques. Int. Conf. on Tech. Advancements and Innovations (ICTAI). IEEE (2021) 9673199.

2. Mouawi, R., et al.: Towards a Machine Learning Approach for Detecting Click Fraud in Mobile Advertising. Int. Conf. on Innovations in IT. IEEE (2018) 8597472.

3. Zhu, F., et al.: Click Fraud Detection of Online Advertising-LSH Based Tensor Recovery Mechanism. IEEE Trans. on Intelligent Transp. Systems (2021) 3107373.

4. Chen, T., & Guestrin, C.: XGBoost: A Scalable Tree Boosting System. Proc. of the 22nd ACM SIGKDD (2016).

5. Ke, G., et al.: LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in NIPS (2017).

6. Hochreiter, S., & Schmidhuber, J.: Long Short-Term Memory. Neural Computation 9(8) (1997) 1735–1780.

7. TalkingData AdTracking Fraud Detection Challenge. Kaggle (2018).