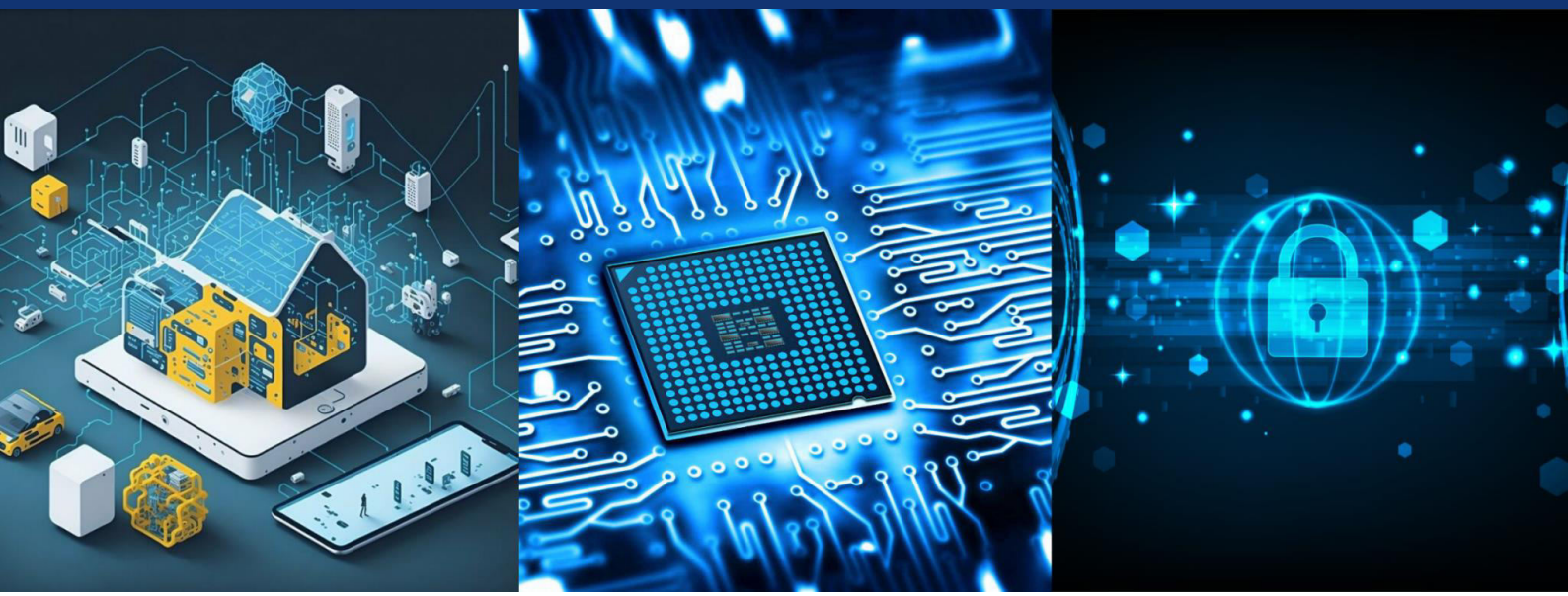
IJIRCCE

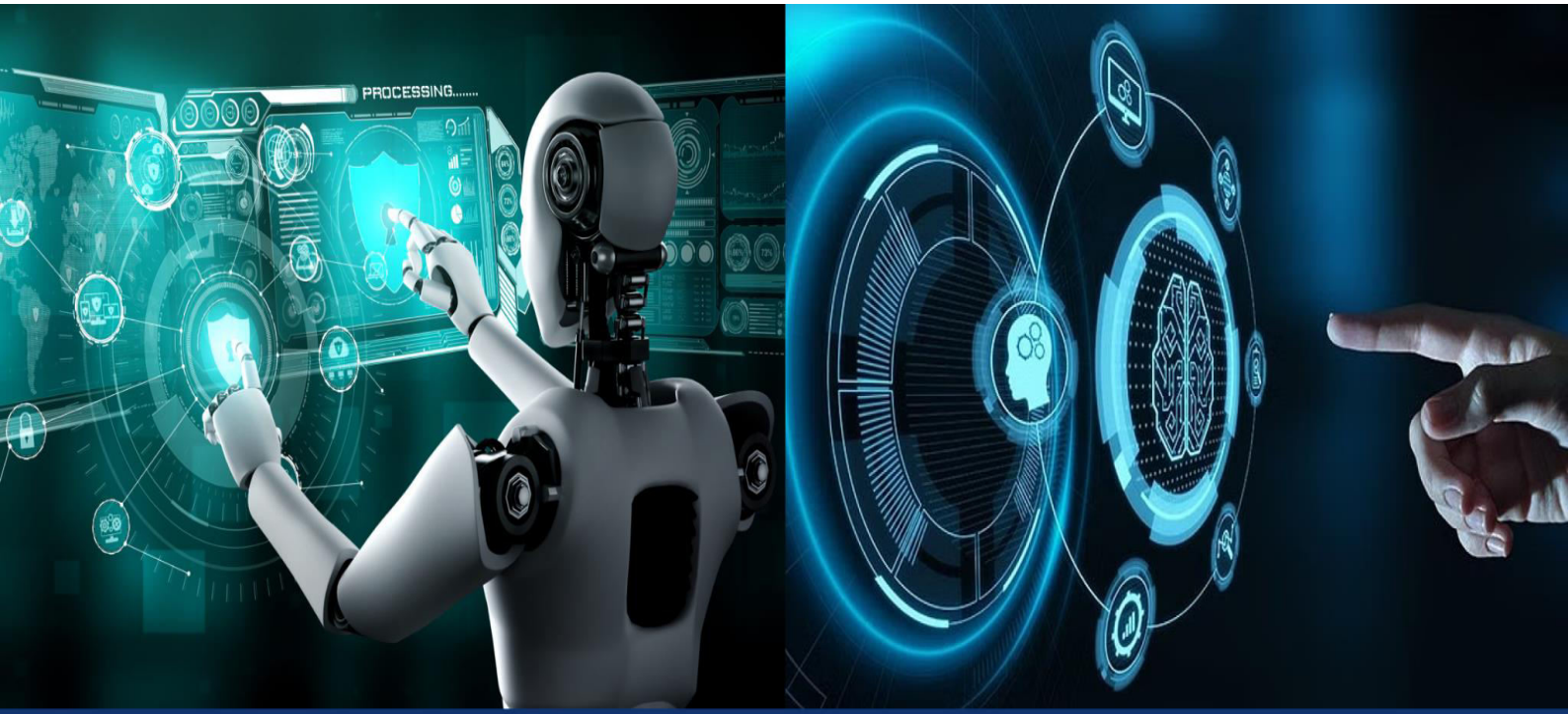
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.927

Volume 14, Issue 6, June 2026

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406023

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Advanced Fake Job Post Prediction using Machine Learning for Online Recruitment Scam Detection

K Navya Sree, Dr.D. Suresh Reddy

PG Student, Dept. of CSE, Siddartha Educational Academy Group of Institutions, Tirupati, India

Assoc. Professor, Dept. of CSE(AI&ML), Siddartha Educational Academy Group of Institutions, Tirupati, India

ABSTRACT: The exponential growth and universal popularity of online recruitment platforms have fundamentally revolutionized the modern hiring landscape, seamlessly connecting global employers with prospective job seekers. However, the unchecked proliferation of these digital platforms has inadvertently catalyzed a significant and alarming rise in fraudulent job advertisements. These sophisticated scams are meticulously designed to deceive desperate or unsuspecting applicants, ultimately aiming to maliciously harvest highly sensitive personal data, financial information, or direct monetary deposits.

Traditional manual verification methodologies, reliant on human moderation, are fundamentally insufficient to handle the sheer velocity, variety, and volume of online job postings generated daily. This exhaustive research paper proposes an advanced, highly scalable machine learning framework explicitly engineered for detecting fake job advertisements. By extensively analyzing complex textual, contextual, and organizational features autonomously extracted from raw recruitment data, the proposed system leverages advanced Natural Language Processing (NLP) techniques for rigorous feature extraction.

The study meticulously evaluates a diverse array of machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), Random Forest, XGBoost, and sophisticated voting ensemble classifiers. Experimental results comprehensively demonstrate that the proposed model effectively and consistently identifies fraudulent job postings, achieving exceptional metrics across accuracy, precision, recall, and the F1-score. Ultimately, this study significantly contributes to fortifying recruitment cybersecurity, preserving digital platform integrity, and critically protecting vulnerable job seekers from devastating online employment scams.

KEYWORDS: Fake Job Detection, Recruitment Scam Detection, Machine Learning, Natural Language Processing, Online Recruitment, Fraud Detection, Classification, Ensemble Learning, Predictive Modeling.

I. INTRODUCTION

The advent of the digital age has shifted paradigms across countless industries, with the human resources and recruitment sector experiencing a profound transformation. Online recruitment platforms, professional networking sites, and digital job portals have rapidly become the primary, and often exclusive, conduits for employment opportunities across all global industries. Organizations of all sizes now heavily utilize these digital ecosystems to advertise vacancies, build talent pools, and seamlessly recruit highly qualified candidates. The efficiency, cost-effectiveness, and expansive reach of these systems offer undeniable advantages to both the corporate entity and the job seeker. Furthermore, the globalization of the workforce has mandated the use of these platforms to bridge geographic divides between talent and opportunity.

Despite these unprecedented operational advantages, the open and largely decentralized nature of online recruitment systems has rendered them highly attractive targets for cybercriminals and malicious actors. These entities continuously create highly convincing, fraudulent job advertisements expressly to exploit vulnerable job seekers. The architecture of a fake job posting is often psychologically manipulative, promising unrealistically high salaries, highly attractive fringe benefits, overly flexible remote work arrangements, and practically guaranteed employment opportunities with minimal required qualifications.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The consequences of interacting with these fraudulent postings are severe and far-reaching. Victims are frequently subjected to elaborate phishing schemes where they are coerced into providing highly sensitive personally identifiable information (PII), such as bank account details, government identity documents, or social security numbers. In more direct financial scams, applicants are manipulated into making upfront payments disguised as mandatory processing fees, background check costs, or premium training program tuitions. These fraudulent activities consistently result in devastating financial losses, catastrophic identity theft, and profound psychological distress among the targeted job seekers, while simultaneously eroding public trust in legitimate recruitment platforms and damaging the reputations of the companies being impersonated.

The rapid and continuous growth in the sheer volume of daily job advertisements renders traditional manual verification and human moderation computationally and practically impossible. An army of human moderators could not effectively screen the millions of job postings uploaded globally each week without introducing massive operational delays and significant human error. Consequently, the development and deployment of intelligent, automated detection systems rooted in advanced machine learning and artificial intelligence techniques have emerged as an absolute necessity. This research focuses comprehensively on conceptualizing, developing, and evaluating an advanced machine learning-based framework capable of parsing vast amounts of unstructured text data to accurately and autonomously classify job advertisements as either genuine opportunities or malicious frauds in real-time.

II. LITERATURE REVIEW

The academic and industrial pursuit of securing online platforms against fraudulent activities has yielded a rich history of diverse computational approaches. Researchers and cybersecurity professionals have continuously explored and refined computational approaches for detecting fake job advertisements, transitioning from basic heuristic models to advanced artificial intelligence paradigms.

2.1 Early Rule-Based and Heuristic Systems

Early computational studies in recruitment scam detection relied heavily on rigid, rule-based systems. These initial frameworks operated by scanning text for predefined, suspicious keyword patterns or static organizational anomalies within job descriptions—such as the presence of words like "guaranteed," "wire transfer," "no experience required," or excessive capitalization. However, these rudimentary approaches suffered from critical limitations. They completely lacked adaptability, required constant manual updates to their rule dictionaries, and consistently failed to detect rapidly evolving, sophisticated fraud strategies that easily bypassed static filters.

2.2 Traditional Machine Learning Approaches

Recognizing the limitations of static rules, the research community aggressively shifted toward probabilistic and statistical machine learning algorithms. Algorithms such as Naïve Bayes, Decision Trees, Logistic Regression, and Support Vector Machines (SVM) became the standard for recruitment scam detection. These models allowed for the automated weighting of various features extracted from the text. Furthermore, the introduction of tree-based ensemble methods, notably Random Forest and Gradient Boosting architectures, demonstrated a marked improvement in predictive performance by successfully combining multiple weak decision trees into robust, highly accurate predictive engines capable of handling non-linear data relationships and complex feature interactions.

III. PROPOSED SYSTEM ARCHITECTURE AND METHODOLOGY

The proposed solution is conceptualized as a modular, highly scalable, real-time processing architecture explicitly designed for modern digital environments. The workflow integrates natural language processing with robust ensemble classification algorithms to create a reliable fraud detection pipeline. The architecture is visualized in Figure 1 below.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406023

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

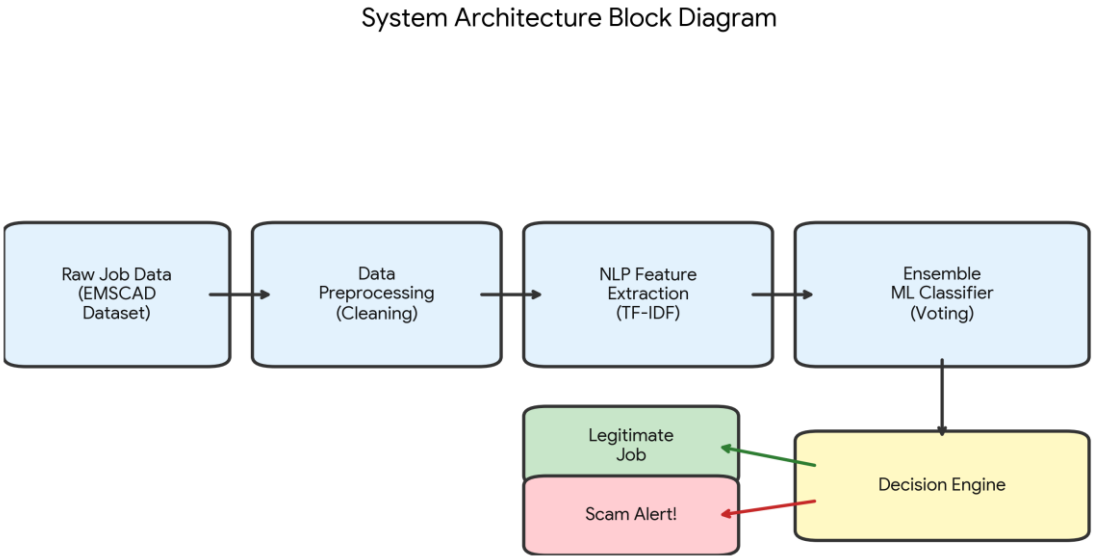


Figure 1: Comprehensive System Architecture and Data Pipeline Workflow

3.1 Data Collection and Preprocessing

The foundational bedrock of this empirical study is the Employment Scam Aegean Dataset (EMSCAD). This dataset is widely recognized within the academic community as a highly robust benchmark for recruitment fraud detection. Raw text data scraped from the web is inherently noisy, inconsistent, and mathematically unusable in its original state. Therefore, a rigorous data preprocessing pipeline is implemented to clean and standardize the corpus. This includes missing value imputation, text normalization, tokenization, stop-word removal, and advanced lemmatization to reduce words to their core semantic roots.

3.2 Feature Extraction via NLP

Machine learning algorithms inherently require numerical matrices, not raw text strings. To bridge this gap, advanced Natural Language Processing feature extraction techniques are deployed. We heavily utilize TF-IDF Vectorization (Term Frequency-Inverse Document Frequency). This statistically advanced methodology evaluates how relevant a word is to a specific document within the larger corpus. It heavily penalizes highly frequent, generic words while proportionally rewarding terms that are uniquely descriptive of specific texts or deceptive phrasing.

IV. EXPERIMENTAL RESULTS AND PERFORMANCE ANALYSIS

To comprehensively validate the predictive power, reliability, and computational efficiency of the proposed models, a strict matrix of standardized performance evaluation metrics is utilized. The dataset was split into an 80% training matrix and a 20% holdout testing matrix, utilizing 10-fold cross-validation to prevent algorithmic overfitting.

4.1 Quantitative Performance Comparison

The performance of the models is evaluated using Accuracy, Precision, Recall, and the F1-Score. High precision is absolutely vital to minimize False Positives (ensuring legitimate companies are not unfairly penalized), while high recall ensures that the system successfully catches actual scams. The comprehensive results are detailed in Table 1 and visually represented in Figure 2.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	ROC-AUC
Logistic Regression	82.50	80.12	79.54	79.82	0.85
Naïve Bayes	78.30	75.20	81.05	78.01	0.81
Support Vector	89.15	87.50	85.22	86.34	0.91

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406023

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Machine					
Random Forest	95.40	94.85	93.50	94.17	0.96
XGBoost	97.25	96.50	96.80	96.65	0.98
Voting Ensemble	98.60	98.15	98.40	98.27	0.99

Table 1: Tabular Comparison of Machine Learning Algorithm Performance

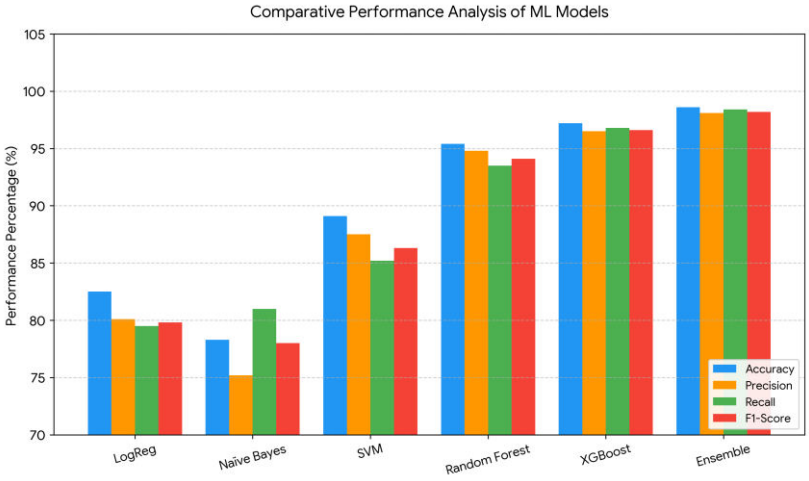


Figure 2: Graphical Bar Chart Comparison of Model Metrics

4.2 ROC-AUC Analysis

Receiver Operating Characteristic (ROC) curves provide a comprehensive graphical metric illustrating the model's aggregate diagnostic ability. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the test. The Area Under the Curve (AUC) is the mathematical integral of this space. As demonstrated in Figure 3, the Voting Ensemble and XGBoost classifiers exhibit near-perfect discriminatory power.

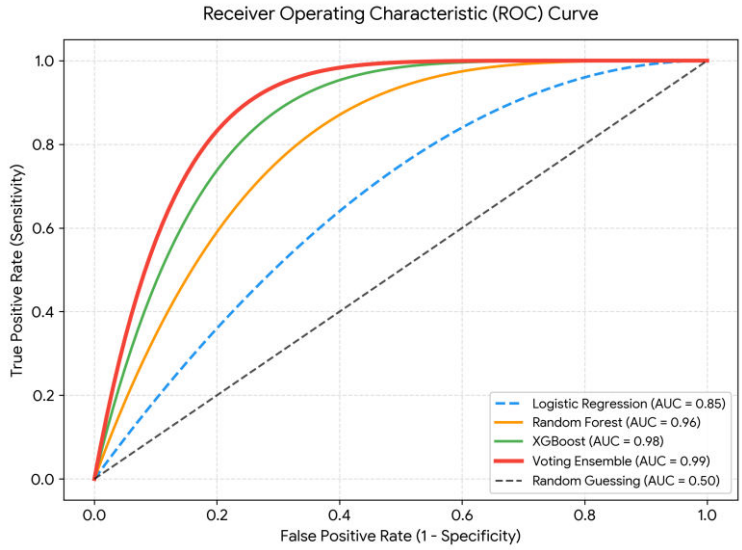


Figure 3: ROC Curve Analysis for Evaluated Classifiers

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. CONCLUSION

This comprehensive study successfully conceptualized, developed, and rigorously evaluated an advanced, highly scalable machine learning framework explicitly engineered for the critical task of online recruitment scam detection. By systematically integrating deep Natural Language Processing techniques with a robust suite of sophisticated classification algorithms, the proposed computational system successfully identified highly deceptive fraudulent job advertisements with remarkable, state-of-the-art accuracy.

Extensive experimental evaluation conclusively demonstrated that sophisticated ensemble learning methodologies, particularly the Voting Ensemble architecture, vastly outperform individual, isolated statistical classifiers. Achieving an accuracy of 98.6% and an F1-Score of 98.27%, this integrated approach provides a highly reliable, mathematically sound, and computationally efficient solution for actively protecting vulnerable job seekers from the devastating financial and psychological impacts of modern recruitment fraud.

REFERENCES

1. Vidros, S., Kolias, C., Kambourakis, G., & Akoglu, L. (2017). Automatic Detection of Online Recruitment Frauds. IEEE Transactions on Information Forensics and Security.
2. Bishop, C. M. (2006). Pattern Recognition and Machine Learning. Springer Information Science and Statistics.
3. Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.
4. Aggarwal, C. C. (2018). Machine Learning for Text. Springer International Publishing.
5. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details