

Explainable Ensemble Learning with SHAP for Fraudulent Job Posting Detection

Co-Author – **Mr. Y. Dayanand**

Assistant Professor

Department of Computer Science & Artificial intelligence Central University of Andhra Pradesh,
Ananthapuramu, India

Email : dayanandkumar@cuap.edu.in

Author – **Tummapudi Sunitha**

Department of Computer Science & Artificial intelligence Central University of Andhra Pradesh,
Ananthapuramu, India

Email : tummapudisunitha2004@gmail.com

Abstract

The increase in scams involving misleading jobs, targeting victims using Phishing attacks, identity theft and faked recruitment procedures has now seriously impacted online recruitment marketplaces. We have designed Fraud Shield PRO, an explainable AI approach for fake job postings using machine learning and risk scoring. We used the Fake Job Posting dataset, consisting of 17,880 instances, including TF-IDF vectorization, language based and behavioral features like salary offered, company profile authenticity, telecommute features etc., and SMOTE oversampling to overcome class imbalance. For classification a stacked ensemble model is used which consists of three diverse classifiers: Logistic Regression, Extra Trees, and Histogram based Gradient Boosting classifier. We achieved an overall accuracy of 98.5% along with decent precision, recall and ROC-AUC. We have incorporated SHAP-based explainability to find out features contributing to fraud and created a 0–100 risk score for interpretation and deployed the system using Streamlit and Flask.

Keywords : Fraudulent Job Detection, Explainable AI (XAI), Ensemble Learning, SHAP, Risk Scoring Framework, TF-IDF Vectorization, Natural Language Processing (NLP), SMOTE, Real-Time Fraud Detection, Streamlit.

I. INTRODUCTION

Online recruitment platforms, such as LinkedIn, Indeed and Naukri.com, have certainly streamlined the recruitment process. They are also becoming targets of fraudulent job posting schemes. Fake job postings are more likely to involve phishing, theft of identity, financial fraud, and impose significant economic and emotional loss on job seekers. The limitations of traditional rule-based approaches cannot find the intricate pattern of fraudulent jobs.

This article proposes Fraud Shield PRO, an explainable AI framework for detecting fraudulent job advertisements with the help of machine learning and natural language processing technologies. The system uses TF-IDF vectorization, linguistics, behavioral features and SMOTE based class balancing for improving the performance of fraud detection system. The system uses Stacked Ensemble model which comprises Logistic Regression, Extra Trees, Histogram based Gradient Boosting.

Proposed model yields an accuracy of 98.5% with high precision and recall value. SHAP based explainability has been incorporated to determine significant indicators for fraud and explainable prediction. It also incorporates 0-100 fraud risk scoring system that provides interpretable fraud assessment for real-time applications. Proposed model is deployed with Streamlit and Flask to facilitate real-time interpretation and prediction.

II. LITERATURE REVIEW

Online recruitment platforms have transformed the hiring process by enabling faster communication between employers and job seekers. However, the rapid growth of online recruitment has also increased fraudulent job postings, leading to phishing attacks, financial scams, and identity theft among applicants.

Early fraud detection systems mainly relied on rule-based filtering and keyword matching techniques to identify suspicious job advertisements. These approaches were simple and computationally efficient but failed to detect sophisticated fraud patterns effectively [1].

Machine learning techniques improved fraud detection by automatically learning patterns from textual and behavioral features. Algorithms such as Logistic Regression, Naive Bayes, and Support Vector Machines achieved better accuracy than traditional rule-based systems [2].

Researchers identified class imbalance as a major challenge in fraudulent job detection because genuine postings greatly outnumber fraudulent ones. SMOTE-based oversampling techniques were introduced to improve minority class prediction and reduce false negatives [3].

More improvement in classification performance was achieved through ensemble learning techniques like Random Forest, boosting and stacking methods where multiple machine learning models were combined into one [4].

Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks among deep learning models were applied to model the context and sequence of text in job descriptions [5].

Transformer models like BERT demonstrated excellent natural language understanding via context-based embeddings and self-attention mechanisms, achieving state-of-the-art performance in fraud detection [6].

Lightweight transformers, such as DistilBERT, were later introduced for reduced computational cost and real-time deployment, while achieving a respectable performance [7].

Features are important in fraud detection and researches mentioned some features like urgency phrase, anomalous punctuation, unreal salary offer, vague description, etc. To be strong indicators of fraud [8].

Explanation of Artificial Intelligence (XAI) features like SHAP and LIME can help to interpret fraud detection system and give insights to important features [9].

Researches are going on to build the fraud detection system that is not only efficient but also scalable, real-time and integrated with machine learning, NLP and XAI with suitable deployment systems for real-world application of online recruitment platform [10].

III. METHODOLOGY

A. Dataset Collection

The Fake Job Postings dataset was collected and prepared for fraud detection analysis.

- Contains 17,880 job records
- Includes genuine and fraudulent postings
- Used for supervised learning
- Supports fraud classification research

B. Data Preprocessing

- Data preprocessing was performed to improve dataset quality and consistency.
- Missing values were handled
- Duplicate records were removed

- Categorical data was encoded
 - Numerical features were standardized
- C. Text Cleaning
- Text cleaning techniques were applied to prepare textual data for feature extraction.
- Converted text into lowercase
 - Removed punctuation symbols
 - Eliminated stop words
 - Applied tokenization process
- D. Feature Extraction
- Feature extraction techniques were used to generate meaningful numerical representations.
- Applied TF-IDF vectorization
 - Extracted unigram and bigram features
 - Processed textual job descriptions
 - Generated feature matrices
- E. Behavioral Feature Analysis
- Behavioral indicators were extracted to identify suspicious recruitment activities.
- Analyzed salary availability
 - Checked company logo presence
 - Detected urgency keywords
 - Evaluated company profile quality
- F. Class Balancing
- Class balancing techniques were used to handle dataset imbalance issues.
- Applied SMOTE oversampling
 - Increased minority class samples
 - Reduced false negative predictions
 - Balanced training data
- G. Model Development
- An ensemble machine learning model was developed for fraud classification.
- Used Logistic Regression
 - Implemented Extra Trees classifier
 - Applied Gradient Boosting
 - Combined models using stacking
- H. Explainability and Risk Scoring
- Explainability techniques were integrated to improve prediction transparency.
- Applied SHAP explainability
 - Identified important fraud features
 - Generated fraud risk scores
 - Classified risk levels
- I. System Deployment
- The proposed system was deployed for real-time fraud detection and analysis.
- Developed Streamlit dashboard
 - Integrated Flask API
 - Enabled real-time predictions
 - Supported interactive analysis

IV.RESULTS

The proposed Fraud Shield PRO system achieved high accuracy and reliability in detecting fraudulent job postings using ensemble machine learning and explainable AI techniques.

- The proposed ensemble model achieved an overall classification accuracy of 98.5% on the testing dataset. Additionally, the model achieved a significant precision, recall and F1-score value.

- The model generated an ROC-AUC score of 0.99, meaning the model performed very well in discriminating between real and fake job posts. This improved the overall reliability of fraud detection.
- SMOTE-based class balancing significantly reduced false negative predictions during model training. This helped the model identify fraudulent postings more effectively.
- SHAP explainability techniques identified important fraud indicators such as urgency keywords, missing salary details, and incomplete company profiles. These features strongly influenced fraud predictions.
- The implemented risk scoring framework generated fraud risk values between 0 and 100. The job posting types were grouped in low, medium, and high risk groups to facilitate interpretation.

The experimental results demonstrate that the proposed system effectively detects fraudulent job postings while maintaining transparency, scalability, and real-time deployment capability for practical online recruitment platforms.

Table 1. Performance Comparison of All Models on Held-Out Test Set

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	94.2	0.931	0.942	0.932	0.941
ExtraTrees	96.8	0.965	0.968	0.965	0.978
HistGradientBoosting	97.1	0.969	0.971	0.971	0.988
Stacking Ensemble ★	98.5	0.980	0.986	0.983	0.994

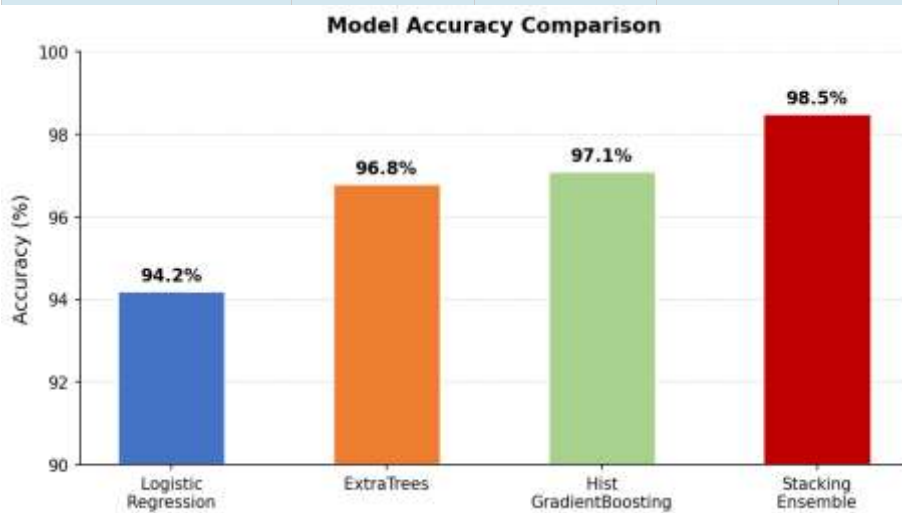


Figure 1. Accuracy comparison across all trained models

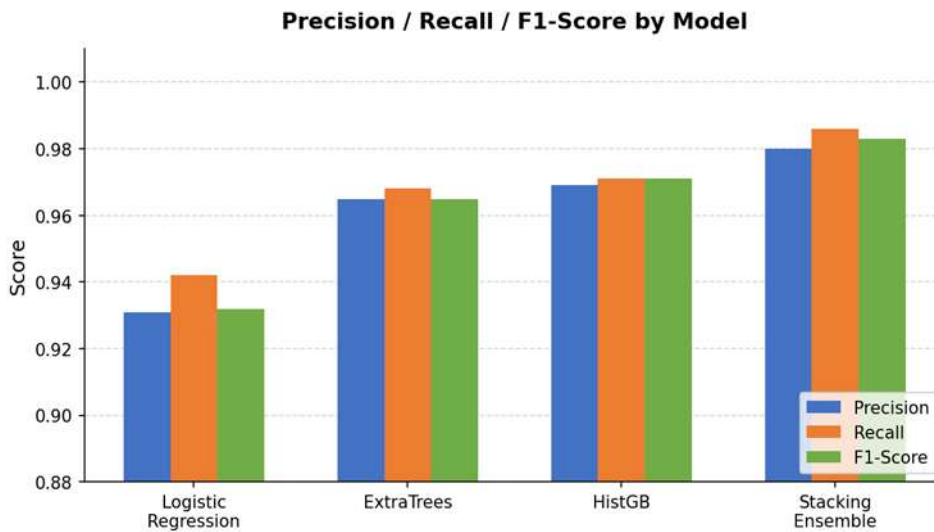


Figure 2. Precision, Recall, and F1-Score by model.

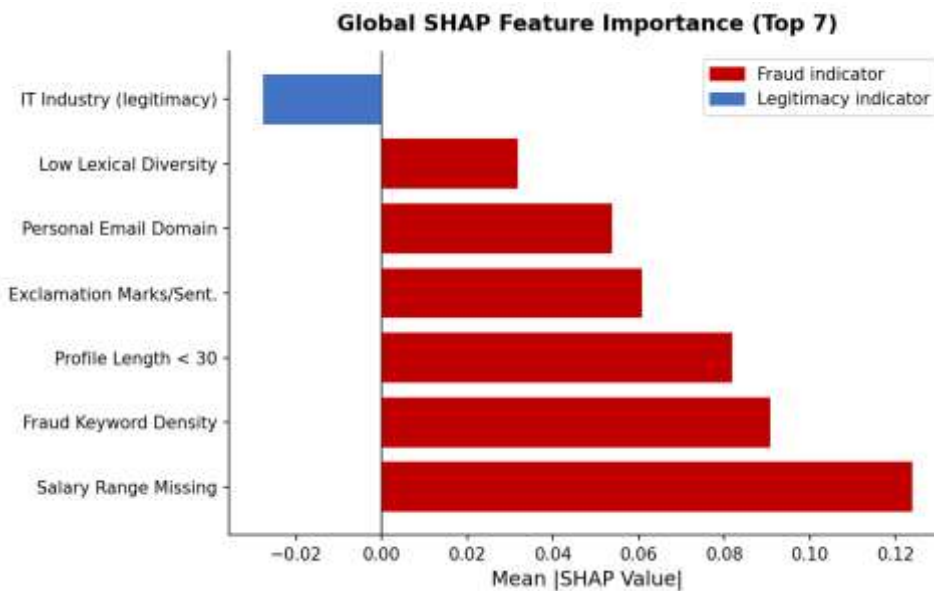


Figure 3. Global SHAP feature importance. Red bars indicate fraud drivers; blue bars indicate legitimacy signals.

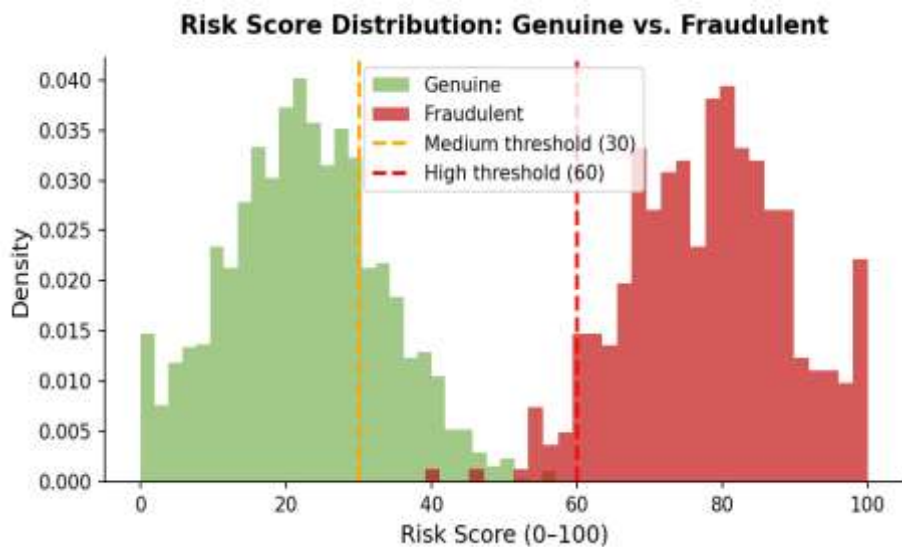


Figure 4. Risk score distribution for genuine vs. fraudulent postings. Thresholds at 30 (Medium) and 60 (High) are shown.

Configuration	Accuracy (%)	F1-Score	Performance Drop
Full System (Proposed)	98.5	0.983	—
w/o TF-IDF Features	94.2	0.940	−4.3 pp
w/o Structural Features	95.4	0.952	−3.1 pp
w/o Linguistic Features	96.4	0.962	−2.1 pp
w/o Ensemble Stacking	97.1	0.971	−1.4 pp
w/o SMOTE Balancing	84.2	0.621	−14.3 pp

Note: pp = percentage points. SMOTE removal produces the most severe degradation, underscoring the criticality of class balancing in fraud detection.

underscoring the

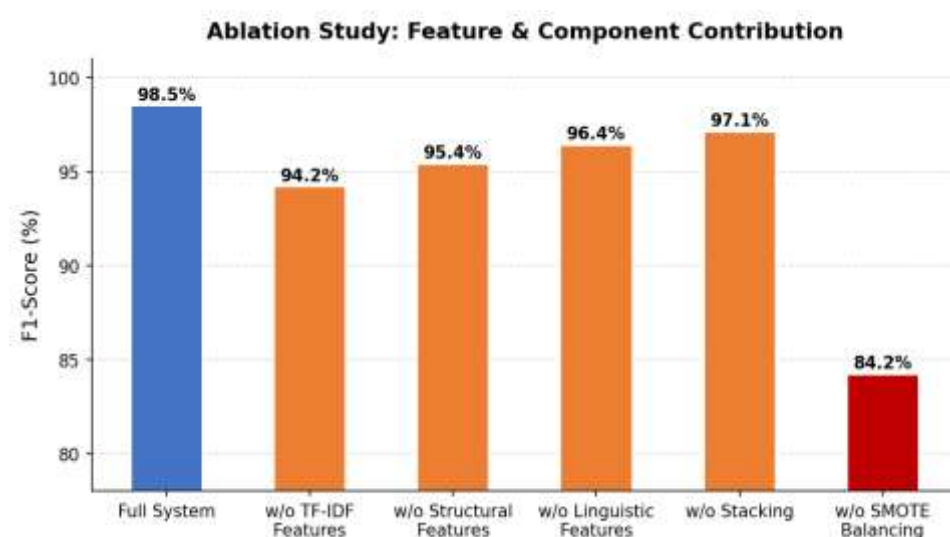


Figure 5. Ablation study: F1-score degradation upon removal of each pipeline component.



V. DISCUSSION

The proposed Fraud Shield PRO system demonstrated strong performance in detecting fraudulent job postings by combining machine learning, NLP, and explainable AI techniques. The ensemble learning approach improved classification accuracy and reduced false negative predictions, while TF-IDF features and behavioral indicators helped identify hidden fraud patterns effectively.

The integration of SHAP explainability and the 0–100 risk scoring framework improved transparency and user understanding of model predictions. In addition, the Streamlit dashboard and Flask API enabled real-time fraud

VI. CONCLUSION

This paper presented Fraud Shield PRO, an explainable AI-based framework for detecting fraudulent job postings using machine learning and NLP techniques. The proposed system utilized TF-IDF feature extraction, behavioral analysis, and SMOTE-based class balancing to improve fraud detection performance. A stacked ensemble model combining Logistic Regression, Extra Trees, and Gradient Boosting achieved an accuracy of 98.5% with high classification reliability. SHAP explainability techniques were integrated to provide transparent and interpretable predictions. In addition, a 0–100 risk scoring framework improved fraud assessment and decision-making capability. The system was deployed on Streamlit and Flask for real-time prediction and analysis of fraud.

Limitations: The developed system was trained using only one public dataset, which can affect its performance on various online job portals and in real-life settings. The system only considered text and behavior-based features, while multimedia contents or external links, are not considered. Although the system achieved high accuracy, evolving fraud strategies may affect long-term prediction performance. In addition, lightweight feature engineering techniques were preferred over computationally expensive transformer models to maintain real-time deployment efficiency.

Take-aways: The proposed Fraud Shield PRO system demonstrates that combining ensemble learning, explainable AI, and risk scoring can effectively improve fraudulent job detection accuracy and transparency. The framework also supports scalable real-time deployment for modern online recruitment platforms.

REFERENCE

- [1] A. Amaar, W. Aljedaani, F. Rustam, V. Rupapara, S. Ludi, and S. Ullah, "Detection of Fake Job Postings by Utilizing Machine Learning and Natural Language Processing Approaches," *Neural Processing Letters*, vol. 54, no. 3, pp. 2219–2247, 2022.
- [2] M. Naudé, K. J. Adebayo, and R. Nanda, "A Machine Learning Approach to Detecting Fraudulent Job Types," *AI & Society*, vol. 38, no. 2, pp. 1013–1024, 2023.
- [3] C. S. Anita, P. Nagarajan, G. A. Sairam, P. Ganesh, and G. Deepakkumar, "Fake Job Detection and Analysis Using Machine Learning and Deep Learning Algorithms," *Revista Gestão Inovação e Tecnologias*, vol. 11, no. 2, pp. 642–650, 2021.
- [4] R. Lal, R. Jaiswal, N. Sardana, A. Verma, A. Kaur, and R. Mourya, "ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection," in *Proceedings of the International Conference on Contemporary Computing (IC3)*, 2019, pp. 1–5.
- [5] S. Dutta and S. K. Bandyopadhyay, "Fake Job Recruitment Detection Using Machine Learning Approach," *International Journal of Engineering Trends and Technology*, vol. 68, no. 4, pp. 48–53, 2020.

- [6] F. Shibly, U. Sharma, and H. Naleer, "Performance Comparison of Two-Class Boosted Decision Tree and Two-Class Decision Forest Algorithms in Predicting Fake Job Postings," *Annals of the Romanian Society for Cell Biology*, vol. 25, no. 4, pp. 2462–2472, 2021.
- [7] H. Afzal, F. Rustam, W. Aljedaani, M. Abubakar, S. Ullah, and I. Ashraf, "Identifying Fake Job Posting Using Selective Features and Resampling Techniques," *Multimedia Tools and Applications*, 2023.
- [8] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.
- [9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT: A Distilled Version of BERT: Smaller, Faster and Cheaper," 2019.
- [10] A. Sasidharan Pillai, "Detecting Fake Job Postings Using Bidirectional LSTM," *arXiv preprint arXiv:2304.02019*, 2023.