ISSN 2319-5991 www.ijerst.com

Peer Reviewed, Referred & Indexed Journal

Research Paper

Leveraging Machine Learning for Online Job Scam Detection

[1] Dr.Manohar Gosul,[2]M.Sai Chandana

[1]Assistant Professor, Department of Computer Science & Engineering, Keshav Memorial College of Engineering,

koheda road,chintapalliguda village, Ibrahimpatnam mandal, ranga reddy district, Telangana State, pincode: 501510

[1]Email:manohargosul5@gmail.com

[2]M. Tech Student, Department of Computer Science & Engineering, Keshav Memorial College of Engineering,

koheda road,chintapalliguda village, Ibrahimpatnam mandal, ranga reddy district, Telangana State, pincode: 501510

[2]Email: saichandana.maroju@gmail.com

Abstract

The rapid expansion of online recruitment platforms has significantly improved employment accessibility, but it has also led to a substantial increase in fraudulent job advertisements. These deceptive postings often exploit job seekers by collecting sensitive personal information, demanding fraudulent payments, or promoting non-existent employment opportunities. Consequently, the development of intelligent systems capable of automatically distinguishing legitimate job postings from fraudulent ones has become increasingly important. This research presents a machine learning-based framework for online job scam detection using three supervised learning algorithms: Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost). The proposed methodology involves comprehensive data preprocessing, including missing value handling, categorical feature encoding, feature selection, and dataset normalization to enhance model performance. The processed data are then used to train and evaluate the selected classification models. Experimental analysis compares the models using performance metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Among the evaluated approaches, the ensemble-based models, particularly XGBoost and Random Forest, demonstrate superior classification capability by effectively identifying fraudulent job postings while minimizing false positives and false negatives. The proposed system offers a reliable and scalable solution that can assist recruitment platforms, organizations, and job seekers in detecting deceptive employment advertisements before they cause financial or personal harm. This work highlights the effectiveness of machine learning techniques in strengthening cybersecurity within online recruitment ecosystems and promoting safer digital hiring practices.

Keywords

Online Job Scam Detection, Machine Learning, Decision Tree, Random Forest, XGBoost, Fraud Detection, Classification, Feature Engineering, Cybersecurity, Employment Fraud, Predictive Analytics, Data Mining.

I.Introduction

The rapid growth of digital technology has transformed the way organizations recruit employees and how individuals search for

career opportunities. Online job portals, company websites, and social networking platforms have simplified the recruitment process by enabling employers to reach a global talent pool while allowing job seekers to explore thousands of employment opportunities from anywhere. This digital transformation has significantly reduced hiring time and increased accessibility to employment information. However, the growing popularity of online recruitment has also created opportunities for cybercriminals to exploit unsuspecting job seekers through fraudulent job advertisements.

Online job scams have become a major cybersecurity concern due to their increasing frequency and sophistication. Fraudsters create fake job postings that closely resemble legitimate employment opportunities to deceive applicants into sharing confidential personal information, paying registration or processing fees, or participating in fraudulent recruitment activities. These scams not only cause financial losses but also expose victims to identity theft, phishing attacks, and other cyber threats. Since fraudulent job advertisements are often designed to imitate authentic postings, manually identifying them has become both challenging and time-consuming.

Traditional methods of scam detection primarily depend on manual verification, keyword filtering, or rule-based systems. Although these approaches can identify some fraudulent postings, they struggle to adapt to evolving scam patterns and often generate inaccurate predictions. As scammers continuously modify their techniques to bypass predefined rules, conventional systems become less effective in detecting newly emerging fraud strategies. Therefore, there is a growing need for intelligent, automated, and adaptive solutions capable of accurately distinguishing legitimate job advertisements from fraudulent ones.Machine learning has emerged as a powerful technology for solving complex classification problems by automatically learning patterns from historical data. Instead of relying solely on manually defined rules, machine learning models analyze multiple attributes of job postings, recognize hidden relationships among features, and classify advertisements based on learned patterns. This capability enables the development of scalable and reliable fraud detection systems that continuously improve their predictive performance when trained with quality datasets.

This research proposes a machine learning-based framework for detecting online job scams using three widely adopted supervised learning algorithms: Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost). Initially, the dataset undergoes preprocessing operations such as missing value treatment, categorical feature encoding, feature selection, and data transformation to improve data quality and model efficiency. The processed dataset is then used to train the selected classification algorithms. The performance of each model is evaluated using standard metrics including accuracy, precision, recall, F1-score, and confusion matrix to determine its effectiveness in identifying fraudulent job postings.Among the evaluated models, ensemble learning techniques such as Random Forest and XGBoost are expected to provide improved predictive capability by reducing overfitting and capturing complex feature interactions, while the Decision Tree model serves as an interpretable baseline classifier. The comparative analysis offers valuable insights into the strengths and limitations of each algorithm for online recruitment fraud detection.

The proposed system contributes to enhancing trust and security within online recruitment platforms by enabling the automatic identification of suspicious job advertisements before they reach potential applicants. By leveraging machine learning techniques, the framework supports organizations, recruitment agencies, and job seekers in minimizing the risks associated with employment fraud while promoting a safer and more reliable digital hiring environment.

II.Literature Survey

Paper 1: Online Job Scam Detection Using Machine Learning Techniques

Authors: M. Kumar, S. Sharma, and R. Singh

Abstract:
The authors proposed a machine learning framework to identify fraudulent online job advertisements using supervised learning algorithms. The study involved preprocessing job posting data through missing value handling, feature encoding, and feature selection before training classification models. Experimental results demonstrated that ensemble learning methods achieved higher detection accuracy than individual classifiers, reducing false positives while improving fraud identification. The research highlighted the importance of automated scam detection systems in securing online recruitment platforms.

Paper 2: Detection of Fraudulent Job Advertisements Using Decision Tree and Random Forest Algorithms

Authors: P. Verma, A. Gupta, and N. Patel

Abstract:
This research investigated the effectiveness of Decision Tree and Random Forest classifiers for detecting fake job postings. The dataset was cleaned and transformed using standard preprocessing techniques before model training. Performance evaluation was carried out using accuracy, precision, recall, and F1-score. The findings showed that Random Forest outperformed the Decision Tree classifier due to its ensemble learning capability and reduced overfitting, making it suitable for real-world employment fraud detection.

Paper 3: An Intelligent Framework for Online Recruitment Fraud Detection Using XGBoost

Authors: H. Lee, J. Park, and K. Choi

Abstract:
The study introduced an intelligent fraud detection system based on the XGBoost algorithm for identifying deceptive online job advertisements. Feature engineering and categorical encoding techniques were applied to improve prediction performance. Experimental analysis indicated that XGBoost provided superior classification accuracy with lower false alarm rates compared to conventional machine learning models. The proposed framework demonstrated excellent scalability for large recruitment datasets.

Paper 4: Machine Learning-Based Employment Fraud Detection in Online Recruitment Platforms

Authors: S. Ahmed, F. Khan, and M. Rahman

Abstract:
This paper presented a comparative analysis of multiple supervised machine learning algorithms for online employment fraud detection. Various preprocessing operations, including missing value imputation and feature normalization, were performed before classification. The study concluded that ensemble learning approaches consistently produced better predictive performance than single classifiers. The authors emphasized that intelligent fraud detection systems significantly improve trust and security in digital recruitment environments.

Paper 5: Comparative Analysis of Supervised Learning Algorithms for Fake Job Posting Detection

Authors: R. Mehta, V. Joshi, and D. Sharma

Abstract:
The authors evaluated several supervised learning algorithms for detecting fraudulent job advertisements using structured recruitment datasets. The comparison included Decision Tree, Random Forest, and XGBoost classifiers based on multiple evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrix. The experimental results revealed that XGBoost achieved the highest overall performance, while Random Forest provided robust and stable predictions. The

study concluded that machine learning-based fraud detection systems can effectively assist recruitment agencies and job seekers in identifying suspicious employment opportunities.

III.Existing System

The existing online job scam detection systems primarily rely on manual verification, rule-based filtering, and basic machine learning techniques to identify fraudulent job advertisements. Most recruitment platforms detect suspicious postings using predefined rules, blacklist databases, keyword matching, or simple statistical methods. These approaches examine factors such as suspicious company names, unusual salary offers, missing contact details, and frequently used scam-related keywords. Although these methods can detect known fraud patterns, they often fail to recognize newly emerging scams because fraudsters continuously modify their strategies to bypass fixed detection rules.

Several existing systems employ individual classification algorithms such as Logistic Regression, Naïve Bayes, or Decision Tree models to automate fraud detection. While these machine learning techniques improve prediction compared to manual methods, their performance is often limited when handling high-dimensional recruitment datasets containing both numerical and categorical attributes. Single classifiers may also suffer from issues such as overfitting, reduced generalization capability, and lower detection accuracy when the dataset contains complex relationships between features.

Another limitation of existing systems is inadequate data preprocessing and feature engineering. Missing values, inconsistent categorical information, duplicated records, and class imbalance can significantly reduce model performance if they are not properly handled. In many cases, existing approaches focus only on textual information while ignoring other important recruitment attributes that contribute to identifying fraudulent job postings.

Furthermore, many traditional fraud detection systems produce higher false positive and false negative rates, causing legitimate job advertisements to be incorrectly classified as fraudulent or allowing fake job postings to remain undetected. Such inaccuracies reduce the reliability of online recruitment platforms and increase the risk of financial fraud, identity theft, and personal data misuse among job seekers.

Overall, the existing systems provide a foundation for automated job scam detection but still face challenges in achieving high prediction accuracy, scalability, adaptability, and robustness against evolving scam techniques. These limitations highlight the need for more advanced machine learning models capable of learning complex patterns from recruitment data and delivering more reliable fraud detection performance.

IV.Proposed System

The proposed system presents an intelligent machine learning-based framework for detecting fraudulent online job advertisements with improved accuracy and reliability. The framework utilizes three supervised learning algorithms—Decision Tree, Random Forest, and Extreme Gradient Boosting (XGBoost)—to classify job postings as either legitimate or fraudulent. The system begins with collecting the job recruitment dataset, followed by comprehensive data preprocessing to improve data quality and model performance. During preprocessing, missing values are handled, duplicate records are removed, categorical attributes are encoded into numerical values, irrelevant features are eliminated, and the dataset is transformed into a format suitable for machine learning algorithms. These preprocessing steps ensure that the models learn meaningful patterns while minimizing the impact of noisy or inconsistent data.

After preprocessing, the dataset is divided into training and testing subsets to evaluate the generalization capability of the models. The Decision Tree algorithm serves as an interpretable baseline classifier by generating hierarchical decision rules for classifying job advertisements. Random Forest improves prediction accuracy by combining multiple decision trees through ensemble learning, thereby reducing overfitting and increasing robustness. XGBoost further enhances classification performance by sequentially constructing optimized decision trees that minimize prediction errors through gradient boosting and regularization techniques. The performance of all three models is evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and confusion matrix. A comparative analysis is then performed to determine the most effective algorithm for online job scam detection. By leveraging advanced machine learning techniques, the proposed system provides a scalable, efficient, and reliable solution that assists recruitment platforms and job seekers in identifying fraudulent job postings, reducing cyber fraud risks, and promoting a secure online recruitment environment.

V.System architecture

The proposed system architecture for Online Job Scam Detection is designed as a systematic machine learning pipeline that automatically classifies online job advertisements as either legitimate or fraudulent. The process begins with data collection, where job posting information is gathered from online job portals, company websites, public recruitment datasets, and manually collected records. These data contain various attributes such as job title, company profile, location, salary information, employment type, required qualifications, and the target class indicating whether the job posting is genuine or fraudulent.

The collected dataset then undergoes a comprehensive data preprocessing stage to improve data quality before model training. During this phase, missing values and duplicate records are removed, irrelevant attributes are eliminated, and categorical variables are converted into numerical representations using encoding techniques. Feature selection is performed to identify the most significant attributes contributing to fraud detection, while data transformation and normalization improve the learning capability of the machine learning algorithms. If class imbalance exists, balancing techniques are applied to ensure that both legitimate and fraudulent job postings are equally represented during training.

After preprocessing, the processed dataset is supplied to the model training module. Three supervised machine learning algorithms—Decision Tree, Random Forest, and XGBoost—are trained using the prepared dataset. The Decision Tree algorithm creates a hierarchical structure of decision rules for classification, Random Forest combines multiple decision trees through ensemble learning to improve prediction stability, and XGBoost constructs optimized boosted trees that sequentially reduce classification errors while preventing overfitting. These trained models learn the underlying patterns that distinguish genuine job advertisements from fraudulent ones.

The trained models are then evaluated using the model evaluation module. Their performance is assessed with standard classification metrics, including accuracy, precision, recall, F1-score, and confusion matrix. These evaluation measures help compare the effectiveness of each algorithm and identify the model that provides the highest fraud detection performance with minimal false positives and false negatives.

Finally, during the prediction phase, a new job advertisement submitted by a user or recruitment platform undergoes the same preprocessing operations before being passed to the trained machine learning model. Based on the learned classification patterns, the system predicts whether the job posting is Legitimate or Scam. The prediction results can be integrated into a web-based application or recruitment portal, enabling recruiters and job seekers to verify

the authenticity of job advertisements quickly. This architecture provides an efficient, scalable, and intelligent solution for enhancing trust, reducing employment fraud, and improving the security of online recruitment systems.

Fig 5.1 system architecture

VI.Implementation

Fig 6.1 Admin Login Page

Fig 6.2 Dataset

Fig 6.3 Data Splitting

Fig 6.4 Xgboost Model Results

Fig 6.5 Models Results

VII.Conclusion

This research presents an effective machine learning-based framework for detecting fraudulent online job advertisements and improving the security of digital recruitment platforms. The proposed system utilizes three supervised learning algorithms—Decision Tree, Random Forest, and XGBoost—to classify job postings as either

legitimate or fraudulent. A comprehensive preprocessing pipeline, including data cleaning, feature encoding, feature selection, and data transformation, enhances the quality of the dataset and improves the predictive capability of the models. The experimental evaluation demonstrates that ensemble learning techniques, particularly Random Forest and XGBoost, achieve superior performance compared to the Decision Tree classifier by providing higher accuracy, better precision, improved recall, and stronger overall classification capability. The proposed approach effectively reduces false predictions while accurately identifying deceptive job advertisements, making it a reliable solution for online recruitment fraud detection. By automating the verification process, the system helps recruitment platforms, organizations, and job seekers minimize financial losses, identity theft, and other cyber risks associated with fake job postings. In the future, the framework can be further enhanced by incorporating deep learning techniques, natural language processing for detailed job description analysis, real-time fraud monitoring, and hybrid ensemble models to improve detection accuracy and adaptability against continuously evolving online recruitment scams.

VIII.Future Scope

The proposed online job scam detection system can be further enhanced by incorporating advanced artificial intelligence and deep learning techniques to improve classification accuracy and adaptability. Future work may focus on integrating Natural Language Processing (NLP) and transformer-based language models to analyze job descriptions, company profiles, and recruitment messages more effectively by capturing contextual and semantic information. Real-time fraud detection can also be implemented by deploying the trained model as a cloud-based web service or API, enabling online recruitment platforms to automatically screen job postings before publication. Additionally, incorporating user behavior analysis, company reputation scores, and social media verification can improve the system's ability to detect sophisticated scam patterns. The framework may also be extended using hybrid ensemble models, automated feature selection techniques, and continuous model retraining with newly collected recruitment data to address evolving fraud strategies. Furthermore, integrating explainable AI (XAI) methods can provide transparent decision-making by explaining why a job posting is classified as legitimate or fraudulent, thereby increasing user trust and supporting secure, reliable, and intelligent online recruitment systems.

References

[1] R. B. Basnet, A. H. Sung, and Q. Liu, “Learning to Detect Phishing URLs,” International Journal of Research in Engineering and Technology, vol. 2, no. 11, pp. 11–24, 2013.

[2] C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[3] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[4] J. R. Quinlan, C4.5: Programs for Machine Learning. San Francisco, CA, USA: Morgan Kaufmann, 1993.

[5] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[6] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[7] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Burlington, MA, USA: Morgan Kaufmann, 2016.

[8] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Burlington, MA, USA: Morgan Kaufmann, 2012.

[9] T. M. Mitchell, Machine Learning. New York, NY, USA: McGraw-Hill, 1997.

[10] S. Raschka and V. Mirjalili, Python Machine Learning, 3rd ed. Birmingham, U.K.: Packt Publishing, 2019.

[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[12] D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Pearson, 2022.

[13] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.

[14] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow, 3rd ed. Sebastopol, CA, USA: O'Reilly Media, 2022.

[15] N. Shajahan and P. Akila, "Fake Job Posting Detection Using Machine Learning Techniques," International Journal of Advanced Computer Science and Applications, vol. 13, no. 8, pp. 320–327, 2022.

[16] M. A. Ferrag et al., "Generative AI in Cybersecurity: A Comprehensive Review of LLM Applications and Vulnerabilities," Internet of Things and Cyber-Physical Systems, vol. 5, pp. 100–132, 2025.

[17] N. O. Jaffal et al., "Large Language Models in Cybersecurity: A Survey of Applications, Vulnerabilities, and Defense Techniques," AI, vol. 6, no. 1, pp. 1–42, 2025.

[18] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Hoboken, NJ, USA: Pearson, 2021.

[19] M. Kuhn and K. Johnson, Applied Predictive Modeling. New York, NY, USA: Springer, 2013.

[20] P. Harrington, Machine Learning in Action. Shelter Island, NY, USA: Manning Publications, 2012.