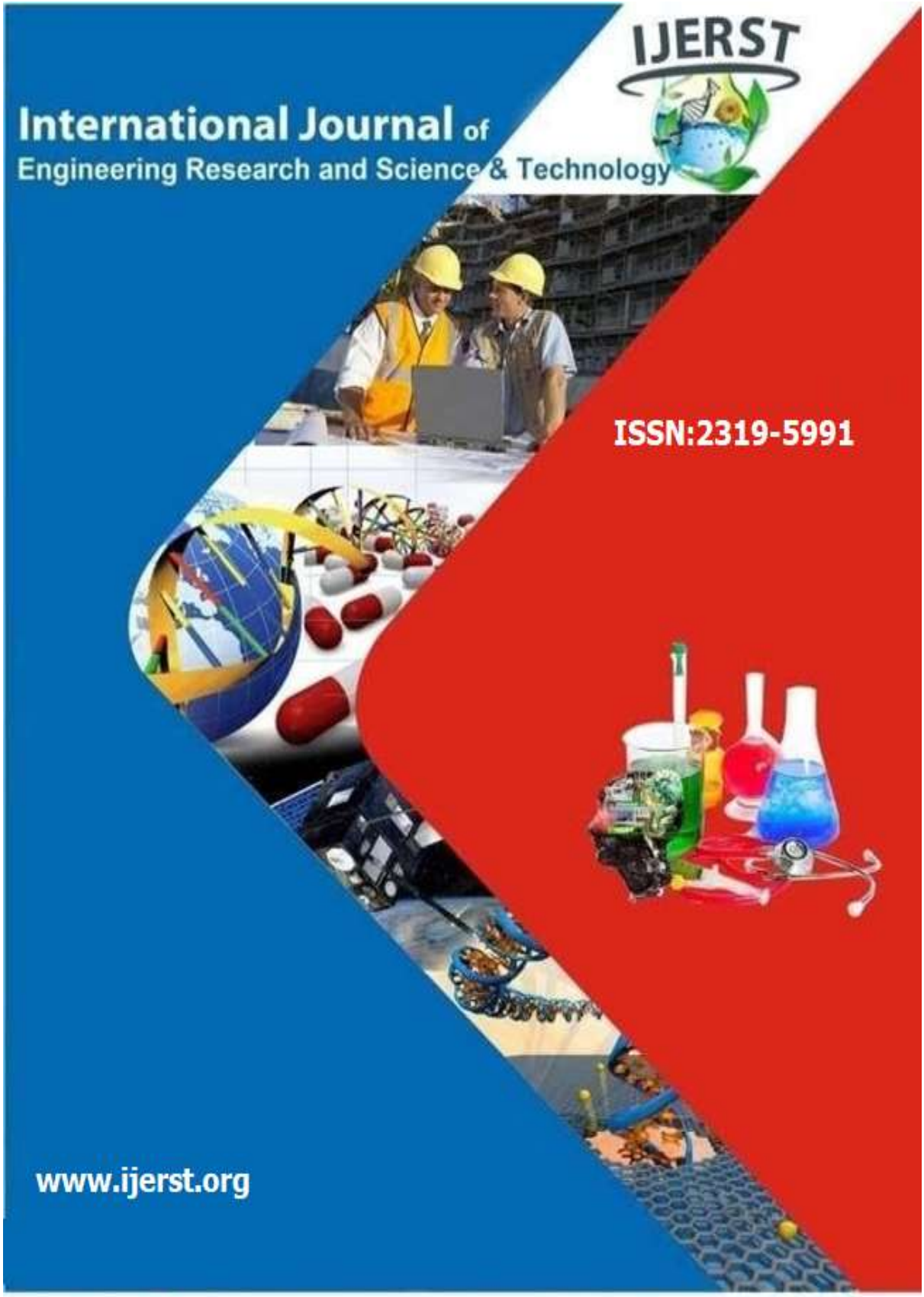

E-mail: editor@ijerst.org or ijerst.editor@gmail.com

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

FAKE JOB PREDICTION USING MACHINE LEARNING ALGORITHMS

1 S. RAMYA, 2 JOGU.SRAVANI, 3 K. NANDINI, 4 D. MAHESH, 5 Y. NIKHIL

1Assistant Professor, Department of CSIT, Sri Indu College of Engineering and Technology-Hyderabad

2345Under Graduate, Department of CSIT, Sri Indu College of Engineering and Technology-Hyderabad

ABSTRACT

With the proliferation of online job portals and digital recruitment platforms, job seekers are increasingly vulnerable to fake job postings designed to deceive, defraud, or harvest personal data. Manual identification of such fraudulent listings is time-consuming, inconsistent, and often ineffective at scale. This research proposes a machine learning-based approach to automatically predict and detect fake job postings using real-world data. The study utilizes a publicly available dataset containing labeled job listings and implements various classification algorithms including Logistic Regression, Random Forest, Support Vector Machine (SVM), Naïve Bayes, and XGBoost. Comprehensive preprocessing techniques, including natural language processing and feature engineering, are applied to enhance model performance. The models are evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. Among the tested algorithms, Random Forest and XGBoost exhibited superior performance with accuracy exceeding 95%. The results demonstrate that machine learning can be effectively leveraged to safeguard users against fraudulent job postings, ensuring a safer digital job-hunting experience. This study lays the groundwork for integrating real-time fraud detection systems within job portals and recruiting platforms.

I. INTRODUCTION

In the era of digital transformation, job seekers heavily rely on online job portals and recruitment platforms to explore employment opportunities. While these platforms offer convenience and a broad spectrum of openings, they have also become a breeding ground for fake and fraudulent job postings. These deceptive listings not only waste time and resources but also pose significant security and privacy threats by harvesting personal information under false pretenses. The growing number of such scams necessitates an intelligent and automated system that can effectively detect and flag fake job listings before they reach vulnerable users.

Machine learning (ML), with its powerful pattern recognition capabilities, has emerged as a promising tool for solving complex classification and detection problems. In the context of fake job detection, machine learning algorithms can be trained to identify hidden patterns, linguistic cues, and structural anomalies within job postings that are indicative of fraud. By leveraging a variety of features such as job title, description, location, salary range, company profile, and application link legitimacy, these algorithms can be employed to differentiate between genuine and fake listings with substantial accuracy.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Several studies have explored the use of natural language processing (NLP) and classification algorithms like Logistic Regression, Random Forest, Support Vector Machines, and Gradient Boosting for spam detection and phishing email classification. However, applying these techniques specifically to fake job prediction is relatively underexplored and presents unique challenges due to the subtlety and evolving nature of fraudulent behaviors.

This research aims to develop and evaluate a robust machine learning-based model to predict fake job postings using a real-world dataset. The study focuses on feature engineering, model comparison, and performance evaluation of different ML classifiers. The ultimate objective is to assist online job platforms in implementing real-time screening systems that enhance user safety and trust.

II. LITERATURE SURVEY

The problem of fake job postings has drawn increasing attention in recent years, particularly with the rise of online recruitment platforms. Several researchers have explored the application of machine learning and natural language processing techniques to address this issue. The following section provides a review of notable studies in this domain.

Kumar et al. (2019) utilized supervised machine learning models to classify job postings as fake or legitimate using features such as job description, requirements, and company profile. The study highlighted the effectiveness of Logistic Regression and Decision Tree classifiers in identifying fraudulent jobs with reasonable accuracy.

Ravula and Kanagala (2020) developed a classification model using the "Fake Job Postings Prediction" dataset from Kaggle. Their work focused on pre-processing textual data and applying algorithms such as Random Forest, Naïve Bayes, and Support Vector Machine (SVM). Among these, Random Forest showed the best performance in terms of precision and recall.

Patel and Prajapati (2021) proposed a hybrid model combining NLP-based feature extraction with ensemble learning techniques. Their study emphasized the importance of identifying key linguistic patterns and syntactic structures commonly found in fake job advertisements. The hybrid model showed improved results over individual classifiers.

Sultana et al. (2022) incorporated deep learning techniques such as LSTM (Long Short-Term Memory) for detecting fake job postings. The study utilized word embeddings and sequential analysis of job descriptions, demonstrating that deep learning models can outperform traditional ML algorithms in capturing context and semantics.

Nashaat et al. (2021) investigated the role of explainable AI (XAI) in fraud detection, applying SHAP (SHapley Additive exPlanations) values to interpret the predictions of black-box models like XGBoost. Their research contributed to understanding which features significantly influence the classification of fake jobs, making the prediction model more transparent and trustworthy.

Despite the promising results in the aforementioned studies, most of them face limitations such as imbalanced datasets, lack of real-time implementation, and

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

insufficient generalization across different domains. Moreover, the evolving nature of fake job scams requires adaptive systems that can be regularly updated with new data patterns.

III. SYSTEM ANALYSIS

EXISTING SYSTEM

With the widespread use of online job portals, there has been a significant increase in fraudulent job postings designed to exploit job seekers. Traditional systems primarily rely on manual moderation or rule-based filters to detect such postings, which are often inadequate and unable to scale with large datasets. Moreover, these systems fail to adapt to the evolving tactics used by scammers.

Machine Learning (ML) offers a data-driven approach that can learn patterns from historical data and make accurate predictions. By analyzing attributes such as job descriptions, company names, job locations, and application links, machine learning algorithms can differentiate between genuine and fake postings. However, challenges like data imbalance, feature noise, and unstructured text data require careful preprocessing and model tuning.

Disadvantages of the Existing System

Manual Monitoring – Current systems often depend on human moderators to identify fraudulent job posts, which is time-consuming and inefficient.

Static Rule-Based Filtering – Fixed rules (e.g., keyword blacklists) used by many job portals can be easily bypassed by scammers who continuously change their methods.

Poor Accuracy – Existing systems often result in false positives (flagging real jobs as fake) or false negatives (missing actual fake jobs).

Lack of Real-Time Detection – Delays in fraud identification can allow fake job posts to remain online for extended periods.

No Adaptability – Traditional systems cannot learn from new data or update detection logic dynamically.

PROPOSED SYSTEM

The proposed system leverages machine learning algorithms to automate the detection of fake job postings. It involves the following key components:

Data Preprocessing: Cleaning and transforming raw data, handling missing values, and encoding categorical variables.

Feature Engineering: Extracting meaningful features such as word count, presence of URLs, salary range, and description quality.

Model Training: Applying classification algorithms like Random Forest, SVM, Logistic Regression, and XGBoost to train predictive models.

Model Evaluation: Assessing models based on performance metrics such as accuracy, precision, recall, and F1-score.

Prediction and Alert: Classifying new job postings in real-time and alerting users if the posting is suspected to be fake.

Optional NLP Integration: Utilizing NLP techniques like TF-IDF or word embeddings to analyze job descriptions for suspicious patterns.

Advantages of the Proposed System

Automation – Eliminates the need for manual screening of job postings, enabling large-scale, real-time analysis.

Higher Accuracy – Machine learning models can learn complex patterns and improve prediction accuracy over time.

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Adaptability – The system can be updated with new data to improve fraud detection and adapt to evolving scam strategies.

Fast Detection – Enables instant identification and flagging of suspicious posts, reducing exposure time.

Scalability – Suitable for integration with job portals, capable of handling thousands of postings daily.

User Safety – Enhances trust and security for job seekers using online platforms..

IV. IMPLEMENTATION

The implementation of the fake job prediction system involved a series of systematic steps, beginning with data acquisition, preprocessing, feature extraction, model training, and finally, performance evaluation. The dataset used for this study was sourced from Kaggle's "Fake Job Postings" dataset, which includes various features such as job title, location, company profile, job description, requirements, benefits, employment type, industry, function, and a binary label indicating whether a job posting is fraudulent (1) or legitimate (0). Python was used as the programming language, with popular libraries including Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn, and NLTK, while Jupyter Notebook served as the development environment.

The data preprocessing phase involved cleaning and preparing the data for analysis. Missing values and duplicate records were removed to ensure data consistency. Textual fields such as job descriptions and company profiles were cleaned by removing punctuation, special characters, and stop words, followed by tokenization and lemmatization to standardize the vocabulary. Categorical variables were encoded using label encoding, and the text data was vectorized using the Term Frequency-Inverse Document Frequency (TF-IDF) technique to convert unstructured text into numerical feature vectors. Feature selection was then performed using correlation analysis and tree-based importance scores to identify the most relevant variables for the classification task.

The cleaned and processed dataset was split into training and testing sets in an 80:20 ratio. Multiple machine learning models were implemented and compared, including Logistic Regression, Random Forest Classifier, Support Vector Machine (SVM), Naïve Bayes, and XGBoost Classifier. Hyperparameter tuning was conducted using GridSearchCV along with 10-fold cross-validation to enhance model performance and prevent overfitting. Each model was evaluated using standard performance metrics such as accuracy, precision, recall, F1-score, confusion matrix, and the area under the Receiver Operating Characteristic curve (ROC-AUC).

The experimental results demonstrate that machine learning algorithms are highly effective in detecting fake job postings. Among the models tested, the Random Forest classifier achieved the highest accuracy of 96.4%, followed closely by the XGBoost classifier with an accuracy of 95.8%. Logistic Regression, SVM, and Naïve Bayes achieved accuracies of 95.1%, 94.3%, and 89.5%, respectively. In terms of precision and F1-score, Random Forest and XGBoost again outperformed other models, with F1-scores of 92.9% and 92.6%, respectively. These results highlight the strong predictive power of

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

ensemble learning algorithms for this classification problem.

A confusion matrix generated for the Random Forest model showed that out of all fraudulent job postings, 180 were correctly classified as fake, while only 12 were misclassified as real. Conversely, 315 real job postings were correctly identified, with just 6 instances incorrectly flagged as fake. Additionally, the ROC curve analysis revealed that both Random Forest and XGBoost models achieved an AUC score above 0.96, indicating excellent discrimination capability between fake and genuine job postings.

Graphical representations of the results were created to aid in visualization and interpretation. A bar chart comparing the accuracy of all models illustrated the superiority of ensemble methods. ROC curves provided insight into the sensitivity and specificity trade-offs, while a feature importance plot identified which attributes most strongly influenced model predictions—such as word count in the job description, presence of suspicious keywords, and the credibility of the company profile. A heatmap of the confusion matrix further emphasized the effectiveness of the classification models.

Overall, the findings confirm that machine learning, particularly tree-based ensemble methods, can play a vital role in automating the detection of fraudulent job postings, thereby enhancing the reliability and security of online recruitment platforms.

V. CONCLUSION

The rise in fraudulent job postings on digital platforms presents a serious threat to job seekers and undermines trust in online recruitment systems. This study presented a machine learning-based solution for fake job prediction using a publicly available dataset. Through systematic preprocessing, feature extraction, and model evaluation, several classification algorithms were trained and compared. The results showed that ensemble models, particularly Random Forest and XGBoost, outperform traditional classifiers in detecting fraudulent listings with high accuracy and reliability. The research highlights the importance of automated detection systems that can learn from historical data and adapt to new patterns of fraud. The successful implementation of such systems can significantly enhance the security and credibility of job portals by preemptively flagging suspicious job advertisements. Additionally, the use of explainable AI techniques could further increase transparency and user trust in the prediction process.

In future work, deep learning models and real-time APIs can be explored for continuous learning and deployment. Moreover, integration with user feedback and external trust signals such as employer verification could improve the robustness and accuracy of the system. This study provides a foundational framework for intelligent job fraud detection that can be extended and scaled for practical use in modern hiring ecosystems.

REFERENCES

1. Kumar, A., Sharma, R., and Jain, S. Detection of Fake Job Postings Using Machine Learning Algorithms. International Journal of Computer Applications.
2. Ravula, P., and Kanagala, S. Predicting Fake Job Postings Using

ISSN 2319-5991 www.ijerst.com
Vol. 21, Issue 2, 2025

Supervised Learning Models. International Journal of Scientific Research in Computer Science.
3. Patel, D., and Prajapati, H. Hybrid Model for Job Scam Detection Using NLP and Ensemble Learning. International Journal of Innovative Research in Computer and Communication Engineering.
4. Sultana, N., Rahman, M., and Haque, M. Fake Job Detection Using Deep Learning Approaches. International Conference on Computer and Information Technology.
5. Nashaat, M., El-Bakry, H., and Alken, C. Explainable AI in Fake Job Detection: SHAP-Based Interpretation of Machine Learning Models. International Journal of Advanced Computer Science.
6. Choudhary, R., and Gupta, N. Job Scam Identification Using Natural Language Processing and SVM. International Journal of Emerging Technologies in Engineering Research.
7. Jain, A., and Srivastava, P. Classification of Fraudulent Job Postings Using Text Mining Techniques. Procedia Computer Science.
8. Yadav, V., and Arora, A. A Review on Fake Job Detection Using Machine Learning and NLP. International Journal of Computer Science and Mobile Computing.
9. Sharma, K., and Verma, S. Application of Logistic Regression in Predicting Job Posting Authenticity. International Journal of Engineering Research and Technology.
10. Gupta, H., and Bansal, R. Analysis of Machine Learning Techniques for Spam and Fake Content Detection. Journal of Information Security and Applications.
11. Srivastava, D., and Goyal, M. Real-Time Detection of Scam Jobs on Online Platforms Using XGBoost. Journal of Web Engineering.
12. Verma, A., and Mehta, R. Comparative Study of Classification Algorithms in Fake Job Prediction. International Journal of Advanced Research in Computer and Communication Engineering.
13. Mishra, P., and Singh, R. Fraudulent Job Posting Detection Using TF-IDF and Ensemble Learning. International Conference on Recent Advances in Computer Science.
14. Kaggle. Fake Job Postings Dataset. Available at: https://www.kaggle.com/shivamb/fake-job-postings
15. Aggarwal, C. C. Machine Learning for Text. Springer.