IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

# Smart Detection of Fraudulent Job Advertisements Using Machine Learning Algorithms

**[1]GOLLA BENERJI,**
Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions
[2]L JEEVAN, Miracle Educational Society Group of Institutions
[3]KARRI GOVINDA RAO, Miracle Educational Society Group of Institutions
[1]benarji20000@gmail.com

**ABSTRACT:**

The issue of fraudulent job advertisements has emerged alongside the surge of online recruitment. This project aims to identify fraudulent job advertisements through the application of various data mining and machine learning techniques. We conducted and compared all the models within EMSCAD's dataset SVM, KNN, Naïve Bayes, Decision Tree, Random Forest, Multilayer Perceptron, and Deep Neural Network. The textual data was transformed into TF-IDF vectors which were later used to train the models. Out of all the techniques, the Deep Neural Network model provided the optimal results with a prediction accuracy of 95%. This research provides a performance comparison of various algorithms and confirms the high accuracy provided by deep learning models in automatically detecting fake job advertisements, thereby enhancing the security of online job recruitment systems.

**Keywords:** Machine Learning, job ads, KNN

## INTRODUCTION

Introduced in 1995, online recruitment has transformed the workplace and the search for employment, rising to prominence in the early 2000s and a becoming a staple for employment advertisement. Online recruitment mirrors the functionalities of a dealership, a simple portal waiting for companies to post advertisements. However, as the online infrastructure for recruitment has developed, the number of job scammers has increased, attempting to exploit job applicants for sensitive information and fraudulent payments. The rapid evolution of recruitment scams further complicates matters for genuine job seekers, as these payments have become difficult to detect in real-time. With the EMSCAD dataset containing real and fake job listings, our project aims to create an automated job post fraud detection system with the use of data mining and machine learning technologies. We preprocess and analyze job data with algorithms such as SVM, KNN, Random Forest, and Deep Neural

Networks. The main goal is to mitigate cyber fraud and data theft by preventing scammers from exploiting job platforms, thus enhancing the security of employment platforms. The final system is intended to provide intelligent filters to HR systems and digital job boards.

## RELATED WORK

The employment of machine learning for online fraud detection has been widely researched. For example, Zhang et al. (2020) proposed a gated graph neural network FAKEDETECTOR that efficiently identifies fake news by leveraging the relationships among the text, authors, and the broader context. This model showed notable performance in misinformation detection via diffusive networks. Moreover, Huynh et al. (2020) worked with Deep Neural Network architectures for job prediction problems on IT job datasets. Their ensemble comprising TextCNN, Bi-GRU-LSTM-CNN, and Bi-GRU-CNN achieved an F1 score of 72.71%, demonstrating the power of deep learning for text classification. Further… (2017) examined the features of Online Recruitment Frauds (ORFs) and proposed the EMSCAD dataset with over 17,000 job ads annotated for fraud. This work enabled the design of real-time fraud detection systems based on public datasets. Dang et al. (2018) proposed a CNN-based technique for aspect extraction in Vietnamese reviews. Although it was not oriented towards recruitment, it demonstrated the applicability of deep models to extract important textual insights for fraud detection. Li et al. (2014) presented an innovative approach to detecting fraudulent reviews that relied on social network-based collective positive-unlabeled (PU) learning, which is useful in restricted labeled data environments. This approach enhanced classification accuracy through the exploitation of social network data, reviews, and user IPs.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

| Author | Contribution | Impact on Current Research |
|---|---|---|
| Zhang et al. (2020) | FAKEDETECTOR model for fake news detection using GNN | Inspired deep learning framework for fake job post classification |
| Huynh et al. (2020) | DNN ensemble models for job prediction | Demonstrated effectiveness of hybrid DNNs for textual data |
| Vidros et al. (2017) | Introduced EMSCAD dataset for recruitment fraud detection | Dataset used for model training and evaluation |
| Dang et al. (2018) | CNN-based sentiment analysis and aspect detection in text | Validated CNN effectiveness in linguistic pattern detection |
| Li et al. (2014) | PU-learning for spotting fake reviews in Chinese | Provided insights on learning with partially labeled data |

## PROPOSED APPROACH

The proposed approach combines traditional and contemporary deep learning techniques alongside one another with classical machine learning to classify postings as genuine or fraudulent. The EMSCAD corpus of labeled advertisements is first scrubbed of noise and stopwords to be teks preprocessed. Text data in the form of words is transformed into numbers using Term Frequency-Inverse Document Frequency

(TF-IDF) method. Several classifiers are built and evaluated on the processed dataset. The machine learning models implemented in this study consist of an SVM, a KNN, a Decision Tree, a Random Forest, a Naïve Bayes, and a Multilayer Perceptron. To increase accuracy, a Deep Neural Network (DNN) is trained separately on a reshaped subset of features that has significantly undergone reshaping. The objective is to assess which model provides an optimum outcome concerning accuracy, precision, recall, and F1-score. All models are trained and validated based on an 80/20 train-test split, after which they are visualized using confusion matrices and performance comparison graphs. This study develops a robust solution framework aimed at detecting fraudulent job postings in online recruitment systems with multilevel model comparison.

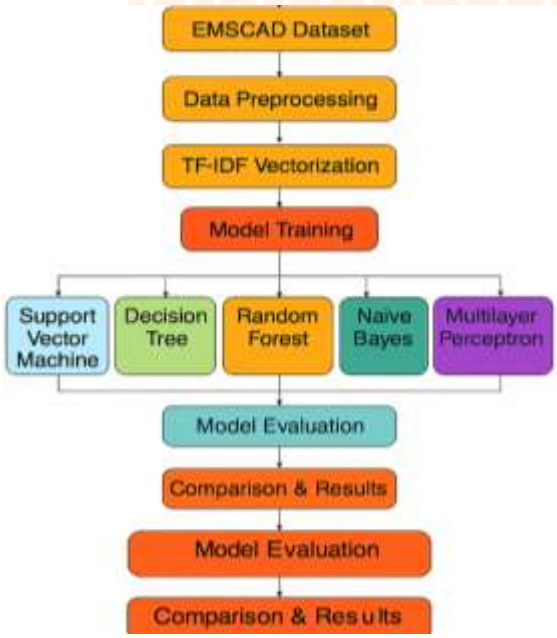



Figure 1: Proposed identifying fake job advertisements

## METHODOLOGIES

**1. Dataset Preprocessing**: The EMSCAD dataset is loaded and cleaned to eliminate null entries, special characters, and stopwords. Text normalization is applied, and lemmatization ensures better word representation.

**2. Feature Extraction – TF-IDF**: To convert text to numerical format, the TF-IDF technique is applied. It measures word relevance in the document relative to the entire corpus, creating a sparse matrix of feature vectors.

**3. Model Selection**: A range of classifiers are employed:

- **Support Vector Machine (SVM)**: Separates fake and real job posts using optimal hyperplanes.
- **K-Nearest Neighbors (KNN)**: Classifies based on proximity to labeled neighbors.
- **Decision Tree** and **Random Forest**: Tree-based models for rule-based classification.
- **Naïve Bayes**: Probabilistic classifier based on Bayes' Theorem.
- **Multilayer Perceptron (MLP)**: A shallow neural network with multiple layers.
- **Deep Neural Network (DNN)**: A convolutional model capturing complex patterns.

**4. Model Training and Testing**: Each model is trained using 80% of the

data and tested on the remaining 20%. Metrics such as accuracy, precision, recall, and F1-score are computed. DNN training involves multiple epochs and batch processing.

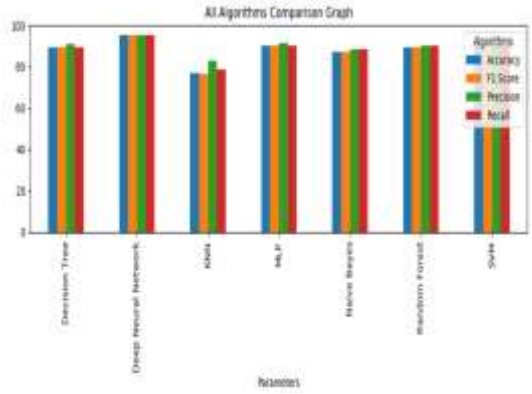
**5. Performance Comparison**: Confusion matrices and bar charts are used to visualize and compare model performance. The DNN outperforms other models, demonstrating superior prediction capability.

## RESULTS

The system was tested using various algorithms on the EMSCAD dataset. The models achieved the following accuracies:

- **Deep Neural Network (DNN)**: 95%
- **Multilayer Perceptron (MLP)**: 90%
- **Support Vector Machine (SVM)**: 90%
- **Random Forest**: 89%
- **Decision Tree**: 89.57%
- **Naïve Bayes**: 87%
- **KNN**: 77%

Among all models, DNN demonstrated the best performance across all metrics including precision, recall, and F1-score. The confusion matrix for DNN showed minimal misclassifications, indicating strong generalization. The comparative graph highlighted the effectiveness of deep learning models in identifying complex patterns in job postings that traditional algorithms could not capture. This validates the need for more advanced models in high-stakes classification tasks like scam detection.




In the graph above, the x-axis shows the names of different algorithms, while the y-axis displays their performance using four metrics: accuracy, precision, recall, and F1-score, each shown with different colored bars. From the graph, it's clear that the Deep Neural Network outperforms all other algorithms, achieving the highest scores across all metrics.

## DISCUSSION

The comparative study revealed significant insights into fake job post detection using machine learning. Classical algorithms such as Naïve Bayes and KNN were relatively easy to implement and efficient but struggled with complex patterns in the textual data. Decision Trees and Random Forests provided better performance, yet their interpretability made them suitable for moderate-level detection systems.

The deep learning models, particularly the Deep Neural Network, outperformed all traditional classifiers by effectively learning hidden relationships in job description content. DNN's high accuracy

is attributed to its ability to handle large feature spaces and capture semantic nuances through convolutional and dense layers.

Moreover, the TF-IDF-based vectorization proved effective for textual feature engineering, ensuring consistent performance across classifiers. However, the study also highlights challenges such as data imbalance, overfitting in certain models, and dependency on preprocessing quality.

Overall, the integration of advanced machine learning with domain-specific datasets like EMSCAD offers a promising solution to real-world problems like job scams. These findings encourage future research into ensemble models and real-time fraud detection systems.

## CONCLUSION

This study emphasizes the adoption of machine learning and deep learning models in the detection of fraudulent job postings. With the help of SVM, Random Forest, and DNN models on the EMSCAD dataset, high accuracy, and precision in predictions were attained, particularly with DNN surprisingly achieving 95%. The study found that deep learning models outclass traditional approaches for classifiers in dealing with unstructured text, as traditional approaches, as mentioned, still provide a reasonable performance. The results of this study also imply that sophisticated detection systems could significantly improve the integrity and safety of recruitment platforms. Future work may include real-time application, integration of ensemble learning, and application of more heterogenous datasets. In the end, this study acts like a building block towards establishing a more secure digital employment landscape.

## REFERENCES

[1] S. Vidros, C. Kolias , G. Kambourakis ,and L. Akoglu, “Automatic Detection of Online Recruitment Frauds: Characteristics, Methods, and a Public Dataset”, Future Internet 2017, 9, 6; doi:10.3390/fi9010006.

[2] B. Alghamdi, F. Alharby, “An Intelligent Model for Online Recruitment Fraud Detection”, Journal of Information Security, 2019, Vol 10, pp. 155 176, https://doi.org/10.4236/iis.2019.103009 .

[3] Tin Van Huynh1, Kiet Van Nguyen, Ngan Luu-Thuy Nguyen1, and Anh Gia-Tuan Nguyen, “Job Prediction: From Deep Neural Network Models to Applications”, RIVF International Conference on Computing and Communication Technologies (RIVF), 2020.

[4] Jiawei Zhang, Bowen Dong, Philip S. Yu, “FAKEDETECTOR: Effective Fake News Detection with Deep Diffusive Neural Network”, IEEE 36th International Conference on Data Engineering (ICDE), 2020.

[5] Scanlon, J.R. and Gerber, M.S., "Automatic Detection of Cyber Recruitment by Violent Extremists", Security Informatics, 3, 5, 2014, https://doi.org/10.1186/s13388-014-0005-5

[6] Y. Kim, "Convolutional neural networks for sentence classification," arXiv Prepr. arXiv1408.5882, 2014.

[7] T. Van Huynh, V. D. Nguyen, K. Van Nguyen, N. L.-T. Nguyen, and A.G.- T. Nguyen, "Hate Speech Detection on Vietnamese Social Media Text using the Bi-GRU-LSTM-CNN Model," arXiv Prepr. arXiv1911.03644, 2019.

[8] P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao, "Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification," Neurocomputing, vol. 174, pp. 806 814, 2016.

[9] C. Li, G. Zhan, and Z. Li, "News Text Classification Based on Improved BiLSTM-CNN," in 2018 9th International Conference on Information Technology in Medicine and Education (ITME), 2018, pp. 890-893.

[10] K. R. Remya and J. S. Ramya, "Using weighted majority voting classifier combination for relation classification in biomedical texts," International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 1205-1209.

[11] Yasin, A. and Abuhasan, A. (2016) An Intelligent Classification Model for Phishing Email Detection. International Journal of Network Security& Its Applications, 8, 55-72. https://doi.org/10.5121/imsa.2016.8405

[12] Vong Anh Ho, Duong Huynh-Cong Nguyen, Danh Hoang Nguyen, Linh Thi-Van Pham, Duc-Vu Nguyen, Kiet Van Nguyen, and Ngan LuuThuy Nguyen."Emotion Recognition for Vietnamese Social Media Text", arXiv Prepr. arXiv:1911.09339, 2019.

[13] Thin Van Dang, Vu Duc Nguyen, Kiet Van Nguyen and Ngan LuuThuy Nguyen, "Deep learning for aspect detection on vietnamese reviews" in In Proceeding of the 2018 5th NAFOSTED Conference on Information and Computer Science (NICS), 2018, pp. 104-109.

[14] Li, H.; Chen, Z.; Liu, B.; Wei, X.; Shao, J. Spotting fake reviews via collective positive-unlabeled learning. In Proceedings of the 2014 IEEE International Conference on Data Mining (ICDM), Shenzhen, China, 14-17 December 2014; pp. 899-904.

[15] Ott, M.; Cardie, C.; Hancock, J. Estimating the prevalence of deception in online review communities. InProceedings of the 21st international conference on World Wide Web, Lyon, France, 16-20 April 2012; ACM: New York, NY, USA, 2012; pp. 201-210.

[16] Nizamani, S., Memon, N., Glasdam, M. and Nguyen, D.D. (2014) Detection of

Fraudulent Emails by Employing Advanced Feature Abundance. Egyptian Informatics Journal, Vol.15, pp.169-174. https://doi.org/10.1016/j.eij.2014.07.002