IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 10, October 2026

An Explainable Machine Learning Approach for Detecting Fraudulent Online Job Postings

Sherkar Akanksha[1], Dhanore Vishakha[2], Thube Sakshi[3], Payal Ulhare[4]

Adsul Technical Campus, Chas[1,2,3]

Assistant Professor, Adsul Technical Campus, Chas[4]

Abstract: Online recruitment platforms make it easier for applicants to discover vacancies, but they can also be misused by people posing as legitimate employers. Such schemes may seek advance fees, personal information, or other benefits from job seekers. Automated screening could help moderators prioritise advertisements that deserve closer inspection. A useful system should also provide reasons for its predictions, rather than returning a risk label alone. This paper outlines an explainable machine-learning workflow that combines advertisement text—including job titles, descriptions, requirements and company profiles—with structured posting attributes. The proposed process includes data auditing, text representation, leakage-aware data splitting, methods for addressing class imbalance, baseline classifiers, and post-hoc explanations with SHAP or LIME. Precision, recall, F1-score, PR-AUC and a confusion matrix are planned for evaluation. The Employment Scam Aegean Dataset (EMSCAD) is a candidate source, subject to checking the downloaded version and its usage conditions. Because implementation and testing have not yet been completed, this paper reports no measured performance. The intended outcome is a decision-support approach that makes evidence and uncertainty visible while keeping final judgements with human reviewers.

Keywords: online recruitment fraud, fake job postings, explainable AI, machine learning, natural language processing, SHAP, LIME, class imbalance

1. INTRODUCTION

Online recruitment services allow organisations to publish vacancies and help applicants search across many opportunities. Their scale and open access, however, can also make it possible to circulate deceptive advertisements. Potential warning signs include demands for payment before hiring, promises that seem implausible, inconsistent employer information, or requests for sensitive data early in the process. These indicators should prompt verification, but none is sufficient on its own to establish that a posting is fraudulent.

Reliable automated detection remains challenging because deceptive advertisements can imitate the wording and layout of genuine vacancies, while fraudulent examples form only a small part of many benchmark datasets. Consequently, a model may achieve strong overall accuracy and still fail to identify a substantial share of scams. Vidros et al. introduced the Employment Scam Aegean Dataset (EMSCAD) for research on online recruitment fraud [1]. Other work has examined machine learning for identifying bias in job advertisements [2], while more recent research has explored contextual language models for fake-job detection [3].

Prediction transparency is another concern. Applicants and moderators may find a risk score difficult to act on if the system does not explain what influenced it. Explainable-AI techniques can highlight words or recorded attributes that contributed to a model output. Such explanations describe the model's behaviour; they cannot establish a recruiter's intent or prove that a posting is fraudulent. For this reason, the proposed workflow pairs classification with explanations that a person can inspect alongside other evidence.

1.1 Research Problem

This study addresses how online job advertisements can be classified as potentially genuine or potentially fraudulent while providing understandable evidence for each prediction. The proposed workflow must account for imbalanced labels, mixed text and structured inputs, possible data leakage, and the different consequences of missed scams and incorrect warnings.

1.2 Aim and Objectives

The aim is to develop a transparent machine-learning workflow for screening online job advertisements. The objectives are to: (i) examine relevant text and metadata features; (ii) compare baseline supervised classifiers; (iii) produce post-hoc explanations using SHAP or LIME; (iv) plan evaluation with measures suited to imbalanced labels; and (v) document limitations and safeguards for responsible use.

1.3 Research Questions

- **RQ1:** How well can supervised models separate potentially fraudulent ads from genuine ones using text and structured features?
- **RQ2:** Which features contribute to individual predictions?
- **RQ3:** Can local explanations help a reviewer inspect a prediction?
- **RQ4:** How do label imbalance, historical data and changing scam patterns affect reliability?

2. RELATED WORK

Vidros et al. presented EMSCAD and described characteristics associated with online recruitment fraud [1]. The dataset provides a shared benchmark for studying this problem. However, the advertisements were collected in an earlier period, so their language and fraud patterns may differ from those seen on current platforms. An implementation should record the exact dataset version and avoid treating performance on a historical benchmark as proof of present-day effectiveness.

Frissen et al. investigated machine-learning methods for recognising bias and discrimination in job advertisements [2]. Although their focus differs from fraud detection, the study illustrates how recruitment language can be analysed computationally. Taneja et al. later presented Fraud-BERT, a transformer-based approach to detecting fake job advertisements [3]. These studies support evaluating simple baseline models alongside more complex language models. Greater model complexity alone, however, does not guarantee clearer explanations or more reliable decisions.

LIME approximates a model locally around a particular example to explain that prediction [4]. SHAP uses Shapley-value concepts to estimate how individual features contribute to an output [5]. In text-classification tasks, these methods can identify influential words or attributes. Their results still require caution: explanations may vary with the model, feature representation and explanation settings, and a feature's influence does not show that it causes fraud. In this study, explanations are intended to support review, not to serve as proof.

3. FUNDAMENTAL CONCEPTS

3.1 Fraudulent Online Job Postings

For this study, a fraudulent posting means an advertisement that misrepresents an employment opportunity with the intention of deceiving applicants. Possible warning signs include requests for upfront payments, employer details that cannot be independently confirmed, implausible promises, or pressure to disclose sensitive information. Yet legitimate employers may also have incomplete profiles, and scammers may copy authentic job descriptions. The system should therefore consider multiple signals and communicate uncertainty rather than treating any single attribute as decisive.

3.2 Supervised Classification

In supervised binary classification, each training example has a known label. The proposed label scheme assigns 0 to genuine advertisements and 1 to fraudulent ones. Candidate approaches include Logistic Regression as a baseline, Multinomial Naive Bayes for sparse text representations, and a tree-based classifier for structured or combined features. The final model should be selected using validation evidence, the available data and computational constraints—not an assumption that one algorithm will necessarily perform best.

3.3 Explainable Artificial Intelligence

A global explanation describes patterns across a dataset, whereas a local explanation focuses on one prediction. The proposed system may use SHAP or LIME to display which text tokens or metadata values influenced the classification of a selected advertisement. Explanations should reflect the trained model, remain understandable to the intended audience, and be accompanied by a clear reminder that predictions can be wrong.

4. PROPOSED METHODOLOGY

The planned work is an applied machine-learning study implemented in Python with tools such as pandas and scikit-learn. SHAP and/or LIME will be considered for the explanation component. The following subsections set out the intended procedure; they do not indicate that model training or evaluation has already taken place.

4.1 Dataset Selection and Audit

EMSCAD is the proposed starting point for the study [1]. The dataset publication reports 17,880 job advertisements:

866 labelled as fraudulent and 17,014 labelled as genuine [1]. Before analysis, the downloaded file will be checked for its version, column definitions, duplicates, missing values, class distribution and licence or terms of use. Counts included in the final report must be calculated from the exact file used in the experiment.

4.2 Preprocessing and Feature Construction

The text fields may include the job title, company profile, description, requirements and benefits. These fields can be represented separately or combined, depending on validation results. Initial preprocessing may involve handling missing values, removing HTML, standardising whitespace and transforming text into TF-IDF features. Stop-word removal and stemming should be evaluated rather than applied automatically, since they can remove useful context. Structured inputs may include employment type, experience, industry, location, and indicators such as whether a company logo or profile is present. Features that reveal the target label or rely on information unavailable at prediction time must be excluded to reduce leakage.

4.3 Data Splitting and Class Imbalance

The dataset should be divided into training and test subsets using stratification so that the class proportions remain similar across the splits. Model selection and threshold adjustment should be performed with training data and, where feasible, a separate validation set or cross-validation. Vectorisers and other transformations that learn from data must be fitted only on the training portion. Class weighting and resampling can be compared, but any resampling must occur within the training workflow—not before the test set is separated.

4.4 Model Training

A starting pipeline can pair TF-IDF text features with Logistic Regression. Multinomial Naive Bayes and a tree-based classifier may be added where the chosen feature representation is suitable. Hyperparameters should be selected using training or validation data. The operating threshold should reflect the trade-off between missed fraudulent posts and false alarms; it should not be assumed in advance that 0.50 is the most appropriate threshold.

4.5 Explainability Layer

For selected advertisements, SHAP or LIME can generate a ranked list or visualisation of features that influenced the model output. A reviewer-facing result could show a risk category, a probability only if calibration has been checked, and a concise explanation tied to the advertisement. The system must not fabricate reasons, imply that association proves causation, or present a model output as a definitive fraud finding.

4.6 Evaluation Plan

Accuracy by itself is insufficient when fraudulent examples are uncommon. The planned evaluation will report precision, recall, F1-score, a confusion matrix and PR-AUC; ROC-AUC may be included as an additional measure. Recall indicates the share of fraudulent advertisements detected, while precision indicates the share of flagged advertisements that are actually fraudulent. Numerical results should be added only after the experiment is run, along with the test procedure and decision threshold.

Table 1. Planned workflow and outputs

Stage	Activity	Output
1. Audit	Verify the file, labels, missing values, duplicates and data version.	Documented dataset
2. Prepare	Clean the text, encode structured fields and build TF-IDF features.	Feature pipeline
3. Train	Fit baseline models using stratified splits and training-only imbalance handling.	Candidate models
4. Evaluate	Measure precision, recall, F1, PR-AUC and the confusion matrix.	Measured test metrics
5. Explain	Generate SHAP/LIME explanations for selected predictions.	Inspectable evidence
6. Review	Inspect errors, limitations and possible fairness issues.	Error analysis

5. EXPECTED CONTRIBUTION AND DISCUSSION

The proposed contribution is a repeatable workflow that combines fraud-risk classification with explanations for individual predictions. Rather than returning only a binary label, the system is intended to show reviewers which recorded features contributed to a warning and help them decide whether further checks are needed. Comparing baseline classifiers may also clarify whether a more complex model improves detection of the minority fraud class. These potential contributions must be assessed through implementation and evaluation.

Automated screening should remain separate from a final judgement about whether a recruiter has committed fraud. Labels such as "potentially suspicious" or "requires review" are more appropriate than making an accusation. Explanations should refer to information present in the advertisement and be followed by human verification of the employer and vacancy.

6. CHALLENGES AND LIMITATIONS

6.1 Class Imbalance

A small number of fraudulent examples can make training less stable, and a high accuracy score may conceal poor detection of scams. Class weighting or oversampling may improve learning from the minority class, but either approach can also increase false positives. Evaluation should therefore examine relevant decision thresholds and report the precision–recall trade-off.

6.2 Dataset Age and Generalisation

EMSCAD contains advertisements gathered in an earlier period [1]. Fraud tactics, platform policies and recruitment language can change, so performance on this dataset may not transfer directly to current postings. Where possible, future work should test the approach on newer, independently labelled examples and monitor performance drift over time.

6.3 Limits of Explanations

SHAP and LIME explain aspects of a model's behaviour; they do not reveal a recruiter's actual intention. A feature may correlate with the dataset label without being a dependable indicator of fraud, and explanations can change when modelling choices change. Their stability and usefulness should be assessed before explanations are used in a real moderation process.

6.4 Privacy, Fairness and Misuse

Job advertisements may include contact details or other personal information. Data should be handled under applicable privacy requirements, and explanations or logs should avoid exposing unnecessary personal details. Language, location and employment type should not be treated as conclusive evidence of fraud. Periodic audits can help identify patterns of false positives that unfairly affect genuine employers or applicants.

7. CONCLUSION

This paper has set out a proposed explainable machine-learning workflow for screening potentially fraudulent online job advertisements. The plan combines textual and structured features, compares baseline supervised classifiers, addresses class imbalance and data leakage, and considers SHAP or LIME for explaining individual predictions. Its central principle is that a risk estimate should be accompanied by inspectable evidence and should assist—not replace—human judgement. Since the pipeline has not yet been implemented and tested, its actual performance remains unknown. The next stage is to build the workflow, record reproducible evaluation results, examine errors and assess whether the approach generalises to more recent data.

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, vol. 9, no. 1, article 6, 2017. https://doi.org/10.3390/fi9010006.

[2] R. Frissen, K. J. Adebayo, and R. Nanda, "A machine learning approach to recognize bias and discrimination in job advertisements," AI & Society, vol. 38, pp. 1025-1038, 2023. https://doi.org/10.1007/s00146-022-01574-0.

[3] K. Taneja, J. Vashishtha, and S. Ratnoo, "Fraud-BERT: Transformer based context aware online recruitment fraud detection," Discover Computing, vol. 28, article 9, 2025. https://doi.org/10.1007/s10791-025-09502-8.

[4] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

[5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.